Welcome Message for the AMIA 2017 Annual Symposium Scientific Program Committee Chair

On behalf of the Vice Chairs and the Scientific Program Committee, I am delighted to present these proceedings of the AMIA 2017 Annual Symposium!

No single meeting captures the breadth and depth of informatics better than the AMIA Annual Symposium. Since the first convening of this meeting, more than forty years ago, the use of informatics has grown exponentially to improve health and to make better healthcare decisions. Today, informatics is the key to accelerating the current goals of healthcare reform.

We received a record number of submissions this year, and I am duly impressed by the high quality of work selected for these proceedings that reflect the breadth of topics across the five domains of informatics: translational bioinformatics, clinical research informatics, clinical informatics, consumer health informatics, and public health informatics.

While it is impossible to fully convey the excitement and energy of the in-person events and presentations of the meeting, these proceedings reflect the core scientific content that were essential for the success of AMIA 2017. These proceedings therefore reflect the continuation of AMIA meetings past, and the beginnings of future discussions of how informatics can truly impact health and health care.

Enjoy!

Neil Sarkar, PhD, MLIS, FACMI
Chair, 2017 Scientific Program Committee
Brown University
2017 AMIA Board of Directors

Officers

Thomas H. Payne, MD, FACMI
University of Washington
Chair

Peter J. Embi, MD, MS, FACMI
Regenstrief Institute
Chair-elect

Titus K. Schleyer, DMD, PhD, FACMI
Regenstrief Institute/Indiana University School of Medicine
Treasurer

Eneida A Mendonça, MD, PhD, FACMI, FAAP
University of Wisconsin – Madison
Secretary

Directors

Tiffany J. Bright, PhD
Agency for Healthcare Research and Quality

Wendy W. Chapman, PhD, FACMI
University of Utah

Sarah A. Collins, RN, PhD
Partners Healthcare System/Brigham and Women’s Hospital/Harvard Medical School

Theresa Cullen, MD, MS
Veterans Health Administration

Paul Fu, Jr., MD, MPH, FAAP
Los Angeles County Harbor-UCLA Medical Center/University of California, Los Angeles

Cynthia S. Gadd, PhD, FACMI
Vanderbilt University

Gretchen Purcell Jackson, MD, PhD
Vanderbilt University Medical Center

Wanda Pratt, PhD, FACMI
University of Washington
Neil Sarkar, PhD, FACMI
Brown University

Dean F. Sittig, PhD, FACMI, FHIMSS
University of Texas Health Sciences Center at Houston

Jessica Tenenbaum, PhD
Duke Translational Medicine Institute

Adam Wright, PhD
Brigham and Women’s Hospital/Harvard Medical School

Ex-Officio Board Members

Christopher G. Chute, MD, DrPH
American College of Medical Informatics President
Johns Hopkins University

Douglas B. Fridsma, MD, PhD, FACP, FACMI
President and Chief Executive Officer
AMIA

Karen A. Monsen, PhD, RN, FAAN
Academic Forum Executive Committee Chair
University of Minnesota School of Nursing

Michael Weiner, DO
Industry Advisory Council Chair
IBM

Laura K. Wiley, PhD
Student Working Group Representative
University of Colorado
AMIA 2017 Annual Symposium Scientific Program Committee

Chair
Neil Sarkar, PhD, MLIS, FACMI
Brown University

Vice Chairs
Jessica Ancker, PhD, MPH, FACMI
Weill Cornell Medical College

Brian Dixon, MPA, PhD
Regenstrief Institute, Inc.

Peter Embi, MD, FACMI
Regenstrief Institute, Inc.

Jessie Tenenbaum, PhD, FACMI
Duke University

AMIA 2018 Chair
Patricia Dykes, RN, PhD, MA, FACMI
Brigham and Women’s Hospital

Members
Jos Aarts, PhD, FACMI
Erasmus University

Elske Ammenwerth, PhD
University for Health Sciences, Medical Informatics and Technology

Nicholas Anderson, PhD
University of California, Davis

James Andrews, PhD
University of South Florida

Mary Regina Boland, MA, MPhil, PhD
University of Pennsylvania

Elizabeth Chen, PhD, FACMI
Brown University

Licong Cui, PhD
University of Kentucky
George Demiris, PhD
University of Washington

Julie Doberne, MD, PhD
Duke University

Robert Freimuth, PhD
Mayo Clinic

Paul Fu, Jr., MD, MPH
Harbor-UCLA Medical Center

Kate Fultz Hollis, MS
Oregon Health & Science University

Kin-Wah Fung, MD
National Library of Medicine

Davera Gabriel, RN
Duke Clinical Research Institute

Jennifer Garvin, PhD
University of Utah

Denise Goldsmith, RN
Brigham and Women's Hospital

Zhe He, PhD
Florida State University

Vitaly Herasevich, MD, PhD
Mayo Clinic

James Hoffman, PharmD
St. Jude Children's Research Hospital

Pei-Yun Sabrina Hsueh, PhD
IBM T.J. Watson Research

Kun Huang, PhD
The Ohio State University
<table>
<thead>
<tr>
<th>Reviewer Name</th>
<th>First Name</th>
<th>Last Name</th>
<th>First Name</th>
<th>Last Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aakre, Christopher</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aaron, Skye</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aarts, Jos</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Abastillas, Glenn</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Abbaszadegan, Hamed</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Abdelrahman, Samir</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Abedtash, Hamed</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Abend, Aaron</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Abhyankar, Swapna</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Abujarad, Fuad</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acharya, Amit</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aczon, Melissa</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adam, Terrence</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adams, Samantha</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adibuzzaman, Mohammad</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Afzal, Naveed</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Agarwal, Kanupriya</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Agresta, Thomas</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ahmad, Suzan</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ahmed, Ali</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Airan-Javia, Subha</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aithem, Sukumar Rao</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Akbilgic, Oguz</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Al Assad, Wasim</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alafaireet, Patricia</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alameddine, Sarah</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Albers, David</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aldekhyyel, Raniah</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alekseyenko, Alexander</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alexander, Gregory</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alhuwail, Dari</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ali, Sana</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alipanah, Neda</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Allin, Sonya</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Almalki, Manal Ali</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Almeida, Mauricio</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Almoeen, Abdulgader</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alonso, William David</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alrige, Mayda Abdullateef</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Al-Taie, Zainab Salam</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alves, Danielle Santos</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alzeer, Abdullah</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Amadi, Tochi</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Amado, Mary</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ambite, Jose-Luis</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Amland, Robert</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ammenwerth, Elske</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ananiadou, Sophia</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Andersen, Tariq</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Andrews, James</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Annechiarico, Robert</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anoop, Shah Dinesh</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Antani, Sameer</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Apathy, Nate</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Araujo, John</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ariosto, Deborah</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Arnold, Corey W</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aronsky, Dominik</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aronson, Samuel</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Arruda-Olson, Adelaide</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asaro, Phillip</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ash, Joan</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Askari, Marjan</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Atnoor, Deven</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aucoin, Jennifer</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Averitt, Amelia Jean</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Avrunin, George</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Azadmanjir, Zahra</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Azevedo, Renato</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aziz, Momin Al</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bailey, Charles</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bailey, Steffani</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bakken, Suzanne</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Balasubramanian, Niranjan</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baldwin, Jessica</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Name</td>
<td>Name</td>
<td>Name</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----------------------------</td>
<td>-----------------------------</td>
<td>-----------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Doing-Harris, Kristina</td>
<td>Eisenberg, Floyd</td>
<td>Fernández Breis, Jesualdo</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Donevant, Sara</td>
<td>Eisenstein, Eric</td>
<td>Tomas</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dong, Ming</td>
<td>Ekberg, Joakim</td>
<td>Figueroa, Rosa Liliana</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dorr, David</td>
<td>Elhadad, Noemie</td>
<td>Filannino, Michele</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dowding, Dawn</td>
<td>El-Kareh, Robert</td>
<td>Finan, Sean</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dreiseitl, Stephan</td>
<td>Elkin, Peter</td>
<td>Finkelstein, Joseph</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Driscoll, Patricia</td>
<td>ElMesiry, Adel</td>
<td>Finnell II, John</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Du, Fan</td>
<td>Eran, Alal</td>
<td>Fisher, Arielle</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Du, Jingcheng</td>
<td>Estiri, Hossein</td>
<td>Fisher, Jason</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Du, Nan</td>
<td>Evans, Scott</td>
<td>Flannery, David</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Du, Shimeng</td>
<td>Exposito, Ernesto</td>
<td>Flegel, Willy</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Du, Xinsong</td>
<td>Fan, Jung-wei</td>
<td>Fleisher, Linda</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dua, Prerna</td>
<td>Farcas, Claudiu</td>
<td>Florez-Arango, Jose</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Duan, Rui</td>
<td>Farcas, Emilia</td>
<td>Flynn, Allen</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Duarte, José Márcio</td>
<td>Farineli, Fernanda</td>
<td>Fodeh, Samah</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Duke, Jon</td>
<td>Farri, Oladimeji</td>
<td>Foraker, Randi</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dullabh, Prashila</td>
<td>Farrugia, Sherry</td>
<td>Fore, Ian</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dumitrache, Anca</td>
<td>Fathiamini, Safa</td>
<td>Forlizzi, Jodi</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dunlea, Robert</td>
<td>Faysel, Mohammad</td>
<td>Fort, Daniel</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dunn, Adam</td>
<td>Fedorov, Andriy</td>
<td>Franklin, Amy</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dunn Lopez, Karen</td>
<td>Feld, Shara</td>
<td>Franzke, Laura</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dutra, Ines</td>
<td>Feldman, Henry Joel</td>
<td>Freimuth, Robert</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dzomba, Bari</td>
<td>Felmeister, Alex</td>
<td>Freundlich, Robert</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Edinger, Tracy</td>
<td>Feng, Gang</td>
<td>Friedman, Carol</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Edmunds, Margo</td>
<td>Feng, Yuanyuan</td>
<td>Fu Jr, Paul</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Effken, Judith</td>
<td>Fenton, Susan</td>
<td>Fujita, Katsuhide</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Einbinder, Jonathan</td>
<td></td>
<td>Fuller, Theresa</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Name</td>
<td>Name</td>
<td>Name</td>
<td></td>
<td></td>
</tr>
<tr>
<td>--------------------------------</td>
<td>--------------------------------</td>
<td>--------------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fung, Kin Wah</td>
<td>Gernant, Stephanie</td>
<td>Gopalan, Raj</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Furberg, Robert</td>
<td>Ghersi, Dario</td>
<td>Gordon, Geoffrey</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Furniss, Stephanie</td>
<td>Giannaris, Pericles</td>
<td>Gordon, William</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gabb, Henry</td>
<td>Giardina, Traber</td>
<td>Gordon, Zachary</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gabriel, Rodney</td>
<td>Gibaud, Bernard</td>
<td>Goswami, Chirayu</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gahbiche-Braham, Souhir</td>
<td>Gilder, Jason</td>
<td>Gotz, David</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Galichi, Leila</td>
<td>Giunti, Guido</td>
<td>Goudey, Benjamin</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gallo De Moraes, Alice</td>
<td>Giuse, Dario</td>
<td>Gourab, Krishnaj</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Galusca, Dragos</td>
<td>Gligorijevic, Djordje</td>
<td>Gouripeddi, Ramkiran</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Galvez, Jorge</td>
<td>Goel, Anupam</td>
<td>Gourley, Gato</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gamache, Roland</td>
<td>Goel, Satyender</td>
<td>Goyal, Nikhil</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gao, Cheng</td>
<td>Goggins, Kathryn</td>
<td>Goyal, Swati</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gao, Grace</td>
<td>Gold, Jeffrey</td>
<td>Grabar, Natalia</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Garcia, Diana Paola</td>
<td>Gold, Rachel</td>
<td>Graetz, Ilana</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gardner, Bret Jarret</td>
<td>Gold, Sigfried</td>
<td>Grasso, Michael</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gardner, Peter</td>
<td>Goldschmidt, Yaara</td>
<td>Green, Beverly</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gartrell, Kyungsook</td>
<td>Goldshtein, Inbal</td>
<td>Griffin, Ashley</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gaudioso, Carmelo</td>
<td>Goldstein, Ayelet</td>
<td>Grosjean, Julien</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gavino, Alex</td>
<td>Goldstein, Benjamin</td>
<td>Gross, Brian</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gawelek, Becky</td>
<td>Gollub, Randy</td>
<td>Grouin, Cyril</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gaynor, Mark</td>
<td>Gong, Yang</td>
<td>Groza, Tudor</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gazarian, Priscilla</td>
<td>Gonna, Hanney</td>
<td>Gu, Yang</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ge, Wendong</td>
<td>Gonzalez Hernandez, Graciela</td>
<td>Guattery, Jason</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ge, Yaorong</td>
<td>Goodman, Kenneth</td>
<td>Guida, Kari</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gennari, John</td>
<td>Goodwin, Rebecca</td>
<td>Guidry, Alicia</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Georgsson, Mattias</td>
<td>Goodwin, Travis</td>
<td>Guimarães, Marcelo</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gerkovitch, Mary</td>
<td></td>
<td>Gundlapalli, Adi</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Name</td>
<td>Name</td>
<td>Name</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----------------------</td>
<td>---------------------------</td>
<td>-----------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hsu, John</td>
<td>Jaana, Mirou</td>
<td>Johnson, Stephen</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hsu, Tsan-sheng</td>
<td>Purcell Jackson, Gretchen</td>
<td>Johnson, Steven</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hsu, William</td>
<td>Jacobs, Jason</td>
<td>Jones, Josette</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hu, Lu</td>
<td>Jacobs, Robin</td>
<td>Jonnagaddala, Jitendra</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hu, Peter</td>
<td>Jadhav, Ashutosh</td>
<td>Joseph, Joshua</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hu, Yiying</td>
<td>Jain, Nilesh</td>
<td>Juhn, Young</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hu, Zhen</td>
<td>Jain, Viral</td>
<td>Jung, Changmi</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Huang, Hui</td>
<td>James, Andrew</td>
<td>Jung, Hyunggu</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Huang, Yan</td>
<td>Jamieson, Trevor</td>
<td>Justice, Amy</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Huang, Yang</td>
<td>Jamoom, Eric</td>
<td>Kabel, Margo</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hughes, Kevin</td>
<td>Jawhari, Badeia</td>
<td>Kaelber, David</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Huguet, Nathalie</td>
<td>Jeffery, Alvin</td>
<td>Kahn Jr, Charles</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hulse, Nathan</td>
<td>Jelen, Benjamin</td>
<td>Kakalasaria, Ankita</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hunter-Zinck, Haley</td>
<td>Jenders, Robert Allen</td>
<td>Kale, David</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hurley, Ann</td>
<td>Jenkins, Melinda</td>
<td>Kalenderian, Elsbeth</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hussain, Mustafa</td>
<td>Jeong, In cheol</td>
<td>Kalpathy-Cramer,</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hussain, Syed Ali</td>
<td>Ji, Xiang</td>
<td>Jayashree</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hussein, Rada</td>
<td>Jiang, Guoqian</td>
<td>Kalsy, Megha</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hylock, Ray</td>
<td>Jiang, Xiaoqian</td>
<td>Kamerick, Michael</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hynes, Denise</td>
<td>Jiang, Yun</td>
<td>Kandula, Vijayabhaskar</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hyun, Sookyung</td>
<td>Jimeno Yepes, Antonio</td>
<td>Kane, Bridget</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Imai, Takeshi</td>
<td>Jing, Xia</td>
<td>Kang, Hong</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Iribarren, Sarah</td>
<td>John, Rita Marie</td>
<td>Kang, Tian</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ishikawa, Charlie</td>
<td>Johnson, Andrew</td>
<td>Kang, Youjeong</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Islam, Tasneem</td>
<td>Johnson, Kevin B.</td>
<td>Gichohi Kanga, Samuel</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ito, Márcia</td>
<td>Johnson, Patricia</td>
<td>Kannampallil, Thomas</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Iyengar, Sriram</td>
<td>Johnson, Sharon</td>
<td>Kannan, Vaishnavi</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Kanter, Andrew Stuart  Khetan, Vanitha  Kolb, Edward
Kanthawala, Shaheen  Khodambashi, Soudabeh  Kondo, Yoshiaki
Kartoun, Uri  Khorsandnia, Bayan Vahid  Kondylakis, Haridimos
Kashyap, Rahul  Khurshid, Anjum  Kong, Ha Kyung
Kashyap, Vipul  Kilicoglu, Halil  Konicek, Debra
Kato-Lin, Yi-Chin  Kim, Christopher  Kooienga, Sarah
Kator, Steven  Kim, George  Kookal, Krishna Kumar
Katz, Marilyn  Kim, Hyeoneui  Koola, Jejo
Kaufman, David  Kim, Katherine  Kopf, Lisa
Kavuluru, Ramakanth  Kim, Min  Korkontzelos, Ioannis
Kebapci, Ayda  Kim, Sujin  Koru, Gunes
Keber, Barbara  Kim, Sun  Koski, Eileen
Keebler, Joseph  Kim, Tae Youn  Kotov, Alexander
Keim, Susan  Kim, Yejin  Krause, Tim
Kench, Anuja  Kimura, Eizen  Krauss, John
Kennedy, Diana  Kinsella, John  Krawiec, Conrad
Kennell Jr, Timothy  Kirchner, Les  Krishnan, Rahul
Kerezoudis, Panagiotis  Kirkendall, Eric  Krive, Jacob
Kerley, Denise  Kite, Bobbie  Kroth, Philip
Khader, Shameer  Klann, Jeffrey  Kucharska-Newton, Anna
Khairat, Saif  Klassen, Prescott  Kuhn, Brooks
Khalifa, Mohamed  Klimov, Denis  Kuhn, Thomson
Khalilia, Mohammed  Knoop, Sarah  Kukafka, Rita
Khan, Faiza  Ko, Hon-Sum  Kumar, Manish
Khan, Naqi  Kobayashi, Shinji  Kunjan, Kislaya
Khare, Ritu  Koch, Sabine  Kuo, Pei-Yi
Kharrazi, Hadi  Kochmann, Matthias  Kuo, Tsung-Ting
<table>
<thead>
<tr>
<th>Lourie, Eli</th>
<th>Marini, Simone</th>
<th>McDaniel, Anna</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lovely, Jenna</td>
<td>Markan-Aurora, Sumita</td>
<td>McDonald, Clement</td>
</tr>
<tr>
<td>Lucic, Ana</td>
<td>Markey, Mia</td>
<td>McFarlane, Timothy</td>
</tr>
<tr>
<td>Lundberg, Jonas</td>
<td>Marko-Holguin, Monika</td>
<td>McGarvey, Sunanda</td>
</tr>
<tr>
<td>Luo, Bo</td>
<td>Marquard, Jenna</td>
<td>McGlothlin, James</td>
</tr>
<tr>
<td>Luo, Heng</td>
<td>Marshall, Nell</td>
<td>McInnes, Bridget</td>
</tr>
<tr>
<td>Luo, Yuan</td>
<td>Marshall, Robert</td>
<td>McIntosh, Leslie</td>
</tr>
<tr>
<td>Lussier, Yves</td>
<td>Marsolo, Keith</td>
<td>McKnight, Lawrence</td>
</tr>
<tr>
<td>Lyalin, David</td>
<td>Martin, Ross D</td>
<td>Meehan, Rebecca</td>
</tr>
<tr>
<td>Mabotuwana, Thusitha</td>
<td>Martinez, Gabriela</td>
<td>Mehta, Prachi</td>
</tr>
<tr>
<td>Magge, Arjun</td>
<td>Martinez Costa, Catalina</td>
<td>Mei, Jing</td>
</tr>
<tr>
<td>Mahajan, Satish</td>
<td>Martinez-Romero, Marcos</td>
<td>Mendes, David</td>
</tr>
<tr>
<td>Mahouj, Malika</td>
<td>Martin-Sanchez, Fernando</td>
<td>Meng, Frank</td>
</tr>
<tr>
<td>Major, Vincent</td>
<td>Massengale, Jill</td>
<td>Mentis, Helena</td>
</tr>
<tr>
<td>Malec, Scott</td>
<td>Massoudi, Barbara</td>
<td>Mercincavage, Lauren</td>
</tr>
<tr>
<td>Malmasi, Shervin</td>
<td>Matentzoglu, Nicolas</td>
<td>Mericle, Lee</td>
</tr>
<tr>
<td>Maloney, Christopher</td>
<td>Mathis, Michael</td>
<td>Meyer, Anne-Marie</td>
</tr>
<tr>
<td>Manaktala, Sharad</td>
<td>Mathur, Puneet</td>
<td>Meystre, Stephane</td>
</tr>
<tr>
<td>Manard, William</td>
<td>Matiasz, Nicholas</td>
<td>Meztis, Nicholas</td>
</tr>
<tr>
<td>Manders, Eric-Jan</td>
<td>Maygers, Joyce</td>
<td>Michalowski, Wojtek</td>
</tr>
<tr>
<td>Mankowitz, Scott</td>
<td>McCallie Jr, David</td>
<td>Midboe, Amanda</td>
</tr>
<tr>
<td>Manning, John</td>
<td>McClay, James</td>
<td>Middleton, Blackford</td>
</tr>
<tr>
<td>Manojlovich, Milisa</td>
<td>McClure, Robert</td>
<td>Mikles, Sean</td>
</tr>
<tr>
<td>Marceglia, Sara</td>
<td>McCormick-Ricket, Iben</td>
<td>Miksch, Tim</td>
</tr>
<tr>
<td>Marcial, Laura</td>
<td>McCoy, Allison</td>
<td>Miller, Christopher</td>
</tr>
<tr>
<td>Marcilly, Romaric</td>
<td>McCoy Jr, Thomas</td>
<td>Miller, Timothy</td>
</tr>
<tr>
<td>Marcu, Oana</td>
<td>McCullagh, Lauren</td>
<td>Milosevic, Nikola</td>
</tr>
<tr>
<td>Name</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Miloulis, Stavros</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Min, Hua</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Minich, Thomas</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Minnies, Jessica</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Miotto, Riccardo</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mirkovic, Jelena</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mirro, Michael</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mishra, Ninad</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mishra, Rashmi</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mishra, Vimal</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Misquitta, Leonie</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mitchell, Terri</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mitsch, Christoph</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Modave, Francois</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Moffatt-Bruce, Susan</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mohammad, Amir</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mohan, Santosh</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mollah, Shamim</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Moneda, Manolo</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Monsen, Karen</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Moore, Carlton</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Moore, Jason</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Morales, Miraida</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Morales Nicolás, Antonio</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Morey, Angela</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Morgan, Stephen</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Morid, Mohammad Amin</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mork, James</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Morris, Alan</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Morse, Keith</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mosa, Abu Saleh Mohammad</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mougin, Fleur</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mowery, Danielle</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Muhlbairer, Lawrence</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mukherjee, Sukrit</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Muller, Henning</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Multak, Nina Loren</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mundt, Barbie</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Muñoz, Emir</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Munson, Sean</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Murcko, Anita</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Murphree, Dennis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Murphy, Shawn</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Murphy, Robert</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mussavi Rizi, Seyed Ali</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Musser, R. Clayton</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Muthee, Veronica</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Myerburg, Stuart</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Myers, Risa</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nahlawi, Layan</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nahm, Eun-Shim</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nakayama, Masaharu</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nandigam, Hari Krishna</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nanji, Karen</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Narain, Niven</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Narus, Scott</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nathan, Aaron</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Natsiavas, Pantelis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Naumann, Tristan</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Navar, Ann Marie</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nazi, Kim Marie</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nebekker, Jonathan</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nelson, Stuart</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Neuraz, Antoine</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Neveol, Aurelie</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Newbold, Susan</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Newman, David</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Newman-Griffis, Denis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ng, Kenney</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ngufor, Che</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nguyen, Anthony</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nguyen, Hai-Son</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nguyen, Hien Huy</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nguyen, Vickie</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nho, Kwangsik</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ni, Yizhao</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nicolalde, F. Daniel</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ning, Xia</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Noble, Heather</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nordo, Amy Harris</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Name</td>
<td>Name</td>
<td>Name</td>
</tr>
<tr>
<td>-----------------------------</td>
<td>-----------------------------</td>
<td>-----------------------------</td>
</tr>
<tr>
<td>Norman, Christopher</td>
<td>Ota, Sakiko</td>
<td>Parsad, Nigel</td>
</tr>
<tr>
<td>Norris, Ruth</td>
<td>Overby, Casey Lynnette</td>
<td>Patel, Aditi</td>
</tr>
<tr>
<td>Nosal, Sarah</td>
<td>Overgaard, Shauna</td>
<td>Patel, Jigar Shirish</td>
</tr>
<tr>
<td>Novak, Laurie Lovett</td>
<td>Overhage, J. Marc</td>
<td>Patel, Vimla</td>
</tr>
<tr>
<td>Novince, Ryan William</td>
<td>Ozaydin, Bunyamin</td>
<td>Patel, Jay</td>
</tr>
<tr>
<td>Nunes, Fátima</td>
<td>Ozery-Flato, Michal</td>
<td>Patel, Vipul</td>
</tr>
<tr>
<td>Nystrom, Daniel</td>
<td>Ozbekaynak, Mustafa</td>
<td>Patrick, Timothy</td>
</tr>
<tr>
<td>Nyström, Mikael</td>
<td>Tafti, Ahmad</td>
<td>Patterson, Olga</td>
</tr>
<tr>
<td>Obara, Piotr</td>
<td>Pacheco, Jennifer Allen</td>
<td>Pedersen, Ted</td>
</tr>
<tr>
<td>Obstfeld, Amrom</td>
<td>Paciotti, Brian</td>
<td>Peek, Niels</td>
</tr>
<tr>
<td>Odekunle, Florence</td>
<td>Padula, Michael</td>
<td>Peeples, Malinda</td>
</tr>
<tr>
<td>Ogunyemi, Omolola</td>
<td>Pagliaong, Alessia</td>
<td>Pell, Jonathan</td>
</tr>
<tr>
<td>Oh, Elissa</td>
<td>Pai, Jennifer</td>
<td>Percha, Bethany</td>
</tr>
<tr>
<td>Okamoto, Kazuya</td>
<td>Paik, Kenneth</td>
<td>Perotte, Adler</td>
</tr>
<tr>
<td>Pacheco de</td>
<td>Painter, Ian</td>
<td>Peterson, Josh</td>
</tr>
<tr>
<td>Oliveira, Sergio Ologeanu</td>
<td>Pal, Doyel</td>
<td>Pfeffer, Michael</td>
</tr>
<tr>
<td>Taddei, Roxana</td>
<td>Palchuk, Matvey</td>
<td>Phansalkar, Shobha</td>
</tr>
<tr>
<td>Omoloja, Abiodun</td>
<td>Pan, Chao</td>
<td>Phatakwal, Shantanu</td>
</tr>
<tr>
<td>Onga, Tracy</td>
<td>Pan, Eric Chungren</td>
<td>Phillips, Andrew</td>
</tr>
<tr>
<td>Ong, Irene</td>
<td>Pandolfe, Frank</td>
<td>Phillips, Mark</td>
</tr>
<tr>
<td>Ong, Mei-Sing</td>
<td>Paraiso-Medina, Sergio</td>
<td>Pino, Esteban</td>
</tr>
<tr>
<td>Ongwere, Tom</td>
<td>Paré, Guy</td>
<td>Piscotty Jr, Ronald</td>
</tr>
<tr>
<td>Orenstein, Evan</td>
<td>Parimbelli, Enea</td>
<td>Pivovarov, Rimma</td>
</tr>
<tr>
<td>Orreggio, Giordi</td>
<td>Park, Albert</td>
<td>Plasek, Joseph</td>
</tr>
<tr>
<td>Ortolani, Cristina Lúcia</td>
<td>Park, Anne</td>
<td>Poker, Anna</td>
</tr>
<tr>
<td>Orwoll, Benjamin</td>
<td>Paroubek, Patrick</td>
<td>Polepalli Ramesh, Balaji</td>
</tr>
<tr>
<td>Osterman, Travis</td>
<td>Parry, Dave</td>
<td>Pollack, Ari</td>
</tr>
<tr>
<td>Pollak, JP</td>
<td>Ragaban, Nouran</td>
<td>Redd, Travis</td>
</tr>
<tr>
<td>--------------------------</td>
<td>-------------------------</td>
<td>----------------------</td>
</tr>
<tr>
<td>Polzin, Thomas</td>
<td>Rahmani, Esmaeel</td>
<td>Redd, Douglas</td>
</tr>
<tr>
<td>Poon, Eric</td>
<td>Rahurkar, Saurabh</td>
<td>Reed, Robyn</td>
</tr>
<tr>
<td>Popescu, Mihail</td>
<td>Rajabiyazdi, Fateme</td>
<td>Reeder, Blaine</td>
</tr>
<tr>
<td>Popovic, Jennifer</td>
<td>Rajamani, Sripriya</td>
<td>Reeves, Ruth</td>
</tr>
<tr>
<td>Pore, Meenal</td>
<td>Raje, Satyajeet</td>
<td>Rehman, Shakaib</td>
</tr>
<tr>
<td>Posada, Jose</td>
<td>Ramachandra Rao, Sanjay</td>
<td>Renard, Jean-Marie</td>
</tr>
<tr>
<td>Poterack, Karl</td>
<td>Ramelson, Harley</td>
<td>Reyes Nieva, Harry</td>
</tr>
<tr>
<td>Povalej Brzan, Petra</td>
<td>Ramirez, Gilbert</td>
<td>Reynolds, Tera</td>
</tr>
<tr>
<td>Powers, James</td>
<td>Ramirez, Magaly</td>
<td>Rho, Mi Jung</td>
</tr>
<tr>
<td>Pozhitkov, Alex</td>
<td>Ramljak, Dusan</td>
<td>Richards, Janise</td>
</tr>
<tr>
<td>Pozzi, Giuseppe</td>
<td>Ramly, Edmond</td>
<td>Richardson, Lorilei</td>
</tr>
<tr>
<td>Prager, John</td>
<td>Ramos, S. Raquel</td>
<td>Rick, Steven</td>
</tr>
<tr>
<td>Predmore, Zachary</td>
<td>Ranchal, Rohit</td>
<td>Riedl, Albert</td>
</tr>
<tr>
<td>Pressler, Taylor</td>
<td>Rangachari, Pavani</td>
<td>Rios, Anthony</td>
</tr>
<tr>
<td>Prior, Fred William</td>
<td>Ranganath, Rajesh</td>
<td>Rios Piedra, Edgar</td>
</tr>
<tr>
<td>Pritchard, T. Gail</td>
<td>Rao, Praveen</td>
<td>Rippen, Helga</td>
</tr>
<tr>
<td>Prodduturi, Naresh</td>
<td>Rappoport, Nadav</td>
<td>Ripple, Anna Marie</td>
</tr>
<tr>
<td>Pruinelli, Lisiane</td>
<td>Rashidee, Ali</td>
<td>Ritchie, Marylyn</td>
</tr>
<tr>
<td>Pujades-Rodriguez, Mar</td>
<td>Rashidi, Parisa</td>
<td>Rizk, Stephanie</td>
</tr>
<tr>
<td>Puppala, Mamta</td>
<td>Rasmussen, Luke</td>
<td>Roberts, Catherine</td>
</tr>
<tr>
<td>Purushotham, Sanjay</td>
<td>Raszewski, Rebecca</td>
<td>Roberts, Kirk</td>
</tr>
<tr>
<td>Pyles, Lee</td>
<td>Rauscher, Richard</td>
<td>Robinson, Jamie</td>
</tr>
<tr>
<td>Qin, Zhaohui</td>
<td>Ravkic, Irma</td>
<td>Rocca, Mitra</td>
</tr>
<tr>
<td>Quigg, Richard</td>
<td>Ravvaz, Kouroshe</td>
<td>Roden, Robert</td>
</tr>
<tr>
<td>Quinn, Matt</td>
<td>Razavian, Narges</td>
<td>Roderer, Nancy</td>
</tr>
<tr>
<td>Radhakrishnan, Kavita</td>
<td>Rea, Susan</td>
<td>Rodrigues, Jean-Marie</td>
</tr>
<tr>
<td>Name</td>
<td>Name</td>
<td>Name</td>
</tr>
<tr>
<td>---------------------------</td>
<td>---------------------------</td>
<td>---------------------------</td>
</tr>
<tr>
<td>Rodriguez, Jorge</td>
<td>Russo, Elise</td>
<td>Schneider, Henning</td>
</tr>
<tr>
<td>Rodriguez-Esteban, Raul</td>
<td>Ruttenberg, Alan</td>
<td>Schneider, Jodi</td>
</tr>
<tr>
<td>Rogers, Michelle</td>
<td>Ryan, Beatriz</td>
<td>Schreiber, Richard</td>
</tr>
<tr>
<td>Rogith, Deevakar</td>
<td>Ryu, Euijung</td>
<td>Schreyer, Kraftin</td>
</tr>
<tr>
<td>Rohani Ghahari, Romisa</td>
<td>Sadasivam, Rajani</td>
<td>Schriml, Lynn</td>
</tr>
<tr>
<td>Romano, Joseph</td>
<td>Safarova, Mayya</td>
<td>Scotch, Matthew</td>
</tr>
<tr>
<td>Romero-Gutiérrez, Arturo</td>
<td>Safran, Charles</td>
<td>Scott, Jean</td>
</tr>
<tr>
<td>Ronquillo, Charlene</td>
<td>Sahoo, Satya</td>
<td>Seale, Deborah</td>
</tr>
<tr>
<td>Ronquillo, Jay</td>
<td>Sakakini, Tarek</td>
<td>Segre, Alberto</td>
</tr>
<tr>
<td>Roper, Rebecca</td>
<td>Saleem, Jason</td>
<td>Senathirajah, Yalini</td>
</tr>
<tr>
<td>Rose, Rebecca</td>
<td>Sales, Fernando</td>
<td>Sent, Danielle</td>
</tr>
<tr>
<td>Rosenau, Paul</td>
<td>Saltz, Joel</td>
<td>Seppälä, Selja</td>
</tr>
<tr>
<td>Rosenbloom, S Trent</td>
<td>Saltz, Mary</td>
<td>Sexton, Kevin</td>
</tr>
<tr>
<td>Rosenthal, Alex</td>
<td>Salvi, Elisa</td>
<td>Shah, Ajay</td>
</tr>
<tr>
<td>Ross, Mindy</td>
<td>Sandor, Nicolae</td>
<td>Shah, Harshal</td>
</tr>
<tr>
<td>Rosset, Sophie</td>
<td>Sanger, Patrick</td>
<td>Shahar, Yuval</td>
</tr>
<tr>
<td>Rothman, Brian</td>
<td>Sarantis, Demetrios</td>
<td>Shamsuzzaman, Md</td>
</tr>
<tr>
<td>Ru, Boshu</td>
<td>Sarker, Abeed</td>
<td>Shang, Chao</td>
</tr>
<tr>
<td>Rubenstein, Yaffa</td>
<td>Sathyarayanan, Shrie</td>
<td>Shang, Ning</td>
</tr>
<tr>
<td>Ruiz, Carolina</td>
<td>Raam</td>
<td>Shang, Rung-Ji</td>
</tr>
<tr>
<td>Ruiz Herrera, Victor</td>
<td>Saupp, Christi</td>
<td>Shapiro, Michael</td>
</tr>
<tr>
<td>Ruland, Cornelia</td>
<td>Savage, Erica</td>
<td>Shardlow, Matthew</td>
</tr>
<tr>
<td>Rule, Adam</td>
<td>Schaub, Marc</td>
<td>Sharma, Ashish</td>
</tr>
<tr>
<td>Rumana, Umme</td>
<td>Scheid, Beth</td>
<td>Sharma, Deepak</td>
</tr>
<tr>
<td>Rusin, Craig</td>
<td>Scherpberier, Harm</td>
<td>Shea, Christopher</td>
</tr>
<tr>
<td>Rusincovitch, Shelley</td>
<td>Schlegel, Daniel</td>
<td>Sheets, Lincoln</td>
</tr>
<tr>
<td>Russell, David</td>
<td>Schneider, David</td>
<td>Sheikhtaheri, Abbas</td>
</tr>
</tbody>
</table>
Wang, Yanshan
Wang, Ying
Wangia-Anderson, Victoria
Warner, Jeremy
Wasilewski-Masker, Karen
Wass, Sofie
Webber, Emily
Weber, Griffin
Weber, Ingmar
Weber, Patrick
Wei, Chih-Hsuan
Wei, Duo
Wei, Qiang
Weichelt, Bryan
Weinberg, Stuart
Weiner, Mark
Weinstein, Peter
Weiskopf, Nicole
Welch, Brandon
Wen, Cheng
Wen, Jingran
Weng, Wei-Hung
West, Vivian
Westmorland, Jennifer
Westra, Bonnie
Wi, Chung-II
Wiitala, Wyndy
Wilbanks, Bryan
Wilcox, Adam
Wilgus, Rebecca
Wilk, Szymon
Willett, DuWayne
Williams, Allison
Williams, Jennifer LeAnne
Williams, Marc
Williams, Richard
Williams-Devane, ClarLynda
Wilson, Marisa
Winden, Tamara
Wojtusiak, Janusz
Wood, Elena
Wood, Grant
Woodward, James
Workman, Terri
Wright, Adam
Wu, Chunlei
Wu, Danny TY
Wu, Justina
Wu, Leslie
Wu, Stephen
Wu, Yonghui
Wyatt, Jesse
Xia, Eryu
Xu, Jun
Xu, Songhua
Yahi, Alexandre
Yang, Fei
Yang, Haodong
Yang, Yang
Yanover, Chen
Yao, Lixia
Yao, Patricia
Ye, Cheng
Ye, Ye
Ye, Zhan
Yen, Po-Yin
Yim, Wen-wai
Yin, Xiaoyao
Yorks, Melissa
Yoshida, Eileen
Yu, Denny
Yu, Kun-Hsing
Yu, Sheng
Yu, Xinyu
Yu, Yiqin
Yuan, Chi
Yuen-Reed, Gigi
Yusuf, Rafeek
Zaleski, John
Zavaroni, Carlo
Zayas-Cabán, Teresa  Zhao, Yuehua
Zech, Jennifer Munro Zhao, Shilin
Zech, John Zhen, Yi
Zelmer, Jennifer Zheng, Hua
Zeng, Jia Zheng, Jiaping
Zeng, Jia Zheng, Shuai
Zeng, Ningzhou Zhong, Yizhen
Zeng, Xiaoming Zhou, Leming
Zeng, Yuqun Zhou, Lina
Zeng, Zexian Zhou, Sicheng
Zengul, Ferhat Devrim Zhou, Yujia
Zhang, Enming Zhu, Dongxiao
Zhang, Guo-Qiang Zhu, Haining
Zhang, Hao Howard Zhu, Vivienne
Zhang, Jinghe Zhuang, Yan
Zhang, Lei Zide, Mary
Zhang, Olivia Ruoran Zielinski, Gail
Zhang, Rui Zimolzak, Andrew
Zhang, Songmao Zuckerman, Alan
Zhang, Xi
Zhang, Yaoyun
Zhang, Yi
Zhang, Yili
Zhang, Yiye
Zhang, Yong
Zhang, Yongkang
Zhang, Zhan
NOTICE

Medicine is an ever-changing science. As new research and clinical experience broaden our knowledge, changes in treatment and drug therapy are required. The authors and the publishers of this work have checked with sources believed to be reliable in their efforts to provide information that is complete and generally in accord with the standards accepted at the time of publication. However, in view of the possibility of human error or changes in medical sciences, neither the authors nor the publisher nor any other party who has been involved in the preparation or publication of this work warrants that the information contained herein is in every respect accurate or complete, and they are not responsible for any errors or omissions or for the results obtained from use of such information. Readers are encouraged to confirm the information contained herein with other sources. For example and in particular, readers are advised to check the product information sheet included in the package of each drug they plan to administer to be certain that the information contained in this book is accurate and that changes have not been made in recommended dose or in the contraindication for administration. This recommendation is of particular importance in connection with new or infrequently used drugs.
Enabling Interoperability between Healthcare Devices and EHR Systems

Swapna Abhyankar, MD1, Paul Schluter, PhD2, Kathryn Bennett3, Daniel J. Vreeman, PT, DPT, MS1,4, Clement J. McDonald, MD5

1Regenstrief Institute, Inc, Indianapolis, IN; 2GE Healthcare, Milwaukee, WI; 3IEEE, New York, NY; 4Indiana University School of Medicine, Indianapolis, IN; 5National Library of Medicine, National Institutes of Health, Bethesda, Maryland

Introduction

State-of-the-art healthcare devices are used across healthcare settings and produce an enormous amount of patient data. However, despite the advances in technology and the growing adoption of electronic health record systems (EHRs) in both inpatient and outpatient settings,1 most devices do not communicate with EHRs in a meaningful way. The problem is twofold: 1) device manufacturers use local, often proprietary codes to represent each measurement, so in order for an EHR to interpret device data, the local codes must be mapped to a language that the EHR understands; and 2) every manufacturer or device has its own set of local codes, so a separate map is needed for each device. Most healthcare institutions do not have the resources to create these maps, so the data are lost.

The IEEE Standards Association (IEEE-SA) and Regenstrief Institute, Inc. (RI), are working together to bridge the gap between healthcare devices and EHRs. IEEE-SA is a global standardization body within IEEE that is dedicated to creating consensus-based technology standards,2 including the IEEE device nomenclature, a unified, consensus-based terminology across device manufacturers. RI maintains the Logical Observation Identifiers Names and Codes (LOINC®) terminology, an international standard for health measurements, observations, and documents that is widely used in EHRs and adopted as a national standard in over 25 countries, including the United States.3 Mapping IEEE device nomenclature to LOINC will enable the flow of data from healthcare devices to EHRs and data repositories for use in patient care and research, with the ultimate goal of improving patient outcomes.

Methods

In 2015, RI and IEEE-SA signed a Memorandum of Understanding (MOU) to enable creation of a map between LOINC and the IEEE 11073™ family of standards for medical device communication. We began the mapping effort with the IEEE 11073-10101™ Standard for Health informatics - Point-of-care medical device communication – Nomenclature (“11073-10101”),4 followed by the 11073 10101a™ Nomenclature amendment (“11073-10101a”).5 We specifically included the concepts contained in the harmonized Rosetta (hRTM) content available in the National Institute of Standards and Technology (NIST) Rosetta Terminology Mapping Management System (RTMMS).6

The primary mapping work has been done by RI using: 1) the IEEE 11073 published standard, which includes the systematic name, reference ID, and description for each code; 2) an export of the associated content from NIST RTMMS, including additional vendor-specific descriptions and applicable units of measure; and 3) the LOINC database, which contains the information that defines each LOINC term, as well as relevant metadata.

For each IEEE 11073 concept, we examined all of the available information, and if a matching LOINC concept was found, a map was created between the existing codes. In cases where no appropriate LOINC concept was available but the meaning of the IEEE 11073 concept was clear, a new LOINC term was created. Finally, if the 11073 concept was not entirely clear to the RI team based on the available published resources, we sought input from members of the HL7 Health Care Devices (DEV) Working Group and IEEE 11073 Point of Care Devices (PoCD) Workgroup before mapping.

Results

The first mapping was published with LOINC version 2.54 (December 2015) and was available both as searchable content in the publicly-available LOINC database and as the standalone LOINC/IEEE Medical Device Code Mapping Table file at http://www.loinc.org/downloads. The table contained mappings for variables produced by ventilators, anesthesia gas machines, invasive blood pressure monitoring, and EKG devices. The mapping table includes the LOINC code and Long Common Name, and the IEEE 11073 Nomenclature Reference ID, numeric code (CF_CODE10), dimensionality (DIM), Unified code for units of measure (UCUM) units of measure (UOM), and description (when available). The mapping file was also published in RTMMS.
The initial IEEE 11073 Nomenclature to LOINC mapping included 358 mappings between 348 unique LOINC concepts and 355 out of the 418 active concepts (84.9%) in 11073-10101 that were part of hRTM. Most mappings were 1:1, but in a few cases one IEEE concept was mapped to multiple LOINCs and vice versa. For example, if the DIM of an IEEE concept encompassed more than one LOINC Property, e.g., mass concentration and molar concentration, one IEEE concept was mapped to two LOINC codes, each in a separate row with the appropriate IEEE DIM and UOM specified for each row of the mapping table as shown in Table 1.

Table 1. Example of multiple LOINC codes mapped to a single IEEE Reference ID based on units of measure

<table>
<thead>
<tr>
<th>LOINC code</th>
<th>LOINC Long Common Name</th>
<th>IEEE CF_CODE10</th>
<th>IEEE Reference ID</th>
<th>IEEE DIM</th>
<th>IEEE UOM_UCUM</th>
</tr>
</thead>
<tbody>
<tr>
<td>75928-2</td>
<td>Hemoglobin [Moles/volume] in Arterial blood</td>
<td>CF_CODE10</td>
<td>MDC_CONC_HB_ART</td>
<td>NL-3</td>
<td>mmol/L</td>
</tr>
<tr>
<td>30313-1</td>
<td>Hemoglobin [Mass/volume] in Arterial blood</td>
<td>159764</td>
<td>MDC_CONC_HB_ART</td>
<td>ML-3</td>
<td>g/dL, g/L, g/cL</td>
</tr>
</tbody>
</table>

As we add new LOINC content, both based on IEEE 11073 concepts and requests for LOINC terms independent of this work, we continue to expand the mapping table. New versions are published with each twice-yearly LOINC release as well as added to RTMMS. The mapping file in the December 2016 LOINC release contained 590 rows with 568 unique LOINC concepts mapped to 578 unique IEEE concepts. These 578 concepts represent more than 70% of the approximately 800 active concepts in the combined 11073-10101 and -10101a standards that are included in hRTM and represent expanded coverage of concepts related to anesthesia and physiologic monitoring.

**Discussion**

RI and IEEE have created a mapping that bridges the gap between healthcare devices and EHRs. This map allows information from a variety of devices to flow into and integrate with a patient’s EHR record in a way so that the data are interpretable, actionable, and accessible for research. This is a work in progress, and more than 100 new mappings based on new and existing LOINC terms will be in the June 2017 LOINC release, including content related to ventilator management and personal healthcare devices. We are prioritizing the concepts that are defined clearly and are clinically most important, and some of the IEEE 11073 Nomenclature content, such as device modes, settings, alerts, and alarm triggers, may not need mapping. Future work includes mapping content contained in 10101 amendments in process, including neuromuscular and infusion pump domains.

**Conclusion**

The LOINC/IEEE 11073 mappings provide a path for information to flow from healthcare devices to EHRs.

**References**

Applying a Process-based Framework to examine Interunit Patient Transfers

Joanna Abraham, PhD, Shirley Burton, MPH, Imade Ihianle, MD MPH
University of Illinois at Chicago, Chicago, IL

Introduction
Interunit hospital transfers involve the transfer of patients between clinicians from the sending and receiving units [1, 2]. Approximately on average, 29 million patients in the US are transferred per year from the emergency department (ED) to inpatient clinical units [3] for diagnostic and therapeutic care services. Despite its role in addressing care delivery and management needs [4], the underlying interunit transfer process can be complex and prone to bottlenecks [5, 6], with a detrimental impact on care continuity outcomes [2, 5, 7, 8]. Fragmented interunit transfers have been reported to result in patient non-colocalization (i.e., placing patients in inappropriate beds) [1], inappropriate use of hospital resources, increased length of inpatient stay, patient decompensation leading to emergent transfers to higher-level of care [9], and propagation of a sense of confusion and stress on patients and their families [8]. Several studies on interunit patient transfers have been conducted between emergency and inpatient units [9], between critical care units [10] and between nursing units [8]. The complex nature of interunit transfers can best be characterized by the reported barriers mainly related to communication gaps [7, 9], negotiation failures [3] and care coordination bottlenecks [8] between the units. Although strategies such as the use of standardized protocols [11], incorporation of negotiation skill-based staff training, implementation of inter-service rounds [3] and cross-boundary integrators/mediators for patient transfer coordination [12] have been proposed, we found that interunit transfers continue to pose a threat to patient safety. This can partly be attributed to the fact that these strategies, although useful, serve to work-around the unit-based challenges, without addressing their downstream effects on the rest of the process across the unit boundaries. In order to develop evidence-based guidelines for designing sustainable interunit transfer strategies, we need to gain an in-depth understanding of the entire patient transfer workflow and also identify points of vulnerabilities in the process. To address this, we conducted a multi-site qualitative study of ED-MICU (emergency department to medical intensive care unit) transfer process conducted by residents to examine factors affecting the process and also potential strategies used to address them.

Method
We observed ED-MICU patient transfers at two hospital sites caring for inner-city urban patient population in Illinois. We selected these sites primarily for their varying patient populations and their organizational differences, which may have an impact on the transfer process. Site 1 is a 495-bed acute care hospital and Site 2 is a 210 bed, tertiary care facility for veteran patient population. Care delivery and management activities were handled similarly at both sites by patient care teams. Our primary participants in this study included resident physicians from the ED and MICU at both hospitals. The study was approved by the IRB at both sites and all participants provided written consents for the study. Our data collection was primarily through participant observations of patient transfer activities in the ED and MICU over a 7-month period (500hrs at both sites), followed by semi-structured interviews with residents (n=21- site 1: ED=5; MICU=5; site 2: ED=5 MICU=6). We used the grounded theory analytical approach to code our observational data on ED-MICU transfers and the interview data analysis was informed by the pre-defined codes from the open codes.

Results and Discussion
Using a process-based framework [13], we scoped and characterized the ED-MICU patient transfer process into three phases: pre-transfer (i.e., focused on ED preparatory coordination activities), transfer (i.e., focused on ED-MICU handoffs) and post-transfer (i.e., focused on MICU care management activities). We identified that the inherent activities within and across these phases were interdependent and effective management of interdependencies between the phases led to timely procurement of needed resources, prevention of unnecessary duplicate tests and procedures, patient assignment to appropriate
units, improvements in patient and clinician satisfaction, and also overall care continuity across the ED-MICU care continuum. We also observed that these ED-MICU interdependencies were successfully managed only when there was a seamless and effective transfer of information, responsibility and control from the ED resident to the MICU resident. Interview and observations suggested that the root contributors affecting the transfer of information were primarily related to the use of the telephone medium for verbal communication between the ED and MICU residents and the lack of a standardized handoff template for patient transfer communication and documentation at both hospital sites.

Nevertheless the transfer of responsibility and control, with respect to accountability and authority slightly varied across the sites and was impacted by transfer delays (resulting in ED boarding of MICU patients) due to the lack of available MICU beds.

**Conclusion**

Study findings can have implications for the design and implementation of: (1) structured face-to-face communication, supported by standardized care coordination bridging interventions; (2) process changes related to ED boarding policies and incentive programs for clinician accountability.

**Acknowledgement:** The study was funded through the NSF CISE CRII program (#1463990) to the first author.

**References**

Medical Benefit Drug Claims: Assessing the NDC Documentation Gap
Terrence J Adam, RPh, MD, PhD, Bithia Anderson, PharmD, MPA, Angeline Carlson, PhD, Mahsa Salsabili, PharmD, Glenn Trygstad, MS, Stephen Schondelmeyer, PharmD, PhD.
1. Univ. of Minnesota College of Pharmacy, Minneapolis, MN. 2. Data IQ, Eden Prairie, MN. USA

Introduction: Prescription drugs are typically covered as part of a pharmacy benefit and may be obtained at retail and other pharmacy settings such as community, long term care, mail order, specialty, and clinic pharmacies. In growing frequency, consumers are administered prescription drugs in a range of health care settings such as physicians’ offices, hospital outpatient clinics, emergency departments, urgent care facilities, dialysis clinics, outpatient surgery centers, home health providers, or home infusion centers. These locations administer prescription drugs and related services—generally as a single dose of a drug while the patient is in the office—typically through an insured medical benefit (medical benefit drug claims).

Medical benefit drug claims data is not as easily accessible as traditional pharmacy claims, and key differences exist in how prescription claims are adjudicated under pharmacy benefits as compared with medical benefits. In addition, prescription drug utilization and spending in these non-traditional settings has been increasing substantially. As a result, market reports which typically focus exclusively on the retail pharmacy sector present an incomplete picture and limits our understanding of prescription drug information. This raises issues for health policy, the management of prescription drug spending, and the appropriate use of and accountability for drug therapies as well as the capacity to build a learning health system. In addition, the structured data which is used to identify medication use varies by the setting in which it is delivered. National Drug Codes (NDC) provide the primary data representation for medications in the traditional pharmacy setting with the classification code based on medication labeler (manufacturer or distributor), the product code (identifying the medication strength, dosage and formulation) and package code (package size and type). The codes are defined based on the labeler and Food and Drug Administration assignments. For medical benefit drug claims, the coding for reimbursement utilizes Healthcare Common Procedure Coding System (HCPCS) codes to identify the medication provided to patients which may include Current Procedural Terminology (CPT) codes. Understanding the relationship between NDC and HCPCS codes is a critical unmet need for secondary data analysis of medications.

Methods: Expert review of available sources of medication data was completed using public use and proprietary data sources including the Center for Medicare and Medicaid Services public use HCPCS files, Medi-Span HCPCS Codes Database, Release (9/2015), Medi-Span Price Rx Pro® Release (9/2015, Micromedex, the Minnesota All Payer Claims Database (MN APCD) and expert knowledge (physician and pharmacist experts). Medications, medication administration, medication testing, diagnostic pharmaceuticals and pharmacogenomics testing related services were identified along with their corresponding HCPCS. An iterative development approach was used to develop a final terminology crosswalk file including HCPCS codes, medication class data (Medi-Span generic product identifier) and NDC codes. After developing the medication list, it was then used to identify medication use and associated reimbursement using HCPCS codes in the MN APCD database for validation. In addition, the total cost associated with medical benefit drug claims was also evaluated to identify the approximate healthcare cost burden associated with these medications as compared with traditional pharmacy medications.

Data Source: The Minnesota All Payer Claims Database from the years 2009 to 2013 was used to identify medical benefit drug claims representing over 4 million unique individuals from the state of Minnesota receiving medical care and medication therapy. Data analysis was completed using SAS 9.2 along with Microsoft Excel for data summarization.

Results: The comprehensive HCPCS medication list identified 2175 unique HCPCS codes associated with medical benefit drug claims. In attempting to map the HCPCS codes to NDC codes using database methods, manual review and expert consensus, a total of 2075 of the eligible HCPCS codes were linked to an NDC code with 100 HCPCS lacking an identifiable NDC mapping (4.6% of eligible HCPCS). However, a number of the HCPCS had overlapping mappings to the same NDC, resulting in a one (NDC) to many (HCPCS) relationship for many medications. As a result, a total of 1044 unique NDCs were included in the crosswalk including 1031 instances with multiple HCPCS codes mapping to a single NDC. To evaluate the validity of the medication HCPCS list, the code list was queried against the MN APCD database identifying a total of 1912 of the HCPCS codes which had one or more medical claim instances over the 5 year observation time period or 87.9% of the eligible HCPCS codes.
**Cost Evaluation:** Since the relative numbers of medication codes used with medical benefit drug claims are substantially smaller than for traditional NDCs, the overall medical benefit drug claims and costs were evaluated for their health system relevance. For the last year of the data set (2013), there were 62.1 million traditional pharmacy claims and 14.6 million medical benefit drug claims, representing 19% of total medication claims by volume. For the same year, the cost of traditional pharmacy claims was $4.74 billion and for medical benefit drug claims, the cost was $2.64 billion or 35.8% of medication spending, indicating the medical benefit drug claims had a substantial cost burden and were a large proportion of total medication cost.

**Discussion:** The use of HCPCS provides a means to facilitate reimbursement for medications, but is also creates a substantial problem for secondary data use and quality measures due to insufficient specificity for several classes of medical benefit drug claims. The use of HCPCS for medications also creates a data granularity problem, which is a major concern since the codes may not fit well with the actual drug packaging or typical dosing regimens. The development of higher HCPCS data granularity to better specify dosing related information could enhance data quality. However, the one (NDC) to many (HCPCS) problem occurs, in some instances, as a result of the effort to make the HCPCS function for a variety of dosages which may be used in clinical practice. Further analysis of these instances may help provide insight on potential coding granularity enhancements, which could facilitate better alignment with NDC coding and improve interoperability across the coding systems.

Several HCPCS use the “not otherwise specified”, “not otherwise classified”, or “other” subclasses for certain medications. While providing a coded component for reimbursement, these subclasses fail to provide specific insight on the actual medications utilized. In attempting to find a solution to the low granularity of the HCPCS and to help manage secondary data use, all HCPCS codes were reviewed and linked in a cross-walk file to NDC codes. However, many of the HCPCS did not have a corresponding NDC due to the inclusive nature of the HCPCS coding selection process, which included medication HCPCS codes related to drug administration and drug testing. Outside these “nondrug” elements, which would be expected to have a low propensity to match to NDCs, the ambiguity of the HCPCS code system was a bigger problem. The vast majority of the actual medications likely have an NDC code which could be linked, however, the NDC data was not universally present at the point of claim creation. As a result, the inherent HCPCS ambiguity prevented these medications from being readily linkable retrospectively to other drug information sources even after extensive expert review.

A plausible solution to this linkage problem is the concurrent requirement of NDCs for reimbursement and secondary data set development; however, NDCs may be insufficient by themselves to provide adequate documentation for medical benefit drug claims. The inclusion of additional fields including the quantity dispensed, the quantity administered and the units of measure are needed to complete the documentation of the drug use pathway. The gaps in potential linkage between HCPCS and NDCs are substantial with drug analytic testing, medication administration, and medications used in diagnostics since these medication process components are not generally included in traditional pharmacy practice. Greater granularity of the HCPCS coding or inclusion of other terminologies such as Logical Observation Identifiers Names and Codes (LOINC) may be potential solutions to improve data representation. However, the use of LOINC would create greater complexity in managing comprehensive secondary medication data representation by adding another terminology not typically used for the medication data management process. Potential HCPCS enhancements could focus on aligning the code granularities to actual medication usage patterns which could provide interim improvement in data representation.

Major limitations of medical drug claims data include the lack of direct links in the reimbursement data between the HCPCS and the associated drug NDC, limited drug quantity dispensed/administered data and insufficient units of measure creating major secondary data use problems. These relationships likely exist for most all clinical scenarios, but it was not a vendor requirement in the claims data reporting and potentially may be missing in the payer reimbursement process. The inclusion of these fields is needed for future work to adequately facilitate secondary data use for research, quality improvement and policy evaluation. In addition, for medication-related medical benefit drug claims, the components related to drug administration, testing, and management also need to be evaluated concurrently with the actual drugs to provide a full picture of the medication use process. Given the large component of health cost and clinical utilization volume associated medical benefit drug claims; this component of the medication use cycle needs better data representation for secondary data applications and health policy.

Advanced Use of EHRs in US Hospitals and the Emergence of a Digital "Use" Divide

Julia Adler-Milstein, PhD¹ and A Jay Holmgren, MHI¹
¹University of Michigan, School of Information and School of Public Health, Ann Arbor, Michigan

Introduction

Over the past five years, hospital EHR adoption has become widespread.¹ However, we know little about whether hospitals are using EHRs in advanced ways that are critical to realizing improved patient outcomes. There is concern that while hospitals with fewer resources – small, rural, safety-net² – have largely kept up with EHR adoption¹, they are falling behind in advanced EHR use. In this study, we sought to measure advanced ways in which hospitals use EHRs and to assess evidence of an emerging digital “use” divide.

Methods

We used national hospital data from the 2008 to 2015 AHA Annual and IT Supplement surveys. We first measured adoption rates of “basic” and “comprehensive” EHRs to provide the latest national numbers. We then used newly available questions on the 2015 Supplement to assess hospital use of EHRs and EHR data for (1) ten quality improvement functions and (2) ten patient engagement functions. To assess the potential for a digital “use” divide, we ran logistic regression models to identify hospital characteristics associated with (1) eight or more quality improvement (QI) functions and (2) eight or more patient engagement functions. Analyses were limited to general medical-surgical acute care hospitals in the United States that responded to the 2015 AHA IT Supplement (N=2,803). We used non-response weights to generate nationally representative estimates.

Results

80.5% of hospitals had adopted at least a basic EHR system in 2015, an increase of 5.2 percentage points from 2014.

The most common EHR-supported QI functions were monitoring patient safety (71.4% of hospitals) and creating dashboards of individual provider performance using EHR data (68.1%). (Table 1) The least common function was creating an approach for clinicians to query data (39.8%). Hospitals more likely to have 8 or more EHR-enabled QI functions if they participated in an ACO (OR=1.55; p<0.001), PCMH (OR=1.67; p<0.001), or both (OR=3.06; p<0.001). For-profit hospitals as well as critical access hospitals were less likely to have 8 or more functions adopted (OR=0.33 and 0.58, respectively; p<0.001).

The most common EHR-supported patient engagement functions were patient ability to view data online (95.1% of hospitals) and downloading that information (86.8%). (Table 1) The least common function was submitting patient-generated data (36.7%). Large hospitals (OR=1.97; p<0.001) as well as those participating in an ACO (OR=1.80; p<0.001), PCMH (OR=3.05; p<0.001) or both (OR=3.12; p<0.001) were more likely to have 8 or more functions. Safety-net hospitals (those in the 3rd and 4th quartiles of DSH payments; OR=0.71 and 0.43; p<0.002 and p=0.001, respectively; and critical access hospitals; OR=0.68; p=0.02) and public hospitals (OR=0.59; p<0.001) were less likely to have 8 or more functions.

Table 1. Hospital Use of Advanced Electronic Health Record (EHR) Functions: Quality Improvement and Patient Engagement, 2015

<table>
<thead>
<tr>
<th>Performance Measurement to Support QI</th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monitor patient safety (e.g. adverse drug effects)</td>
<td>71.4%</td>
<td>28.6%</td>
</tr>
<tr>
<td>Support a continuous quality improvement process</td>
<td>71.1%</td>
<td>28.9%</td>
</tr>
</tbody>
</table>

Continued on next page
Hospital EHR adoption is widespread and many hospitals are using EHRs to support QI and patient engagement. However, this is not happening across all hospitals. Participation in payment reforms was associated with broad engagement in both QI and patient engagement functions while critical access hospitals were less likely to be broadly engaged in either domain. This is concerning because EHR use for QI and patient engagement is essential to improving hospital performance. Our results suggest that policy efforts are impacting adoption of these functions. Patient engagement functions included in the meaningful use program were among the most widely adopted functions, and participation in reform programs was associated with greater likelihood of adoption of both QI and patient engagement functions. Policymakers therefore have a set of levers – both direct funding for health IT and indirect incentives through promoting value-based payment and delivery models – that they can continue to use to drive greater adoption. Policymakers may need to consider specific actions to target safety net hospitals, which could include funding as well as technical assistance with implementation.

References

Phenotyping physiologic measurement of lung function in a large electronic health record using automated tools

Kathleen M. Akgün, MD, MS1,2, Keith Sigel3, Kei Cheung, PhD4, Farah Kidwai-Khan6,7, Alex K. Bryant5, Cynthia Brandt, MD, MPH4, Amy Justice, MD, PhD6,7, Kristina Crothers, MD8

1Pulmonary, Critical Care and Sleep Medicine, VA Connecticut Healthcare System; 2Department of Medicine, Section of Pulmonary, Critical Care and Sleep Medicine, Yale University School of Medicine, New Haven, Connecticut; 3Division of General Internal Medicine, Icahn School of Medicine at Mount Sinai, New York, New York; 4Department of Emergency Medicine, Yale University School of Medicine, New Haven, Connecticut; 5Department of Radiation Medicine and Applied Sciences, University of California San Diego School of Medicine, La Jolla, California; 6Department of Medicine, Section of General Medicine, VA Connecticut Healthcare System; 7Department of Medicine, Section of General Medicine, Yale University School of Medicine, New Haven, Connecticut; 8Pulmonary, Critical Care and Sleep Medicine Division, University of Washington, Seattle, Washington

Brief abstract
Chronic obstructive pulmonary disease (COPD) is associated with poor quality of life, hospitalization and mortality. The clinical phenotype of COPD is diagnosed by the finding of significant airflow obstruction on pulmonary function tests using Forced expiratory volume in one second (FEV1). FEV1 values are often difficult to identify in structured electronic health record (EHR) data. We developed an automated tool to improve ascertainment of FEV1 in EHR in the Veterans Aging Cohort Study (VACS).

Introduction
Forced expiratory volume in one second (FEV1) is a measure chronic obstructive pulmonary disease (COPD) severity that is routinely included in pulmonary function tests (PFTs) and is associated with decreased quality of life and increased risk for hospitalizations and mortality.1, 2 PFTs are frequently performed on vendor equipment that may not be linked to the primary electronic health record (EHR) for patients who undergo these studies. Unreliable linkage to the EHR makes recovery of PFT values such as FEV1 challenging for health services researchers. Further, these barriers to accessing PFT results hinder our ability to assess pulmonary function on the scale necessary to develop standard phenotypes of lung diseases outside of dedicated observational studies and clinical trials. Phenotyping physiologic lung function based upon direct measurement of trajectories of FEV1 is important for patients, clinicians and researchers to advance our understanding of the clinical burden of COPD as well as other lung diseases and to inform genetic discovery.

In the Veterans Affairs (VA) EHR, PFTs can be identified using Common Procedure Terminology (CPT) codes, entries into the Corporate Data Warehouse (CDW) files, scanned documents in the EHR, transcriptions into progress notes or through combinations of these approaches. CPT codes do not include quantifiable measurements from PFTs such as FEV1. While some VA PFT equipment directly uploads structured FEV1 measurements into the CDW (the VA central research data warehouse), software requirements prevent most sites from contributing FEV1 measurements to the CDW. To fill this gap, automatic tools can be used to identify FEV1 in the EHR but are not necessarily integrated in real time to be relevant to clinical setting.3 These tools can identify FEV1 values within structured or unstructured data and must be tailored to individual EHR systems. Development of clinically meaningful phenotypes for patients with COPD depends on developing an accurate approach to extracting FEV1 measurements from these different sources of EHR data. We developed and adapted an FEV1 extraction method for this purpose.
Methods
Using data from a nationally representative cohort of HIV-infected and uninfected Veterans (Veterans Aging Cohort Study [VACS]; 1996-2015), we identified patients who had PFTs within the VA system using relevant CPT codes (94010, 94150, 94060, 94726-9; 93720-93722, 94240, 94260, 94350, 94360, 94370, 94720, 94725). To extract FEV1 measurements, we then collected progress notes, which are stored as Text Integration Utilities (TIU) documents, in the CDW. TIU documents for these patients were processed in a structured query language (SQL) database (Microsoft SQL Server) supporting full-text (keyword) search. We processed TIU documents by developing an SQL-based algorithm for automated extraction of FEV1 entities (FEV-1, FEV1, FEV_1; negation for “fever”) and FEV1 values within 20 characters of the FEV1 entity from the TIU documents. We developed an SQL procedure to perform a full-text search to identify documents containing the keywords of “FEV1” or “FEV-1” or with spaces were retrieved. We also included keywords to determine if the FEV1 was measured following administration of bronchodilator medications. Within text identified by the full-text search, we then extracted numeric values within a plausible range and in a format consistent with the reporting of FEV1 values.

To evaluate the performance of this process, we generated a reference standard from a random subset of TIU documents that were identified as having more than one FEV1 entity in the text; a pulmonologist reviewed 128/250 (51%) of records with multiple FEV1 entities. The pulmonologist determined the presence of an FEV1 measurement closest to the date of the progress note reviewed and recorded that value. We compared results from the SQL tool FEV1 value extracted with CDW data for FEV1 to determine whether the SQL tool added substantially to the number of unique FEV1 measurements extracted.

Results
Using the SQL tool among patients identified using CPT codes, an FEV1 measurement was identified in 18,183 TIU documents among 5,958 unique patients. Compared with CDW data with quantifiable FEV1, the SQL tool identified an additional 3,849 patients with numeric values for FEV1 that can be used for phenotyping this population. Of note, these FEV1 values were otherwise unavailable for analyses.

In our reference standard subset (n=128 documents, 117 unique patients), the SQL tool had a positive predictive value of 89% for identifying records containing quantifiable FEV1 values. Extraction of these values yielded a correctly identified FEV1 measurement in 77% of cases with fair-good agreement (Cohen’s kappa=0.66; 95% confidence interval: 0.47-0.74). Cases where both reviewer and the tool identified quantifiable FEV1 values had excellent agreement (Spearman’s coefficient=0.99).

Conclusion
This demonstrates that FEV1 values can be extracted using SQL queries with excellent ascertainment and good accuracy from unstructured text-based documents generated by the VA EHR and increases the yield for the number of complete FEV1 values when added to structured data sources from CDW. Next steps will include characterization of data sources for patients who have CPT codes for PFTs but whose EHR does not includes FEV1 measurements. Future work is needed to further develop the phenotype, apply these rules to other VA populations, and refine/test the SQL tool for FEV1 extraction and ascertainment outside of the VA EHR.

References
Why predicting postprandial glucose using self-monitoring data is difficult

David Albers¹, PhD, Matthew Levine¹, BA, Andrew Stuart², PhD, Bruce Gluckman³, PhD, George Hripcsak¹, MD MS, and Lena Mamykina¹, PhD
¹Department of Biomedical Informatics, Columbia University, New York, NY
²Department of Computing and Mathematical Sciences, Caltech, Pasadena, CA
³Department of Engineering, Pennsylvania State University, College Park, PA

Introduction

Type 2 diabetes (T2DM) is a common and deadly disease if not treated to minimize its effects; it is costly ($245 billion a year in 2012) and prevalent, affecting over 8% of the US population [1]. While T2DM is a very complex disease, the treatment goal of T2DM is relatively simple: reduce hemoglobin A1c (HbA1c) below 6.0-6.5%. However, HbA1c is a slow moving laboratory variable that is roughly the sum effect of the previous three months of glucose dynamics—HbA1c is linearly related to mean glucose over the previous three months—but the current, quickly changing glucose levels must be reduced in order to reduce the slowly evolving HbA1c. The most typical approaches to reducing and stabilizing daily blood glucose levels is through medication and self-management. Typical self-management activities include introducing regular exercise, and altering one’s diet, among several others. For each of these activities, it is important for individuals to anticipate their impact on their blood glucose levels. For example, to make informed nutritional choices, individuals need an ability to anticipate the impact of different meals on their blood glucose. While there exist general guidelines for healthy eating in diabetes, due to the high individual differences in glycemic response to nutrition, each individual needs to make their own informed choices in regards to their meals based on their ability to anticipate the impact. However, previous research showed that making such predictions is challenging for both individuals with diabetes and their healthcare providers [2], [3]. In this research, we investigate novel computational modeling techniques for generating accurate real-time blood glucose forecasts for individuals with diabetes using data assimilation and machine learning approaches. This abstract focuses on reasons why forecasting postprandial glucose is difficult in general and is made more difficult by the presence of T2DM. By leveraging the source of the errors incurred by computational models when forecasting postprandial glucose, we will better understand why it is so difficult for individuals with T2DM to translate general nutrition guidelines into productive meal choices. Finally, we discuss what we believe will be the most accurate glucose forecasting methodology in the future.

Methods

We use nutrition and blood glucose data previously collected by individuals who participated in several ongoing T2DM intervention studies; two individuals included among our data do not have T2DM and have continuous glucose monitor data. The data were collected using a mobile application and constrained to realistic T2DM outpatient data collection, including sparse finger prick glucose measurements before and after meals and pictures of meals whose nutrition content was estimated by trained nutritionists. We use two forecasting engines to predict postprandial glucose. First we use a dynamic Gaussian process regression [7] on moving windows of various lengths of previous meals to train and forecast glucose. Second we use several data assimilation (DA) machines to integrate data with mechanistic models, including a dual unscented Kalman filter (UKF) [6] and other inverse-problem Bayesian methods [3]. The DAs rely on the ultradian endocrine model [4], can take blood glucose, insulin, and carbohydrates as input, and have 30 other physiologic parameters. Using the patient data and the computational engines, we generate personalized, real-time nutrition-based post-meal glucose forecasts. Mean squared error (MSE) of the model forecast compared to measurements is used to evaluate forecast errors [7].

Results

All of the forecasting engines personalized and converged to the individual patients with varying forecast accuracy. The DAs that simulate the endocrine system compare well with continuous glucose monitor data even when trained on sparse finger-prick data. The continuous DA output and the continuous glucose monitor data reveal the first reason why glycemic impact is difficult to predict: glucose oscillates between 60-160 mg/dl over the course of quarter hours and therefore the glycemic impact of a meal may be difficult to understand from a single random postprandial glucose measurement. Further complicating the situation, blood glucose forecasting error for some of the forecasting engines (e.g., the Gaussian process model) increases linearly with mean glucose statistically significantly (p=0.02). Moreover, there is sensitivity to the length of history used to estimate parameters for the forecasting engines.

Conclusion

The forecasting engines are able to forecast postprandial glucose measurements with varying success; DA based methods were the most accurate. Nevertheless, despite accurate pointwise postprandial glucose forecasts, a single postprandial glucose forecast or measurement may not accurately represent the glycemic impact of the nutrition choice because of the complexity of glucose dynamics. This explains the difficulty of generating such predictions for individuals with diabetes, even after reviewing their records collected with self-monitoring. Moreover, uncertainty of postprandial glucose forecasts increase with severity of T2DM, as...
characterized by mean blood glucose levels. Sensitivity to the length of history used to forecast postprandial glucose further complicates the forecasting task of individuals with T2DM. It is likely that an understanding of glycemic impact of nutrition choices will require computational help.

Acknowledgements We acknowledge NLM grant R01 LM06910, NIDDK grant 5R01DK090372, and the Robert Wood Johnson Foundation (Grant No. 73070).

References


Figure 1 The finger-prick and continuous blood glucose measurements and the data assimilation based continuous and on-measurement blood glucose forecasts for days X-Y for the ultradian model. As more measurements are accumulated, the data assimilator adjusts the parameters given the individual’s data allowing for personalization.

Figure 2 Mean squared error between glucose forecasts and measurements for the dynamic Gaussian process model with varying training history length. Individuals P1 and P2 have T2DM while individuals P3 and P4 do not have T2DM. The linear relationship between $\text{HbA1c}$ and forecasting error is increasing (1000) and statistically significant (p=0.02).
Introduction

As healthcare systems have shifted from paper-based documentation to Electronic Health Record (EHR) systems, issues with timely documentation and the burden of documentation on providers have developed. A 2011 American Medical Informatics Association (AMIA) Policy Meeting report identified challenges to fully realizing the potential benefits of EHRs as described by the National Academy of Medicine, including “identifying how electronic documentation of key data can best be integrated into clinical workflow” and “clarifying the roles of care team members, including the patient, in creating and accessing the electronic record”. As a result, healthcare systems have invested in information technology tools that support both providers and other clinician documentation, including specifically nursing documentation but the patient’s role in documentation remains limited in most cases. The goal of this study was to evaluate the effects of an innovative pain management interactive patient care (IPC) tool supporting automatic patient/parent pain documentation in the EHR and automated nurse notification and its effect on nursing pain reassessment documentation rates in an inpatient pediatric hospital.

Methods

The interactive pain management interface implemented at the University of Minnesota Masonic Children’s Hospital is designed to allow for bidirectional communication between the television-based IPC tool, the EHR, the medication dispensing system, and the nursing call system (Figure 1). The medication dispensing system triggers a time-based patient pain rating scale after a medication is dispensed. When patients or parents respond, the system then triggers the nurse’s phone to serve as a reminder to perform a hands-on pain reassessment and documentation.

This study involved a retrospective analysis of inpatient records covering an eleven-month period post interface implementation from January to November 2015. Inpatient records with at least one narcotic medication administered from a formulary developed by the hospital’s pharmacy team for setting the automatic triggers were included and extracted from the University of Minnesota research Clinical Data Repository. Also included in the extracted dataset were values from 14 unique flowsheet rows, representing nursing pain reassessment documentation observations, as well as values from one unique flowsheet row, representing patient/parent pain responses automatic documentation (Figure 2). Records were summarized by frequencies and percentages based on the presence/absence of nurse documentation values and patient/parent pain-rating responses. A chi-square test was used to compare nurse documentation rates to other categorical factors. P-values less than 0.05 were considered statistically significant.
Results
A total of 27,224 records had at least one narcotic administration during the study time period. The total number of patients/parents pain responses through using the IPC tool was n=1,767 with usage rates of (6.5%). On the other hand, nursing pain reassessment documentation values were found in 14,534 (53.4%) of records. Of the records where the patient/parent did provide a pain rating response with the IPC tool, 63.4% (n=1,121) had a nursing documentation value and 36.6% (n=646) did not. Whereas 47.3% (n=12,044) of patients who did not report their pain with the IPC tool had no follow up nursing documentation, this decreased to 36.6% (n=646) for the patients/parents who used the IPC tool, a relative difference of 10.7% (p<0.001) (Figure 3).

Discussion
The study findings illustrate a number of benefits of implementing IPC tools supporting automatic timely documentation through an inpatient pediatric use case. Our implementation of a unique customized IPC tool at the University of Minnesota Masonic Children’s Hospital, demonstrated relatively low use by patients/parents, but also found that among the group of patients/parents that did use the tool, their use was associated with a statistically and clinically significant increase (10.7%) in nursing pain reassessment documentation rates. Increasing the awareness of patient/parents on the important contribution they are able to make by using this tool may increase utilization rates among patient/parents, and potentially further improve nursing documentation rates. Future research focuses on exploring the reliability of the patient/parent response values in comparison with the nurse pain reassessment documentation values, to explore with hospital leadership the use of automatic patient/parent initiated pain documentation in lieu of nursing reported documentation compliance, and understanding barriers and facilitators of adoption of the tool by end users, including both patient/parents and nurses.

Acknowledgements
This research was supported by the National Center for Advancing Translational Sciences of the National Institutes of Health Award Number UL1TR000114, the University of Minnesota Clinical and Translational Science Award (8UL1TR000114-02), and the University of Minnesota Physicians and Fairview Health Services.

References
Quick Sequential [Sepsis-related] Organ Failure Assessment (qSOFA) and St. John Sepsis Surveillance Agent to Detect Patients At-Risk of Sepsis: An Observational Cohort Study

Robert C. Amland, PhD1, Bharat B. Sutariya, MD1
1Cerner Corporation, 2800 Rockcreek Parkway, Kansas City, MO 64117; USA.

Learning Objective
After participating in this session, the learner should be able to evaluate sepsis surveillance systems involving how to measure time to event and compare different alert system clinimetric performance on detection of at-risk patients and outcomes.

Abstract
The 2016 Sepsis-3 guidelines included a new clinical surveillance algorithm to identify patients at-risk of sepsis. A multiple center observational study design was used to compare the utility of Quick Sequential [Sepsis-related] Organ Failure Assessment (qSOFA) to the St. John Sepsis Surveillance Agent, which has been implemented in 550 hospitals in the USA. The St. John Sepsis Surveillance Agent, when compared to qSOFA, activated earlier and was more accurate in predicting patient outcomes.

Introduction
The 2016 consensus guideline on sepsis (Sepsis-3) includes a new sepsis surveillance alert definition termed Quick Sequential [Sepsis-related] Organ Failure Assessment (qSOFA). The purpose of qSOFA is to identify non-ICU patients currently under suspicion of infection, and potentially deteriorating into a sepsis complication. The qSOFA algorithm differentiates from earlier consensus guidelines, such that it incorporates latency and narrows clinical risk factors to tachypnea, hypotension, and altered mentation within the immediate 24 hours after onset of infection. In contrast, the St. John Sepsis Surveillance Agent, developed by Cerner Corporation in 2010, contains two alert mechanisms for sepsis and severe sepsis based on the earlier consensus definitions of SIRS, sepsis, and severe sepsis/shock and provides continuous 24/7 hospital-wide monitoring. The system has now been implemented in 550 acute care facilities in the USA. The objective of the study was to compare the utility of qSOFA to St. John Sepsis Surveillance Agent among patients with suspected infection.

Methods
The study employed a multiple center observational cohort study design to examine the comparative effectiveness of qSOFA and St. John Sepsis Surveillance Agent on the primary outcome of in-hospital mortality, and the secondary outcome of the composite of death or ICU admission. The St. John Sepsis Surveillance Agent applied a binary alarm paradigm with two alert definitions: (1) indications of SIRS (proxy for sepsis) and (2) indications of Sepsis (proxy for severe sepsis). Each alert, unless suppressed by a localized rule, delivered a real-time notification to a provider which included specific clinical criteria responsible for activation. Subsequently, we developed the analytic dataset with a qSOFA flag and its trigger time stamp, and a suspected infection clinical event with its onset time stamp. These two events were joined by examining their respective time stamps to create a qSOFA clinical event. Consistent with Sepsis-3, the qSOFA clinical event was suppressed if the qSOFA triggered before onset of suspected infection, after 24 hours from the onset of suspected infection, or after ICU admission. For each patient, we studied their first alert for each of the two surveillance models. A baseline risk model included age, male sex, recent discharge from hospital, onset of infection in the ED, and admitting facility. The study population comprised 17,044 patients between January and March 2016, of which 35% (n = 5,992) patients met inclusion criteria for suspected infection. For the primary analysis, receiver operator characteristic curves were constructed for patient outcomes using qSOFA and the St. John Sepsis Surveillance Agent, and the areas under the curve were compared against a baseline risk model. Time-to-event clinical process modeling was also applied.

Results
The St. John Sepsis Surveillance Agent (n = 2,228; 37%) identified substantially more patients at-risk of sepsis than qSOFA (n = 649; 11%). This increased sensitivity corresponded to the St. John Sepsis Surveillance Agent having screened-in 32 patients per 1,000 patient days compared to the qSOFA rate of 9 patients per 1,000 days. Approximately 1 in 4 patients screened-in by the St. John Sepsis Surveillance Agent occurred before onset of
infection, while the remainder of patients were mostly detected within one hour after onset; in contrast, most patients with qSOFA activated several hours after onset of infection. Model discrimination for in-hospital mortality was higher for St. John Sepsis Surveillance Agent (AUC = .74 [95% CI = .71 to .77]) compared to qSOFA (AUC = .69 [95% CI = .66 to .73]). Holding specificity constant, the St. John Sepsis Surveillance Agent and qSOFA correctly reclassified, in absolute terms, 13% and 5% patients, respectively, from the baseline risk model. This reclassification of risk translates to a relative improvement in sensitivity ∆↑ 23% for St. John Sepsis Surveillance Agent and ∆↑ 9% for qSOFA, respectively. Applying the Kaplan-Meier survival framework, Figure 3 illustrates the number of hours elapsed after triggering the St. John Sepsis Surveillance Agent or a qSOFA clinical event to the cumulative composite outcome, trimmed at 96 hours after the alert. A difference in estimated survival curves existed between the three patient cohorts (P = .026). Time-to-event measures with cases censored at 96 hours after detection showed the cohort with the St. John Sepsis Surveillance Agent “Sepsis alert” (n = 306 of 930, 33%), median 5.73 (IQR = 4.88 to 6.58) hours after detection, and approximately 1 in 17 (n = 17 of 306, 6%) patients were censored; compared to the cohort with a qSOFA clinical event (n = 163 of 649, 25%), median 5.30 (IQR = 3.46 to 7.14) hours after detection, and had twice the rate of censored cases (n = 19 of 163, 12%); and last, the cohort with a St. John Sepsis Surveillance Agent “SIRS alert” (n = 292 of 1298, 23%), median 6.35 (IQR = 4.68 to 8.02) hours after detection, and had a similar rate of censored cases (n = 37 of 292, 13%) as qSOFA.

Discussion
Comparing earlier consensus definitions on sepsis with the 2016 Sepsis-3 qSOFA, we found the earlier definitions more robust for sepsis surveillance applications. In contrast to qSOFA, the St. John Sepsis Surveillance Agent screened-in more high-risk patients, activated much earlier in their infectious process, and was more accurate in predicting adverse outcomes. Considering consequences of missing patients or the introduction of avoidable delay, qSOFA fell far behind on these objectives while the St. John Sepsis Surveillance Agent was promising.

References
Identifying and Predicting Falls among Elderly Residents of Baltimore City Using Hospital Discharge Summaries and Health Information Exchange Data

Laura Anzaldi1, Elyse Lasser1, Joshua Sharfstein1, Hadi Kharrazi1
1 Johns Hopkins School of Public Health, Baltimore, MD, USA

Abstract
This study evaluates the use of hospital discharge summaries along with health information exchange data to identify and predict falls among elderly in Baltimore City.

Introduction
CDC estimates the direct cost of falls to be $30 billion each year (1). Baltimore Falls Reduction Initiative Engaging Neighborhoods & Data (B’FRIEND) is a public-private partnership in Baltimore City based on innovative use of health data to decrease the rate of falls leading to an emergency room (ER) or hospital admission among elderly. The overall aims of the project include:

- Develop and validate a case identification methodology;
- Develop and validate a fall’s risk prediction model; and,
- Evaluate the fall risk score and disseminate results (not reported in this abstract)

Methods and Study Design
This study is a retrospective analysis of hospital discharge summaries of 327,727 Baltimore residents (2014) provided by Maryland’s Health Services Cost Review Commission (HSCRC) and Maryland’s Health Information Exchange (HIE; aka CRISP). Phenotyping criteria of falls were translated into a scalable query to identify falls among elderly leading to an ER/hospital admission. The study denominator consisted of 29,301 Baltimore residents, age 65+, 61% female, and, with an unintentional fall resulting into an ER admission or hospitalization in 2014. A semi-systematic review of the literature identified potential causes and risk factors of increased falls among elderly. Logistic regression, with L2 regularization, was used in R (GLMULTI package) to assess the explanatory power of the potential risk factors in predicting falls leading to an ER/hospital admission. Model development used a 100-fold 75%/25%-split training and validation method. AIC was used for model selection.

Findings
The case identification process identified 13% of elderly with at least one high-risk fall in Baltimore City in 2014 (40% higher than the national average). Figure 1 shows the rates of fall by neighborhood derived from real-time HIE data in Baltimore City:

![Figure 1. Prevalence of falls among elderly in Baltimore City zip codes (based on HIE cross-sectional data)](image)

The systematic review of literature identified 28 variables as predictors to high-risk falls, which were limited to 19 variables after applying conceptual and algorithmic feature reduction-methods. The logistic regression identified
history of a prior fall to have the highest predictive value (OR 6.02). A history of fracture, substance abuse, Parkinson, kyphoscoliosis, depression, mental illness, age, and sex (female) had a positive predictive value in explaining high-risk falls, while Charlson index, vision impairment, obesity, CVD, hypertension, cancer, low back pain, joint trauma, and lower joint surgery had a negative predictive value (all statistically significant at alpha .05). The model’s AUC was .703 with an AIC of 15495 (Table 1).

| Predictors               | Estimate | Std. error | z value | Pr(>|z|) | Significance | OR 2.50% | 97.50% |
|-------------------------|----------|------------|---------|---------|--------------|----------|--------|
| History of fall         | 1.795    | 0.074      | 24.113  | <2e-16  | ***          | 6.02     | 5.20   | 6.97   |
| Fracture                | 0.604    | 0.082      | 5.821   | 5.85E-09| ***          | 1.83     | 1.49   | 2.24   |
| Substance Abuse         | 0.520    | 0.082      | 6.364   | 1.96E-10| ***          | 1.68     | 1.43   | 1.97   |
| Parkinson               | 0.337    | 0.178      | 1.895   | 0.058056| **           | 1.40     | 0.98   | 1.97   |
| Kyphoscoliosis          | 0.322    | 0.153      | 2.102   | 0.035519| *            | 1.38     | 1.01   | 1.85   |
| Sex (female)            | 0.173    | 0.046      | 3.736   | 0.000187| ***          | 1.19     | 1.09   | 1.30   |
| Depression              | 0.146    | 0.068      | 2.141   | 0.032238| *            | 1.16     | 1.01   | 1.32   |
| Mental Illness          | 0.128    | 0.065      | 1.980   | 0.047652| *            | 1.14     | 1.00   | 1.29   |
| Age                     | 0.038    | 0.003      | 14.895  | <2e-16  | ***          | 1.04     | 1.03   | 1.04   |
| Charlson Index          | -0.053   | 0.009      | -5.711  | 1.12E-08| ***          | 0.95     | 0.93   | 0.97   |
| Vision                  | -0.211   | 0.057      | -3.689  | 0.000225| ***          | 0.81     | 0.72   | 0.91   |
| Obesity                 | -0.251   | 0.076      | -3.311  | 0.00931 | ***          | 0.78     | 0.67   | 0.90   |
| Cardiovascular Disease  | -0.313   | 0.050      | -6.301  | 2.95E-10| ***          | 0.73     | 0.66   | 0.81   |
| Lower Urinary Tract Symptoms | -0.345 | 0.074      | -4.656  | 3.23E-06| ***          | 0.71     | 0.61   | 0.82   |
| Hypertension            | -0.357   | 0.050      | -7.080  | 1.44E-12| ***          | 0.70     | 0.63   | 0.77   |
| Cancer                  | -0.441   | 0.081      | -5.418  | 6.02E-08| ***          | 0.64     | 0.55   | 0.75   |
| Lower Back Pain         | -0.495   | 0.067      | -7.368  | 1.73E-13| ***          | 0.61     | 0.53   | 0.69   |
| Joint Trauma            | -0.526   | 0.197      | -2.674  | 0.007487| **           | 0.59     | 0.39   | 0.85   |
| Lower Extremity Joint Surgery | -1.069 | 0.182      | -5.870  | 4.36E-09| ***          | 0.34     | 0.24   | 0.48   |
| (Intercept)             | -4.372   | 0.197      | -22.249 | <2e-16  | ***          | 0.01     | 0.01   | 0.02   |

Significance codes: 0 '*' 0.001 '*' 0.01 '*' 0.05 '*' 0.1 '*' 1

Std.: Standard; OR: Odds Ratio

Conclusion

B’FRIEND results will guide, trigger and evaluate a variety of clinical and environmental interventions that will be tailored for the subpopulations of similar trajectories. For example, if a community-level trajectory of falls risk is being detected under the B’FRIEND project, housing resources can be deployed to reduce community risks associated with the predicted increased rates of falls. On an individual basis, trajectory alerts will be sent to patients’ primary care physicians via HIE connectivity for additional follow up.

We acknowledge the support of Darcy Phelan-Emrick (Chief Epidemiologist) and Mike Fried (CIO) of Baltimore City Health Department for this project.

Funding Source

RWJF DASH and JHU inHealth

Learning Objectives:

1. Explain the use of hospital discharge summaries and health information exchange data in detecting and predicting high-risk falls among elderly in Baltimore City

2. Define potential sources of data to improve the prediction of high-risk falls among elderly including EHR and geographic/social data

References

Differing patterns of satisfaction and perception among clinical and non-clinical users following replacement of a legacy EHR system

Nate C. Apathy1; Joshua R. Vest, MPH PhD1; Nir Menachemi, MPH PhD1; John W. Putz, PhD2; Justin G. Morea2, DO; Christopher A. Harle, PhD1

1Richard M. Fairbanks School of Public Health, Indiana University, Indianapolis, IN; 2Eskenazi Health, Indianapolis, IN

Introduction

Implementation and adoption of new health information technology (HIT) is complex, time-consuming, difficult, and expensive for health care organizations.1 Despite widespread use of electronic health records (EHRs), organizations will continue to implement new HIT as they upgrade and replace existing systems in response to pressures to improve care quality and reduce costs. For example, the new federal Quality Payment Program builds on EHR Meaningful Use measures and establishes new measures related to HIT use.2 Therefore, it is important to understand the effects that HIT system changes have on health care organizations’ providers and clinical staff.

Recent literature has primarily focused on physicians and other clinical users of new HIT and their perceptions of implementation and HIT system utility.1,3 Less is known about the impact of organization-wide HIT implementation on non-clinical users. These roles comprise important segments of HIT users, especially given the increasing importance of quality reporting, population health, and multidisciplinary care.4 Yet, HIT implementation strategies often de-prioritize non-clinical HIT users.5 Therefore, to examine how HIT changes affect different types of users, this study presents a longitudinal analysis of changes in perceptions and satisfaction among clinical and non-clinical EHR users in a large health system before and after implementation of a new enterprise EHR system.

Study Design and Methods

We conducted a longitudinal, prospective survey of 3,672 clinical and non-clinical EHR users at large public hospital system consisting of a 315-bed hospital and 10 affiliated clinics providing comprehensive primary and specialty services in Indianapolis, IN. In October 2016, the system switched from a homegrown legacy EHR to a commercial vendor’s enterprise EHR system. We developed a survey instrument modeled after a similar study3 to measure user satisfaction and perceptions of the EHR before and after the transition. Providers and staff identified as EHR users received email invitations one month prior to the new system go-live date and one and three months following the go-live. In total, 607 participants responded to all three survey waves (16.5% of those invited to the initial wave). For analysis, respondents were divided into clinical and non-clinical subgroups. The clinical user group included physicians, physician assistants, nurses, medical assistants, nurse practitioners, and other clinical care providers. The non-clinical group included all other roles unrelated to direct patient care, such as registration clerks, and analysts.

The survey included eight questions relevant to both clinical and non-clinical EHR system users and their workflows. These questions were identical for all respondents, regardless of role. We asked respondents about their access to information from other institutions (1 question), new work added or work eliminated by the EHR (2q), availability of support (1q), adequacy of training (1q), and overall satisfaction with both the EHR (1q) and one’s job (2q). All survey questions included a “not applicable option” and used five-point response scales ranging from “strongly disagree” to “strongly agree.” We dichotomized responses into “Agree” (agree or strongly agree) and “Disagree/Neutral” (neutral, disagree, or strongly disagree) and analyzed the proportion of respondents in the “Agree” category at each survey wave. Chi-square analysis was used to determine the significance of differences between the clinical and non-clinical subgroups over the three waves.

Figure 1. Percentage of respondents agreeing or strongly agreeing with items after 3 months. For added new work, more agreement is worse.
Principal Findings

Across all respondents, six of the eight items showed significant (p < 0.001) increases in the percentage of respondents who either agreed or strongly agreed after three months (Figure 1). Five of these six were changes favorable to the newly implemented EHR. The largest favorable increases were in the EHR’s elimination of work (11.1% agreeing pre-transition to 50.8% after three months), ease of access to information from other providers (30.6% to 58.0%), and availability of help (39.8% to 61.8%). The largest single change was in the EHR’s creation of new work (20.9% to 62.5%). No significant changes occurred across all participants or subgroups in overall job satisfaction, which remained above 75% either satisfied or highly satisfied across all survey waves.

The non-clinical user subgroup had 231 respondents (38%). The clinical subgroup had 376 respondents, [201 nurses/medical assistants (33%); 45 physicians/nurse practitioners/physician assistants (7%); 130 other patient care (21%)]. Compared to clinical EHR users, non-clinical users showed significant differences in patterns of satisfaction over time for several items: access to information from other providers, adequacy of training and support, and creation of new work. At baseline, when surveyed about the legacy EHR system, non-clinicians were more likely than clinicians to agree with ease of access (p = 0.016) and less likely to agree with adequacy of training (p = 0.018) and creation of new work (p < 0.001). However, there were no between-group differences in satisfaction after three months with the new system.

Conclusion

We found evidence that user perception and satisfaction around an EHR transition differed for non-clinical users compared to clinical users. Specifically, in the areas of pre-transition access to information from other providers, training adequacy, and the EHR’s creation of new work, non-clinical EHR users indicated differences before the transition that converged with clinical users’ perceptions and attitudes three months after the transition. Understanding how clinical and non-clinical EHR users differ can inform healthcare organizations as they manage HIT system changes. Especially in the area of pre-transition training, our results emphasize that non-clinical users may perceive lower adequacy of training, perhaps due to different levels of focus on training for different roles.

References

Novel Targeting of Clinical Decision Support: Utilizing Machine Learning To Improve Provider Acceptance – Repeat Imaging as an Example
Wasim Al Assad, MD, MSce1, Ivan Ip, MD, MS, MPH1, Li Zhou, MD, PhD1, Ramin Khorasani, MD, MPH1
1Harvard Medical School & Brigham and Women’s Hospital, Boston, MA 02115

Abstract
Clinical decision support (CDS) suffers low provider acceptance which potentially could be improved with machine learning. We extracted repeat imaging decision support data between 2007 and 2014, and built multiple classifiers to predict the acceptance of repeat imaging alerts. Our best classifier (Random Forest) achieved an accuracy of 75.1%. Machine learning could be utilized to improve provider acceptance of CDS alerts by targeting instances with high likelihood of acceptance.

Introduction
Clinical decision support (CDS) remains a challenging component of health informatics practice, continuing to suffer from a high override rate and low acceptance. Some CDS interventions have an acceptance rate of less than ten percent; others, rates of less than five percent1.

Repeat imaging, diagnostic testing that occurs within a specific time interval after an initial imaging study, is of concern on the national level. Imaging utilization has increased over the last few decades2 due to numerous factors. Increased imaging utilization is a contributor to increased healthcare spending that has outpaced the growth of the national GDP in the last decade3.

In prior work, we developed and implemented a CDS intervention with the goal of reducing unnecessary repeat imaging. The intervention consisted of providing a reminder of an existent relevant imaging study within the ninety days prior to the time of order entry, and asking providers to choose a clinical justification reason if they overrode the alert. This intervention reduced repeat imaging studies by nearly 7%. However, providers still overrode repeat imaging alerts over 90% of the time4.

In this paper, we utilize a data-mining approach to analyze CDS interactions. We employ machine learning algorithms in an attempt to improve repeat imaging CDS acceptance by targeting scenarios when acceptance is predicted by our algorithms. We envision that utilizing such an approach will reduce triggering of CDS alerts when acceptance is highly unlikely, thus improving overall acceptance.

This retrospective, Institutional Review Board-approved study was performed in a tertiary academic medical center including a 793-bed acute care hospital, a network of ambulatory clinics, and an emergency department with an annual volume of ~60,000 visits.

Between 2007 and 2014, alerts were triggered for any imaging orders with a prior exam of the same modality and body part performed within the previous 90 days. Alerts were shown to providers through the computerized provider order entry module of the electronic health record system (Percipio, Medicalis5). For each repeat imaging order, the system recorded provider interaction details including ordering provider, user type (physician, mid-levels, support staff), patient age and gender, ordering site, practice setting, and imaging modality. These data were stored in a database (SQL Server, Microsoft, Seattle, WA). We processed this data in SQL Server and built our machine learning models using Sickit-learn6.

During the time period of the active alert, 551,967 imaging orders triggered repeat imaging CDS alerts. Only 39,131 (7.6%) of CDS alerts were accepted with cancellation of the repeat imaging order; the remaining 512,836 alerts were ignored (overridden).

We divided the data into training (70%) and testing (30%) sets. Since the override dataset was nearly 13 times the acceptance dataset, we oversampled the acceptance dataset and matched it to a similar size of the override dataset. We used the following features, as they were available through our data collection system: patient age, patient gender, practice setting, imaging modality, number of similar orders in the prior 90 days, time elapsed since last order, average time elapsed for similar orders in the prior 90 days, user’s imaging CDS acceptance rate, provider medical school attendance (MD or not), and count of all provider imaging orders in our dataset.

We constructed three commonly-used classifiers, namely decision trees (DT), Random Forest (RF), and logistic regression7. We compared them using standard machine learning metrics of accuracy, error, sensitivity, and specificity7. Table 1 compares the results of the four classifiers.
Our best classifier (Random Forest) achieved an accuracy (total true predictions divided by total predictions) of 75.1%, sensitivity (true acceptance predictions out of all acceptance cases in the dataset) of ~70.9%, and a specificity (true rejection predictions out of all rejection cases of the dataset) of 79.3%. The positive predictive value (PPV) of our best classifier (RF) is 22.1%—a significant improvement over the current acceptance rate of nearly 7% for repeat imaging alerts.

<table>
<thead>
<tr>
<th></th>
<th>Decision Trees</th>
<th>Random Forest</th>
<th>Logistic Regression</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>72.2</td>
<td>75.1</td>
<td>70.6</td>
</tr>
<tr>
<td>Error</td>
<td>27.8</td>
<td>24.9</td>
<td>29.4</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>71.4</td>
<td>70.9</td>
<td>64.7</td>
</tr>
<tr>
<td>Specificity</td>
<td>73.0</td>
<td>79.3</td>
<td>76.5</td>
</tr>
</tbody>
</table>

**Conclusion**

In this study we built a classifier to predict repeat imaging CDS acceptance. Limitations include the low number of features in our classifier due to the limitation of our data collection systems, and the lack of clinical patient information involved in the classifier. Additionally, our assessment is based on clinician’s current acceptance of CDS alerts and thus the appropriateness of imaging utilization in cases of alert overrides has not yet been fully studied. Despite the limited features included in our classifier, the noisy nature of repeat imaging (i.e., repeat imaging occurs in the hospital for large set of reasons), and the complexity of predicting response to such a generic intervention; our classifier achieved results with an accuracy of over 70 percent. This indicates that our approach has potential application to improve CDS acceptance.

Implementing this classifier could theoretically improve acceptance of repeat imaging CDS alerts by nearly 3-folds (from 7.6% to 22.1%) by reducing a significant portion of alerts that are likely to be ignored. Future studies would be needed to test this classifier in a clinical implementation to assess its impact on repeat imaging CDS alert acceptance rate.

**References**


22
Clinical Trial Eligibility Criteria Fail to Meet Burden of Generalizability
Amelia J. Averitt, MPH1; Chunhua Weng, PhD1; Adler J. Perotte, MD MA1
1 Columbia University, Department of Biomedical Informatics. New York, NY

Introduction
Evidence-Based Medicine (EBM) requires medical practitioners to consider empirical and experimental evidence when treating their patients. EBM encourages practitioners to seek the most reputable evidence according to a hierarchy, in which randomized controlled trials (RCTs) are regarded as the gold standard source. While RCTs can offer precise and valid insights into the efficacy and safety, the results are often criticized for their poor generalizability and may therefore be unsuitable to serve as evidence for the practice of medicine.1,2 This failure of an RCT to represent an indicated population may be the result of cohort selection for a single condition of interest that is further narrowed by a breadth of eligibility criteria. This final cohort population may stand in stark comparison to the intended target population, which could have multiple problems and comorbidities.3,4 Observational data sources, such as the electronic health record (EHR), are regarded as more representative of the target patient population; but they cannot guarantee distributional similarity on confounding variables.1 We present empirical evidence that the construction of an observational cohort according to RCT eligibility specifications encourages convergence towards the true RCT effect estimate, but that these observational estimates are marked by residual confounding, which suggests that factors beyond eligibility criteria are necessary to determine the applicability of RCT results as clinical evidence.

Methods
Recent research investigates the representativeness of clinical trial participants for the approximation of a target population, however there exists a question of how effect estimates vary as a function of eligibility criteria.5 To address this knowledge gap, we propose two investigations, each of which will employ a unique RCT as a gold standard comparator. The first investigation explores the impact of eligibility criteria on the effect estimate (Study 1), and the second investigation examines potential sources of residual bias in effect estimates (Study 2).

Study 1: We hypothesized that each addition of an operationalized inclusion and exclusion criteria will increase distributional similarity between the observational and experimental cohorts, and will bring the effect estimate from the observational cohort closer to the effect estimates of published RCT outcomes. Inclusion and exclusion criteria extracted from published RCT protocols and literature were incrementally added to an observational cohort that was constructed according to the baseline study population of the RCT. As an example, we evaluated the occurrence of hypoglycemia associated with sitagliptin compared to glimepiride in elderly patients (65-80 years of age) with Type 2 Diabetes Mellitus (T2DM) and inadequate glycemic control.6 The baseline cohort of patients with these requirements was identified in the NewYork-Presbyterian Hospital data (~5.4 million patients). This baseline cohort was queried multiple times, with each query being subject to the addition of new eligibility criteria (See Table 1). With each incremental criterion, the unadjusted relative risk (RR) of hypoglycemia was calculated and compared to the RR that was calculated for this endpoint at the RCT’s close.

Figure 1 and Table 1: Observational RR vs RCT RR under varying eligibility criteria

<table>
<thead>
<tr>
<th>Criteria Eligibility Criteria</th>
<th>RR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 No eligibility criteria</td>
<td>0.675</td>
</tr>
<tr>
<td>2 Criteria #1 + No HIV</td>
<td>0.676</td>
</tr>
<tr>
<td>3 Criteria #2 + No T1DM</td>
<td>0.662</td>
</tr>
<tr>
<td>4 Criteria #3 + No Surgery</td>
<td>0.649</td>
</tr>
<tr>
<td>5 Criteria #4 + No CVD</td>
<td>0.633</td>
</tr>
<tr>
<td>6 Criteria #5 + No Liver Disease</td>
<td>0.632</td>
</tr>
<tr>
<td>7 Criteria #6 + No PVD</td>
<td>0.621</td>
</tr>
<tr>
<td>8 Criteria #7 + No High TGs</td>
<td>0.621</td>
</tr>
<tr>
<td>9 Criteria #8 + No Insulin/GLP-1</td>
<td>0.621</td>
</tr>
<tr>
<td>10 Criteria #9 + No PPAR</td>
<td>0.604</td>
</tr>
<tr>
<td>11 Criteria #10 + No DPP-4</td>
<td>0.590</td>
</tr>
<tr>
<td>12 Criteria #11 + No Cancer</td>
<td>0.595</td>
</tr>
<tr>
<td>13 Criteria #12 + No Heme. Dis.</td>
<td>0.516</td>
</tr>
<tr>
<td>14 Criteria #13 + No GFR &lt; 35</td>
<td>0.534</td>
</tr>
<tr>
<td>15 Criteria #14 + No Hist of Subst</td>
<td>0.498</td>
</tr>
</tbody>
</table>

Study 2: To investigate the etiology of the residual bias to the effect estimate, observational cohorts that were curated according to eligibility criteria were compared to RCTs. Observational cohorts were created using the Observational Health Data Science and Informatics (OHDSI) ATLAS tools. The trial indication identified a core set of patients to which inclusion and exclusion criteria were applied to appropriately narrow the cohort. Subjects who
remained eligible after this pruning stage comprised the observational cohort of interest, which was compared to the RCT’s participant features that are reported in the baseline features table.

Results
Study 1: The unadjusted RR, calculated from observational data without any eligibility criteria, $RR_{obs} = 0.675$. After application of all eligibility criteria, the unadjusted $RR_{obs} = 0.498$. Figure 1 demonstrates the convergence of the effect estimate towards the RCT’s truth ($RR_{RCT} = 0.178$), which empirically confirms our hypothesis. Our research suggests that the incremental addition of RCT eligibility criteria encourages the distribution of meaningful covariates in the observational cohort to mirror that of the RCT, therefore reducing bias, and moving the observational effect estimate closer to the clinical trial estimate.

Study 2: Table 2 presents select features of this analysis using the landmark RENAAL Trial, which compares the effect of losartan and placebo on diabetic nephropathy. The data demonstrates that, despite curation of an observational cohort from NewYork-Presbyterian, statistically significant differences from the RCT exist among key features.

Table 2: RENAAL Trial Observational Cohort with Eligibility Criteria vs RCT

<table>
<thead>
<tr>
<th>Feature</th>
<th>NewYork-Presbyterian Hospital</th>
<th>The RENAAL Trial</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Indic. $^a$</td>
<td>Indic. + Elig. Criteria $^a$</td>
</tr>
<tr>
<td></td>
<td>n=3818</td>
<td>n=72</td>
</tr>
<tr>
<td>Age</td>
<td>63.61</td>
<td>59.27</td>
</tr>
<tr>
<td>Male</td>
<td>40.78%</td>
<td>38.89%</td>
</tr>
<tr>
<td>Female</td>
<td>59.22%</td>
<td>61.11%</td>
</tr>
<tr>
<td>BMI</td>
<td>30.47</td>
<td>34.00</td>
</tr>
<tr>
<td>CR mg/dL</td>
<td>1.89</td>
<td>2.46</td>
</tr>
<tr>
<td>Angina</td>
<td>14.14%</td>
<td>5.56%</td>
</tr>
<tr>
<td>MI</td>
<td>2.02%</td>
<td>2.78%</td>
</tr>
<tr>
<td>Amputation</td>
<td>1.60%</td>
<td>0.00%</td>
</tr>
<tr>
<td>Lipid Disorder</td>
<td>58.15%</td>
<td>43.06%</td>
</tr>
<tr>
<td>Neuropathy</td>
<td>19.83%</td>
<td>11.11%</td>
</tr>
<tr>
<td>Retinopathy</td>
<td>5.40%</td>
<td>4.17%</td>
</tr>
<tr>
<td>Smoker</td>
<td>6.47%</td>
<td>2.78%</td>
</tr>
<tr>
<td>Triglycerides</td>
<td>154.29</td>
<td>156.21</td>
</tr>
<tr>
<td>Hemoglobin</td>
<td>11.35</td>
<td>11.92</td>
</tr>
<tr>
<td>HbA1c (%)</td>
<td>7.60</td>
<td>8.24</td>
</tr>
</tbody>
</table>

Discussion
This research and the brief examples provided above empirically demonstrate that the distribution of the clinical trial participants is different from that of the target patient population, even when subject to all eligibility criteria. Experimental trial participants are not only an inherently poor representation of the target population, but this research also suggests that the application of eligibility criteria may reveal new hidden bias. Based on these results, careful consideration beyond eligibility criteria is necessary to determine whether results of a given RCT are an appropriate source of evidence when considering the care of a given patient. These, and other related methods, will support a framework for causal inference and the practice of EBM. Funding: NIH BD2K; JANSRD CU15-2317

References
USING COMPUTER AGENTS TO EXPLAIN CLINICAL TEST RESULTS

Renato F. L. Azevedo¹, Kuangxiao Gu¹, Yang Zhang¹, Victor Sadauskas², Tarek Sakakini¹, Dan Morrow, PhD¹, Mark Hasegawa-Johnson, PhD¹, Thomas S. Huang, PhD¹, Suma Bhat, PhD¹, Ann Willemsen-Dunlap²,³, Donald J. Halpin²,³, James Graumlich, MD², William Schuh⁴

¹UIUC, IL; ²UICOMP, IL; ³Jump Sim Center, IL; ⁴Carle Foundation Hospital, IL

Introduction

Health information technology and the Internet have tremendous potential to transform health care¹. EHR portals in particular can improve patient care by providing more access to health information¹. However, they are underutilized by older adults, in part because they are hard to use and the information is often hard to understand². Because portals function more as repositories than collaborative tools, they do not engage patients and may not facilitate self-care. Self-managing illness requires understanding and using a wide range of health information. For example, to understand cholesterol test results, people need to interpret the numbers (e.g., HDL) in terms of risk of cardiovascular disease, and decide how to respond to perceived health threats (e.g., taking prescribed medication, changing diet, exercising). Patients with low health literacy and numeracy especially struggle to understand such information. Conversational agents (CAs) may be well suited for explaining complex concepts to patients with limited health literacy by using exemplary communication techniques without time constraints³. CAs can complement face-to-face communication, potentially reducing the amount of time physicians must spend instructing each individual patient. In fact, CAs have been shown to be at least as effective as human professionals in explaining medical information to patients³,⁴ and with promising findings for older adults acceptance⁵,⁶. While CAs are often used and evaluated in education settings⁷,⁸, evaluations of applications in health care are less frequent⁸. As compared to text-based or lower fidelity interfaces, CAs with realistic facial expressions, gestures, and other relational cues improve learning, support a variety of patient goals, and help diverse patients follow self-care recommendations⁴,⁵,⁶. CAs in EHR portals may especially help patients with low numeracy and literacy because, as in face-to-face communication, they can provide commentary with nonverbal cues (prosody, facial expressions) and verbal cues (risk category labels) to guide gist comprehension. In an effort to guide the development of a novel CA to describe clinical test results in portal environments, we assessed patient comprehension, and message satisfaction with this CA, compared to a video (and audio) of a healthcare professional delivering the same information.

Methods

The CA takes the form of a “talking head”, constructed from a 2D image. Specifically, the input to the system is a 2D frontal face image of a person in neutral expression. Landmarks such as lip corner, eye corners and nose tip are detected and located on the input image. A 3D generic model is morphed to match the appearance of the input image⁹. The CA audio is generated by a commercial text-to-speech synthesizer (Vocalware); the pitch and timing are modified to better match the tone (light, grave, or neutral) desired for each interaction. During the animation of the CA, both expressions and lip movements are synthesized by moving a set of control points on the 3D model to match the visemes with corresponding emotions; the CA audio is synchronized with the lip movement by matching the corresponding visemes, as illustrated in Figure 1 (left). We used a Dell 1908 monitor (resolution: 1440x900), with videos and CA messages presented on 1280x720 resolution and 23 fps.

Figure 1. Left: 2D image to 3D CA⁹. Right: Video-enhanced and CA-enhanced formats.

We conducted a study to evaluate the CA relative to the video provider and to a standard portal control condition. Participants were assigned to one of three message conditions: (a) video recordings, (b) synthesized CA constructed from the same physician’s videos (Figure 1: right) or (c) control condition, standard format typically used in actual EHR portals (table of numbers)¹⁰. All messages were about cholesterol test results and the enhanced-formats were accompanied by a graphical representation¹⁰. Older adults (n=108) matched on factors such as gender, age, and education (69% female; mean= 71 years old, 16 years education) watched a set of messages that described results indicating lower, borderline, or higher risk associated with the test results. They answered questions about the risk
level associated with the test results, and message satisfaction. Participants assigned to the CA condition were also asked to directly compare the CA and the video (also included audio) provider at the end of the session.

**Results**

No significant differences were found in message comprehension (Video=87.5%, CA=84.7% accuracy; t(35)<0.7;p=0.5;d=0.2;power=0.1). Compared to a standard format, both formats improved comprehension (CA=84.7%, Standard=67.6% accuracy; t(35)>3;p<.01; d=0.6;power=0.7). Furthermore, no differences were found for message satisfaction (Video=6.1, CA=6.0,9-pt scale; t(35)<0.3;p=0.8;d=0.1;power=0.1). These results reveal that blind to the existence of other formats, participants were equally satisfied with the CA and the videos. Also encouraging, the CA was overall preferred compared with the standard format (CA = 6.0, Standard= 5.1, 9-pt scale; t(35)=2.4;p<.01; d=0.4;power=0.4). However, when prompted to choose between the video and the CA, participants significantly preferred to receive their test results with the video of a real person (Video=81.9%, CA=18.1% preference; t(35) > 9.9;p< .01, d= 2.0; power=1). Participants preferred the video format because the person was more relatable, personable, conversational and with better voice inflections, or it explained more (despite both formats delivering identical content). Out of the justifications, 64% of participants preferred the video because it was delivered by a human, or conversely not by a “robot”. Those preferring the CA mentioned that it was easier to understand, and had superior eye-contact.

**Discussion**

Both video and CA effectively communicated cholesterol test results compared to standard messages. More generally, CAs can produce a significant effect on learning in part by engendering social responses. Nonetheless, often patient self-reported preferences and objective comprehension are not aligned. While the CA was as intelligible as the video, the CA appearance can be improved to boost preference. We next intend to provide choices from a set of CAs that vary in gender, age, and race, which may improve patients’ satisfaction. Further studies should explore different health care applications (e.g., medication instructions) and the appropriate level of CA realism (animated vs static) that produces trust and ‘social stance’. Increasing realism may have drawbacks and produce negative responses (e.g., ‘uncanny valley’ effect) and our photo-realistic CA may fall into that effect, such that less realistic CAs could be more effective. Ultimately, improving cognitive accessibility of health information through CA-based systems should increase portal utilization by older adults by boosting perceived usefulness as well as ease of use of the technology, as suggested by technology acceptance models. For example, older adult beliefs about ease of and usefulness of web-based services tend to predict intent to use. Portals that can be readily used by older adults with limited cognitive ability are likely to be used by a wide range of users.

**References**

EHR-based Quality Measurement to Reduce Antibiotic Use in Children

L. Charles Bailey, MD, PhD1,2, Hanieh Razzaghi, MPH1, Elizabeth R Earley, MPH1, Jeanhee Moon, PhD1, Levon H. Utitdjian, MD, MBI1,2, Jessica Hawkins, BS1, and Christopher B. Forrest, MD, PhD1,2
1The Children’s Hospital of Philadelphia, Philadelphia, PA; 2 Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA

Introduction

Common low-acuity conditions such as respiratory infections are a major driver of antibiotic prescription and usage, leading to significant side effects as well as increasing antibiotic resistance in community flora. In young children, middle ear conditions are important contributors to this phenomenon. Otitis media with effusion (OME), a persistent collection of fluid in the middle ear, occurs due to impaired drainage rather than infection, and clinical guidelines have recommended for over a decade against use of antibiotics to treat OME. Nonetheless, prescription of antibiotics persists. Because OME is largely treated in the outpatient and primary care settings, it is particularly suited to quality measurement and improvement efforts across large populations. To address this requirement, we have developed an EHR-oriented quality measure intended to evaluate and promote antibiotic avoidance in children with OME. In order to test operating characteristics of the measure, we used outpatient data from PEDSnet, a collaboration among large pediatric health systems that includes sharing of structured EHR data standardized in the PEDSnet Common Data Model, an extension of the OMOP/OHDSI Common Data Model version 5.

Methods

Specification. Because OME exists on a clinical continuum with infectious acute otitis media, reliable automated recognition of OME from clinical documentation has proven difficult. We therefore specified the inclusion criteria based on recorded diagnoses reflecting the clinician’s diagnostic judgment. Codesets specific to OME were constructed from ICD-9-CM, ICD-10-CM, SNOMED-CT, and Intelligent Medical Objects (IMO) terminologies, based on medical records review, subject matter expert review, and crosswalks between terminologies. Similarly, codesets for frequently used oral antibiotics were derived from RxNorm and NDC based on review of prescription records in a large pediatric network. Subsequently, diagnoses recorded at visits where antibiotics were prescribed were reviewed, in order to compile codesets in each of the above terminologies reflecting common evidence-based reasons for antibiotic prescription. Measure specifications were expressed in HQMF using the Measure Authoring Tool, as well as in SQL and R for testing.

The measure was specified for outpatient clinical visits at which the patient was aged 2 through 155 months, a diagnosis of OME was recorded, and no other common diagnosis for which antibiotics were indicated was recorded. Each eligible visit was scored for the presence (failure) or absence (success) of an antibiotic prescription. Composite scores for individual clinicians or clinician groups were computed as the proportion of successful visits that were scored as success.

Testing. As a first step, the results of the computed measure were compared to manual record review in one health system (CHOP). A sample of 100 eligible encounters was drawn from CHOP data, with proportional representation of specialties and twofold over-representation of the top and bottom 10% of clinicians, and 125 non-eligible encounters were sampled from overall primary care (100) and otolaryngology (25) visits. Encounters were reviewed for accuracy of both diagnostic coding and medication capture.

In addition, because clinicians enter diagnoses using IMO’s proprietary terminology, which differs in granularity and structure from publically available diagnostic terminologies, results of the measure computed using IMO terms directly was compared to results using ICD-9-CM codes derived for billing.

To further evaluate performance in different contexts, testing was expanded to include data from six pediatric health system members of PEDSnet, using one of three different EHR systems. Score distributions across individuals, departments, and institutions were compared using one-way ANOVA. As a control for reliability, year-to-year score distributions within institutions were also compared using one-way ANOVA. In all cases, entities with <5 eligible visits per measurement year were excluded to limit effects of score quantization.
Results

Manual review. Using the results of manual review as a benchmark, sensitivity of the computed measure across the sample of 225 visits was 0.90; specificity was 0.92.

Diagnostic terminology. While SME review of crosswalks between IMO and ICD-9-CM demonstrated discordance between the terminologies (e.g. IMO “with effusion otitis media” mapped to ICD-9-CM “acute nonsuppurative otitis media”), correlation of scores using SME-assigned IMO terms directly and using mapped ICD-9-CM codes was excellent, (Figure 1).

Reliability and Discriminant Ability. The test dataset contained 197,786 eligible visits spanning 104,649 patients and 5,193 clinicians. Overall clinician performance was good, with mean scores >70%, and the 3rd quartile at 100% for all sites (Table 2). Nonetheless, the range of scores was broad, and ANOVA demonstrated meaningful differences across sites (F=35.62, p<0.001).

<table>
<thead>
<tr>
<th>Site</th>
<th>Min</th>
<th>1st Q</th>
<th>Mean</th>
<th>SD</th>
<th>3rd Q</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.00</td>
<td>75.00</td>
<td>83.20</td>
<td>28.77</td>
<td>100.00</td>
<td>100.00</td>
</tr>
<tr>
<td>B</td>
<td>0.00</td>
<td>66.67</td>
<td>79.95</td>
<td>32.04</td>
<td>100.00</td>
<td>100.00</td>
</tr>
<tr>
<td>C</td>
<td>0.00</td>
<td>50.00</td>
<td>71.04</td>
<td>38.34</td>
<td>100.00</td>
<td>100.00</td>
</tr>
<tr>
<td>D</td>
<td>0.00</td>
<td>60.00</td>
<td>73.57</td>
<td>31.53</td>
<td>100.00</td>
<td>100.00</td>
</tr>
<tr>
<td>E</td>
<td>0.00</td>
<td>100.00</td>
<td>92.36</td>
<td>22.83</td>
<td>100.00</td>
<td>100.00</td>
</tr>
<tr>
<td>F</td>
<td>0.00</td>
<td>100.00</td>
<td>92.94</td>
<td>20.71</td>
<td>100.00</td>
<td>100.00</td>
</tr>
</tbody>
</table>

Table 2. Score distributions for individual clinicians at participating PEDSnet sites

In contrast, year-to-year (2012-01-01 to 2016-06-30) score distributions for clinicians with ≥5 eligible visits in each year were not significantly different for 4 of 6 sites (Table 3a). Similarly, scores from the 9 months before/after the change in billing from ICD-9-CM to ICD-10-CM differed significantly for only 2 sites (Table 3b).

<table>
<thead>
<tr>
<th>Site</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>28.01</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>B</td>
<td>1.359</td>
<td>.245</td>
</tr>
<tr>
<td>C</td>
<td>1.016</td>
<td>.317</td>
</tr>
<tr>
<td>D</td>
<td>21.88</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>E</td>
<td>2.592</td>
<td>.114</td>
</tr>
<tr>
<td>F</td>
<td>.100</td>
<td>.752</td>
</tr>
</tbody>
</table>

Table 3a. Year-to-year ANOVA

<table>
<thead>
<tr>
<th>Site</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>73.83</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>B</td>
<td>.147</td>
<td>.702</td>
</tr>
<tr>
<td>C</td>
<td>.005</td>
<td>.942</td>
</tr>
<tr>
<td>D</td>
<td>13.73</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>E</td>
<td>.823</td>
<td>.366</td>
</tr>
<tr>
<td>F</td>
<td>3.067</td>
<td>.081</td>
</tr>
</tbody>
</table>

Table 3b. Pre- and post-ICD-10-CM ANOVA

Discussion

We describe the development and initial testing of a clinical quality measure tailored to EHR-derived data (an emeasure) intended to detect overuse of antibiotics to treat OME, a non-infectious condition. Physiologic recognition of OME is hard, so we specified the measure using clinician-asserted diagnoses rather than techniques such as text mining of physical examination findings; this approach significantly increases feasibility of the measure, and can be appropriate for measures of therapeutic decision-making rather than disease recognition.

Our testing strategy focused on evaluation of measure operating characteristics over a large test population rather than detailed manual record review. This allows for testing across a wider variety of technical and clinical contexts, but relies on the specification of the measure over well-defined discrete data elements. The terminology standardization provided by the OHDSI CDM further accelerated testing.

Acknowledgments

The authors are grateful for the data contributed by PEDSnet patients and families, and the work of PEDSnet informatics teams in construction of the data network. Work on this project was supported by the Patient-Centered Outcomes Research Institute (PCORI) (CDRN-1306-01556) and the Agency for Healthcare Research and Quality (AHRQ) (U18HS020508).
Usability and Acceptability of a System to Identify Pediatric Patients at Risk of 30-day Hospital Readmission Prior to Discharge

Amie Barda¹, MS, Lingyun Shi, MS¹, Andrew Urbach, MD², Srinivasan Suresh, MD², and Fuchiang (Rich) Tsui, PhD¹

¹Tsui Lab, Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA
²Children’s Hospital of Pittsburgh of UPMC, Pittsburgh, PA

Introduction: In a joint venture between the Biomedical Informatics Department of the University of Pittsburgh and the Children’s Hospital of Pittsburgh (CHP) of the University of Pittsburgh Medical Center (UMPC), the System for Hospital Adaptive Readmission Prediction and Management (SHARP) was developed as part of an effort to reduce 30-day pediatric hospital readmissions. SHARP utilizes a variety of structured and unstructured electronic health record (EHR) data to provide real-time estimates of an inpatient’s risk of readmission. The SHARP interface is integrated into the Cerner Millennium system currently in use at CHP and displays a variety of information related to readmission risk, including plots depicting readmission risk trends for individual patients, high-level categories of factors contributing to risk (e.g., laboratory risk factors), patient lists ranked by decreasing readmission risk, and heatmaps depicting readmission risk distribution within units/wards. The interface was developed iteratively using input from a small clinical team at CHP. It was anticipated that the information in SHARP would allow inpatient care providers to better target high-risk patients and intervene prior to discharge to reduce readmissions.

To prepare users for interacting with SHARP, an educational training video was prepared. The 9-minute video developed includes a brief overview of the system development and a demonstration of how to access and navigate SHARP within the Cerner system. The purpose of this study was to obtain end-user perceptions on the 1) completeness and clarity of the educational video, 2) usability of the SHARP interface, and 3) acceptability of using information in SHARP to change current clinical practice.

Methods: We recruited a convenience sample of inpatient care providers from CHP to participate in a 2-phase study. To recruit participants, we attended existing hospital group meetings (e.g., readmissions workgroup, monthly resident meeting) and asked some of our hospital contacts to recruit colleagues who might be willing to participate. Preference was given to care providers who had not actively participated in SHARP development. In the first phase, participants were asked to partake in a 1-hour session that consisted of watching the 9-minute educational video on SHARP, completing a brief survey, a 15-minute question and answer session, and a 30-minute semi-structured interview. The survey consisted of 4 questions regarding relevant participant characteristics (e.g., clinical position, prior knowledge about SHARP) and 12 Likert-scale questions designed to assess the participants’ initial perceptions on the usability and acceptability of SHARP. After completing the survey, participants were given the opportunity to ask questions regarding the development and use of the SHARP system. Based on participant availability, both individual and group interviews were conducted in which participants were asked about perceived barriers and facilitators to the adoption of SHARP within CHP.

For the second phase of the study, we invited all participants to pilot SHARP in a live clinical setting. Participation was on a volunteer basis only and participants were recruited during the phase 1 session. Willing participants were given access to the system and asked to use it as part of their regular workflow for a 2-week period. At the end of the 2-week period, participants were asked to answer a series of structured and unstructured questions about their experiences with SHARP in the live clinical setting. When possible, questions were asked during 30-minute in-person interview sessions to allow for follow-up questions. In cases where a volunteer was unable to sit for an interview, answers to questions were collected via an online survey with no follow-up questions.

In addition to taking notes, all interview and question-answer sessions were audio-recorded when permitted and later transcribed. Feedback from structured questions was summarized using percentages and basic statistics. Transcribed audio files, interview notes, and free-text feedback from unstructured questions were grouped by study phase and analyzed for underlying themes.

Results: For phase 1, a total of 31 care providers participated in 6 separate sessions (2 individual, 4 group). Recordings were obtained for 4 of the 6 sessions. Participants consisted of at least one representative of each major clinical role at CHP (e.g., resident, nurse, care manager, etc.) and had been practicing clinically anywhere from 6 months to 31 years (median = 2.3 yrs, mean = 5.8 yrs). Most participants had little to no knowledge of SHARP prior to attending the session (81%). Only 2 participants had been actively involved in the initial development of
SHARP. The phase 1 sessions revealed that although participants felt comfortable about interacting with the SHARP interface based on the educational video alone, they wanted more information about the development and validation of the risk prediction model to improve perceived trustworthiness of the system. Additionally, participants expressed concern that the long length of the educational video would prevent many users from watching it in full. It was noted that SHARP seemed intuitive and easy to learn and that the educational video might be too detailed.

Although 81% of participants thought that using SHARP would not disrupt their current workflow, only 22% of participants anticipated using it frequently. The simple, intuitive interface design and integration into existing EHR workflow were noted by most participants as facilitators to use; however, most participants expressed concerns about SHARP’s ability to impact clinical practice based on the current information available. More specifically, the high-level risk factors provided by SHARP were considered clinically irrelevant or not specific enough to be actionable. Some participants stated that they often knew which of their patients would be readmitted and many mentioned that they were unsure what could change about current clinical practice to help alleviate the risk. Furthermore, participants felt that current time-constraints left no room for additional attention to be spent on high-risk patients without knowledge of specific, effective interventions. Some participants stated that SHARP may be useful in streamlining decisions about patient discharges, but most saw little utility for the information in the inpatient setting. Many participants thought that post-discharge interventions (e.g., follow-up phone call) would be the most effective way to reduce readmissions.

For phase 2, only 5 total care providers volunteered to participate (1 attending, 1 care manager, 2 nurses, 1 nursing unit director). Only 1 user had been actively involved in the initial development of SHARP, while the remaining 4 had not been familiar with the system prior to attending the phase 1 training session. Only 2 of the 5 users could provide feedback via an in-person interview. During the 2-week period, 1 user reported using SHARP very frequently, 1 reported using SHARP occasionally, and 3 reported using SHARP infrequently or very infrequently. All users agreed that the SHARP system was well-integrated into Cerner and current EHR workflow, and 2 reported that the visibility of SHARP access points in Cerner prompted them to use the system. None of the users reported having issues understanding the information within SHARP and 4 reported that they felt the information in the system was accurate. Although no usability issues with the SHARP interface were uncovered during piloting, all but 1 user reported that they would not use it daily as the information is either not clinically actionable or not relevant to the specific patient population for which they care (e.g., intensive care patients).

**Discussion and Conclusion:** Overall, both phases of this study revealed that although the SHARP interface is highly usable and well-integrated into EHR workflow, the anticipated acceptability of SHARP to impact clinical practice is not promising. The two main barriers to utilizing SHARP information to change clinical practice were 1) a lack of actionable or relevant information in SHARP and 2) resistance to increased time spent on high-risk patients without specific interventions that have perceivable benefits. Concerns about improper use of the system were further exacerbated by the potential for users to resist the education process due to the video length.

This study suffered from a poor volunteer rate in the second phase. The resulting low number of participants can be attributed to a lack of interest in demoing the SHARP system in the live clinical setting, which further demonstrated the low acceptability of the information provided by the tool. Specific reasons for high readmission risks were not available from the predictive model, thus limiting the amount of actionable information in SHARP. Based on input from the original development team, it was anticipated that providers would still use SHARP to identify high-risk patients and subsequently use the provided risk categories in conjunction with clinical judgement to determine specific reasons for risk and develop an intervention plan. This extra level of work was unacceptable to study participants, which suggests that a more user-centered approach to system design should have been utilized. More specifically, in depth examination of end-user information needs and workflow should have been completed.

Despite the limitation of a low number of participants, the low acceptability found during this study led the technical and clinical development team to agree that additional planning is required before deploying SHARP within the inpatient clinical setting. A new intervention plan is under development, in which the current SHARP system will be used by a small set of users to perform targeted post-discharge interventions (e.g., tailored follow-up phone calls to high-risk patients). Additionally, information needs and intervention strategies for the inpatient setting are currently being reassessed. SHARP will be modified accordingly and the acceptability of the system will then be reevaluated prior to deployment in the inpatient setting. The results of this study can be considered when developing implementation strategies for readmission risk prediction models in clinical settings.
Large-Scale Text Mining of Social Determinants from Electronic Health Records: Case Studies of Homelessness and Adverse Childhood Experiences

Cosmin A. Bejan, PhD¹, John Angiolillo, MD², Douglas Conway, BA³, Robertson Nash, PhD, ACNP², Jana K. Shirey-Rice, PhD², Loren Lipworth, ScD², Robert M. Cronin, MD, MS¹,2,4, Jill Pulley, MBA³, Sunil Kripalani, MD, MS², Shari Barkin, MD, MSHS⁴, Kevin B. Johnson, MD, MS¹,4, Joshua C. Denny, MD, MS¹,2

¹Department of Biomedical Informatics, ²Department of Medicine, ³Institute for Clinical and Translational Research, ⁴Department of Pediatrics; Vanderbilt University, Nashville, TN

Introduction

Social determinants of health (SDH) have recently received a growing interest as they provided evidence for being associated with an early onset and progression of various diseases, and a risk of premature death.[1] This interest was also reflected in a report from the National Academy of Medicine emphasizing the importance of capturing SDH in Electronic Health Records (EHRs).[2] The objective of this study was to implement a big data approach for extracting two severe SHD, homelessness and adverse childhood experiences (ACE), from a large dataset of notes.

Methods

Figure 1 illustrates the architecture of our approach.[3] Using the Vanderbilt de-identified version of the EHR, which includes >100 million notes, we integrated unsupervised learning and information retrieval (IR) methods to capture the two SDH profiles and to identify the patients that best matched these profiles. First, we built a list of query terms for each profile. Since ACE is a complex phenotype with experiences ranging from sexual to psychological abuse, we relied on domain expertise to identify the initial query of this SDH. For homelessness, we employed two data-driven methods, lexical association[4] and word2vec,[5] to expand two seed keywords, homelessness and homeless. We trained word2vec on 10 million notes randomly sampled from the EHR and used a vector dimension of 100. Next, we computed the similarity between a query and the notes of each patient by using an IR vector space model trained word2vec on 10 million notes randomly sampled from the EHR and used a vector dimension of 100. Next, we computed the similarity between a query and the notes of each patient by using an IR vector space model architecture based on the standard TF-IDF weighted cosine metric. We implemented the IR model on top of our EHR database, a secure IBM Netezza data warehouse consisting of a parallel computing architecture, to ensure the scalability of our system. Each search over the entire EHR lasted <20 min. After retrieval, negation detection[6,7] was employed to exclude the patients with a high prevalence of negated terms. Finally, the top retrieved results were manually examined by seven assessors. Query reformulation based on relevance feedback was adopted by analyzing the top 20 ranked results of each retrieval. For homelessness query expansion, the most relevant terms were selected from the top 50 keywords generated by the data-driven methods. After each query was settled, the final retrieval was executed and the top ranked patients were manually assessed. Specifically, for ACE, the top 1,000 patients were categorized as either ACE, not ACE, or undetermined. For homelessness, the assessment was performed at visit level using five categories: homeless, settled, at risk, undetermined, and pediatric. Since our next objective is to investigate associations between homelessness and various diseases using a case-control study design, for assessment, 600 patients were selected from 3 distinct sets: 1) the case set contained the top 200 ranked patients, 2) the fuzzy set–randomly selected patients with a rank >5000, and 3) the control set–randomly selected patients who don’t have homelessness terms in their notes. The evaluation was performed using precision-recall curves, the precision of top k ranked results (P@k), and the area under the precision-recall curve (AUPRC). The 95% confidence intervals (CI) of AUPRC estimators were computed using a bootstrap procedure.

Results

Lexical association and word2vec proved to be viable methods for homelessness query expansion. Table 1 lists the top 8 ranked keywords relevant for this determinant. However, a comparative study on the top 50 keywords extracted by the two methods revealed superior performance values achieved by word2vec (Figure 2 and Table 2).

As a result of query expansion and reformulation, the final query terms used for retrieval assessment were 1) for homelessness: homeless, homelessness, shelter, unemployed, jobless, and incarceration; and 2) for ACE: child shelter, and incarcerated.
abuse, sexual abuse, child neglect, childhood trauma, child protective service, physical abuse, psychological abuse, verbal abuse, poverty, food insecurity, cps supervisor, cps report, cps worker, and cps investigation. Based on these terms, the system retrieved 35,220 (1.3%) and 27,861 (1.1%) patients for homelessness and ACE, respectively. For both SDH, a substantial inter-rater agreement (kappa=0.8) was achieved. Out of the top 200 patients of the homelessness retrieval, 15 were pediatric, and 172 were found as relevant for this determinant (i.e., they have at least one visit assessed as relevant for homelessness). As listed in Table 3, the homelessness precision of the top 185 adult patients was 93% (172/185). This is substantially higher when compared with the precision of 40% achieved on the fuzzy set. For ACE, the system obtained a precision of 70% for the top 1,000 ranked patients. A comparative evaluation of the top 185 patients retrieved for the two SHD (Figure 3) indicated a higher AUPRC of 94% for homelessness identification (see details in Table 3). An analysis of the ACE patients revealed that most adverse experiences included sexual and physical abuse and top prevalent abusers were father and mother (Table 4). An alluvial diagram of the patients assessed for homelessness across consecutive visits (Figure 4) indicated that this condition was chronic for most of the homeless patients. The prominent cyclic transitions between the at risk and homeless categories captured an episodic trend of homeless patients.

Discussion

We described a feasible solution for mining SDH from a large-scale EHR, which opens the way for conducting large phenome wide association studies[8] on the impact of these determinants on the overall quality of life. To the best of our knowledge, this is the first study on SDH text mining from all the notes of a big EHR repository. Our system performance could be improved by integrating structured data in the retrieval model. However, the only structured data we could use from our EHR is the V60.0 administrative code for homelessness, which is not sufficient for extracting this determinant.[9] In this study, we found that only 62.8% of patients assessed as relevant for homelessness have the V60.0 code. Future work could expand the applicability of our system to identify additional determinants of health.

References

A Digital Health Advisor for High-Need, High-Cost Patients: Exploring Needs, Functions and System Constraints

Onil Bhattacharyya, PhD, MD1,3, Arnav Shah, MPP2, Lovisa Gustafsson, MBA2, Kathryn Mossman, PhD3, Eric Schneider, MD, MSc2
1University of Toronto, Toronto, Ontario, Canada; 2The Commonwealth Fund, New York City, NY, USA; 3Women’s College Hospital, Toronto, Ontario, Canada

Introduction

High-need, high-cost (HNHC) patients often have multiple conditions, are seen in a range of different institutions, and account for a disproportionate amount of system cost, yet they are poorly served by a health system built around diseases and institutions1,2,3. Digital tools that support an individual’s system for maintaining their health could help address the needs of frail elderly and people with multiple chronic conditions. Information technology has transformed many industries, empowering consumers and often reconfiguring services around their needs. There are over 165,000 health apps currently available, but a recent review found most available digital tools for chronic disease are piecemeal and do not address the needs of patients with complex chronic conditions4. Electronic health records with integrated health portals exist, but these tools and data are usually the property of a health institution and not owned by patients, who are seeking more customized care for their specific needs. To promote the uptake of integrated tools we need to consider patient concerns as well as those of providers, payers and policymakers. Through qualitative research methods, we explored the needs of HNHC patients that could be addressed by a digital health advisor (DHA), identified key functions of this tool, and the policy and system constraints that need to be overcome to promote its development and widespread dissemination.

Methods

We carried out focus groups and interviews to better understand the perspectives of key stakeholders regarding the key features of a DHA and the barriers that need to be addressed to develop it. We started with system stakeholders to understand current activity related to the DHA, moved to patients and caregivers to see what was desirable in such a tool, and went back to the system stakeholders to understand what was feasible and how to achieve it.

- The first focus group included clinicians, researchers, technologists, health system leaders, regulators, investors, entrepreneurs, patient advocates, and policymakers. It focused on the feasibility of developing such a tool, its potential impact, and key use cases among HNHC patients.
- We then conducted semi-structured interviews with 8 patient-caregiver pairings, focusing on patients with multiple chronic conditions and the frail elderly, to identify a broad set of needs and goals of HNHC patients. They were asked to iteratively prioritize a series of potential functions for a DHA based on illustrative examples.
- The second focus group involved ten stakeholders, including policymakers, regulators, clinicians, experts in informatics, advertising, and health services, as well as patient advocates, a technology industry executive, and health system leaders. We presented the proposed functions of a DHA, explored market opportunities and system constraints and discussed how to overcome them.

We conducted qualitative analysis of the focus group and interview data to identify patterns and key themes, focusing on identifying stakeholder needs, DHA functions, market opportunities and system barriers related to these functions.

Results

Feedback from the first focus group suggested that a DHA is both feasible and necessary, and some isolated components already exist. Participants noted that while the digital tool might face challenges involving security and privacy, these problems could be overcome5. The tool was considered particularly useful for HNHC patients, whose needs should be further explored. In our interviews with HNHC patients and their caregivers, we identified a broad set of needs and goals that a DHA may be able to address. These fell along two dimensions, ranging from functional needs and emotional needs, and medical and personal needs. This included the following themes: “live my life” (manage day to day tasks that are important to them); “love my life” (preserve dignity and connections while adapting to changes in their health status); “manage my health” (feel empowered to make smart medical and health decisions); and “feel understood” (communicate how conditions affect their physical and emotional state to their
family, friends and care team). Based on these needs, we developed a clickable DHA prototype with the following simulated functions:

- Tracking health indicators with a metric dashboard that collects data from apps and wearables, providing simple recommendations to help patients take action to make healthy decisions
- Providing advice on health-related questions either through digital assistant integration, like Siri or Alexa, or by connecting to a medical practitioner by phone or video, drawing on information from medical records, personal metrics, and available community resources
- Providing a care journal to help create a holistic picture to share with practitioners and foster deeper connections with the medical team
- Providing a shared calendar with all medical appointments, a document centre, and a task manager to improve coordination and communication assistance between patients and their care team, including help with scheduling appointments and transportation

In our second focus group, we presented these proposed DHA functions and identified a number of system and policy constraints to consider in developing the DHA. These include:

- The focus of the DHA needs to be on supporting patients in thriving rather than the management of disease
- Health care delivery organizations are critical to this effort, but it also needs to foster support in target communities, including patients, caregivers, families and social networks
- Extensive collaboration is needed to develop a robust DHA, which should be an integrated suite of tools
- Novel analytics alone are not sufficient; data must be made available in a usable format, and it is currently held by many different groups
- A useful DHA is more likely to emerge in response to bottom-up and middle out efforts rather than a top-down approach, which could include unique partnerships with organizations focused on consumer needs, such as consumer advocacy groups, and large retail companies
- We need to go beyond the policy sphere to engage with public markets, which can provide a competitive environment in which to test and refine the DHA
- The potential returns on investment and incentives need to be clear to attract social entrepreneurs and other developers to create a DHA
- The business case and policy incentives need to be developed to encourage broad data sharing among community, government and commercial initiatives

**Discussion**

This research identifies the key needs, functions and system considerations for developing a DHA that address the needs and goals of HNHC patients who could benefit most from this type of digital tool. It highlights the importance of community support and buy-in, along with creative partnerships between diverse organizations and a role for both policymakers and the marketplace in developing a robust DHA. These findings can also inform development of similar technologies for patients and help health system leaders consider the potential and challenges of patient-facing digital tools to meet the needs of their highest-cost patients.

**References**

Introduction

Background: Potential safety events in the outpatient setting are difficult to identify, quantify and categorize, leading to unique challenges in addressing ambulatory risk through quality improvement initiatives. Over the last 7 years, the OpenNotes initiative has sought to empower and engage patients and family members by inviting them to review visit notes through the patient portals, and now reaches over 12 million patients nationwide. Patients who read notes have reported feeling increased control over managing their health care and improved medication adherence, among other benefits. Furthermore, patients and caregivers may also assist in improving the accuracy and quality of ambulatory documentation. In a recent report, 21% of adult primary care patients who provided feedback on their online notes through an electronic reporting tool reported a potential inaccuracy. The majority of these reports (57%) resulted in a change to the medical record. However, no studies have explored the impact beyond the primary care setting and among other patient populations, such as the pediatric population, where parents most often manage care on behalf of the patient.

Objective: As part of an ongoing quality improvement project dedicated to improving the quality of clinic visit notes, the objective of this study was to quantify and characterize patient and parent reports of potential documentation inaccuracies and resultant safety concerns among ambulatory primary care and subspecialty open notes available through an online patient portal at 2 urban academic hospitals.

Methods

Design: A prospective cohort quality improvement (QI) initiative designed with patients and families

Setting: The quality improvement initiative was implemented at an urban academic adult medical center (Beth Israel Deaconess Medical Center; BIDMC) starting in August, 2014 and a free-standing urban academic children’s hospital (Boston Children’s Hospital; BCH) starting in June, 2016.

Patients/participants: Participants at the adult hospital included patients using the online patient portal (PatientSite) who had an eligible clinic note during the study period. Eligible clinic notes included pilot teams at 2 primary care practices and all the obstetrics and gynecology clinics. Participants at the children’s hospital included patients (≥ 13 years old), parents (for patients < 18 years old) and medical guardians of adult patients who had access to an eligible clinic note through their online patient portal (MyChildren’s). Eligible clinics included all primary care and specialty clinics with the exception of clinics with a higher risk of managing confidential or sensitive information who did not share their notes on the portal (i.e. adolescent clinic, adolescent substance abuse clinic, child protection team clinic, psychiatry clinic and gynecology clinic). Providers could also “hide” individual clinic notes so that they did not appear on the portal, if necessary.

Intervention: The intervention was a ≤ 10 item online patient reporting tool that was co-developed and pilot tested by the patient and family advisory council (PFAC), embedded through a link at the end of the note (PatientSite) or in the secure message notification regarding note availability (MyChildren’s). At both sites participants were invited to provide feedback on their note as part of a QI initiative. Reports flowed through a QI algorithm for real-time responses. The reporting tool requested feedback about whether and how well the note reflected what happened at the visit, whether the patient/parent understood the described care plan, whether they would follow the care plan, and whether there were any inaccuracies in the visit note. The reporting tool queried any language that bothered respondents and provided space for voluntary positive feedback about the provider. We defined a potential safety concern as a “No” response to “Did you understand the care plan” or “Do you plan to follow the care plan?”, or a “Yes” response to “Did you find a possible inaccuracy?”. Each safety concern was reviewed by Patient Relations personnel and outreach was made to patients/parents, and providers, as needed. Any resulting changes to the record or practice were coded accordingly.

Outcomes

We received a total of 1,059 patient family reports; 318 at BCH and 741 at BIDMC. Overall, 30% of reports at BCH identified a potential inaccuracy and an additional 3% were unsure if they had found an inaccuracy. At BIDMC, 22.3% identified potential inaccuracies, while 2.4% were unsure. The majority of inaccuracies reported at both sites included inaccuracies in the description of the chief complaint, inaccuracies in the problem list and medication list,
as well as information that was missing from the note (table 1). A few additional notable categories of inaccuracies included errors in laterality and wrong name/wrong patient errors. Most of the inaccuracies resulted in a change to the medical record (70% at BCH and 77.8% at BIDMC) (Table 2). 45% of respondents at BCH and 69% of respondents at BIDMC also included positive feedback for their providers.

**Discussion**

Our findings add to a growing literature that patients and families can identify errors or problems in care and can be valuable partners in quality improvement initiatives. Nearly a third of the patient/family reports identified a potential safety concern, with the majority resulting in an amendment to the medical record or a change in practice. Few patients or parents reported language that was bothersome and over half of the respondents provided positive feedback for their providers.

**References**


| Table 1. Inaccuracies in Ambulatory Visit Notes Reported by Patients and Parents |
|-------------------------------------------------|-----------|-----------|
| Total Number of Reports                         | BCH 318   | BIDMC 741 |
| Number of Reports with Inaccuracies (% of total reports) | 99 (31%)  | 165 (22.3%) |
| Total number of Inaccuracies*                   | 150       | 247       |
| Medications                                     | 10.7      | 16.6      |
| Description of Chief Complaint                  | 17.3      | 14.2      |
| Problem List                                    | 18.7      | 16.6      |
| Social History                                  | 5.3       | 4         |
| Family History                                  | 6.7       | 7.3       |
| Description of Physical Exam                    | 6         | 7.3       |
| Other Missing Information                       | 9.3       | 13.8      |
| Names of Healthcare Providers                   | 6.7       | 0.8       |
| Appointment Scheduling                          | 1.3       | 0.4       |
| Other                                          | 18        | 19        |

| Table 2. Outcome of Safety Concerns Following Review |
|---------------------------------------------------|-----------|-----------|
| Safety Concern Outcome§                           | BCH N=69¥ | BIDMC N=99¥ |
| Declined Intervention                             | 4 (5.7)   | 3 (3)     |
| Defer to Patient Relations                        | 6 (8.7)   | 15 (15.2) |
| Informal Amendment Made                           | 48 (70)   | 77 (77.8) |
| Informal Amendment Declined                       | 10 (14.5) | 3 (3)     |
| Formal Amendment Made                             | 0         | 1 (1)     |

§ Includes both reported inaccuracies and challenges involving the care plan.
¥Safety concerns where Patient Relations review determined that no action was needed, the concern was resolved in conversation, or there was no response from the patient/family when they were contacted, were excluded.

*Some reports contained multiple inaccuracies
Evaluating the Quality of Patient Address Data in an EHR system

Michael N. Cantor, MD MA
NYU School of Medicine, New York, NY USA

Introduction

As the role of community-level social determinants of health (SDH) grows, accurate geolocation data for patients is increasingly important. Geographic information can be used to identify and target high-risk neighborhoods, for example. In densely populated areas, mapping by zip codes may not be specific enough to represent the diversity of the underlying population for these efforts. Mapping by census tract gives more specificity, and also allows for the incorporation of data from the U.S. Census and similar surveys. In support of a project building a database around community-level determinants of health mapped to New York City census tracts, we attempted to map our entire active patient population’s addresses to their respective census tracts. Our EHR system (Epic) imposes no standardized approach for entering address data for either patients or staff, so the address dataset required significant cleansing.

Accurate address data is important throughout the healthcare enterprise, particularly as part of an accurate patient identifier to support care delivery and coordination as well as information exchange. Several vendors provide address cleaning solutions, but these services can involve sharing the information with vendors (e.g. Google Maps) and lead to potential privacy issues, so local solutions are generally preferable. Many researchers assume that address data is standardized and accurate in EHRs. Though there is extensive literature on linking geographic data to clinical data, particularly as related to social determinants of health, we found no articles that mentioned the quality of patient address data that went into their geocoding efforts. Of note, the purpose of this project was to evaluate the quality of address representation, not whether the patient’s address was the correct one.

Methods

This project was a subset of SDH-related work for the New York City Clinical Data Research Network (NYC-CDRN, our PCORnet instance), where the overall aim was to link NYC addresses to their census tracts and a database of community-level determinants (https://github.com/mcantor2/FACETS). With this in mind, we obtained a list of the raw address data from active patients in our EHR system, separated from any other identifying information. Active patients were defined as those whose data had been sent to the NYC-CDRN’s central database. We pre-processed this data using Perl scripts to eliminate common errors and expand common abbreviations (e.g. AVE) to comply with the formatting of publicly available address databases (openaddress.org and NYC’s open data website). We compared both the pre-processed and post-processed street address list (removing apartment information) with the entries in the publicly available databases to determine the number of matches in the dataset and the common factors that led to errors in the address data (e.g non-alphanumeric characters).

Results

We obtained approximately 640,000 unique complete addresses (i.e. with apartment numbers) for 709,000 patients with NYC zip codes in our EHR system. With no processing, we found only 28,600 matches. With pre-processing, we were able to match approximately 595,000 patients’ (84%) street addresses. These addresses represented approximately 557,000 unique address (87%) entries and 297,000 unique street addresses after processing. We had approximately 66,000 unique unmatched street addresses after processing. Within this set of addresses were approximately 37,000 instances of 9250 unique “errors” (e.g. “Strett” is a unique error with 27 instances), which fell into 6 general categories: problem with reference content; algorithmic errors; misspelling; ambiguous abbreviation; PO Boxes; and unstable housing. Samples of these errors may be found in the Table. The Table also shows the prevalence of the different error types, which we obtained by analyzing the most common errors (those with >10 occurrences, comprising 4% of the unique errors and 81% of the total instances) manually.

Discussion

Because patient data can be entered through several different means in our EHR, addresses are susceptible to errors that make them difficult to align with a list of addresses in standard format. One of the main factors behind the low match rate on raw data, for example, is that apartment numbers were often listed as part of the street address, rather than separately. Adding safeguards or other systematic error checking in the EHR, similar to functionality provided by the U.S. Postal Service or other vendors in standardizing addresses when they are entered, could prevent many of
<table>
<thead>
<tr>
<th>Category</th>
<th>Instance</th>
<th>Explanation</th>
<th>Sample prevalence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ambiguous abbreviation</td>
<td>181 West 131ST ST</td>
<td>Suffix vs. Street</td>
<td>40%</td>
</tr>
<tr>
<td>Misspelling</td>
<td>Amster[dam</td>
<td>dame</td>
<td>dams</td>
</tr>
<tr>
<td>Reference content</td>
<td>100 Malcom X Boulevard</td>
<td>Reference entry is 100 Lenox Avenue</td>
<td>18.7%</td>
</tr>
<tr>
<td>PO Box</td>
<td>PO Box 78022</td>
<td>Does not map to street address</td>
<td>8.6%</td>
</tr>
<tr>
<td>Algorithmic error</td>
<td>Apartment</td>
<td>Algorithm only looks at abbreviations</td>
<td>6.7%</td>
</tr>
<tr>
<td>Unstable housing error</td>
<td>Schwartz Building</td>
<td>Homeless Shelter</td>
<td>2.2%</td>
</tr>
</tbody>
</table>

these errors. Some of the matching issues may be unique to a highly concentrated urban environment like NYC, where apartments are more common than private houses. Additionally, idiosyncrasies of NYC addresses, such as the hyphenated street numbers in Queens, may also have led to subtle errors that affected the matching in the same way that idiosyncrasies of the standardized datasets, such as giving street numbers as a range rather than an individual value; streets that were missing from the reference datasets; or streets with more than one official name. The reference data also have their own idiosyncrasies, such as only using numbers in streets names (5 vs Fifth). Similarly, some unmatched addresses were not necessarily erroneous, such as PO boxes or shelters, but also do not give accurate geographical information. Of course, the quality of the representation of the address data is important but ensuring that the address for a patient is correct is ultimately most important when drawing clinical conclusions. Standardized representation is just the first step in this process.

We used exact rather than fuzzy matching at this phase of the project because experience with fuzzy matching found too many possible alternatives, leading to creating as many new errors as those that were solved by process. Additional errors resulted from ambiguities that are unique to address data, such as “ST” representing either “Street,” “Saint,” or a suffix after a number, or non-standard representations of common abbreviations. Vendor systems such as Google Maps do an excellent job of addressing these errors, but, as noted above, may lead to privacy issues and can be expensive. Because the pattern of errors is relatively random, ranging from misspellings to spacing to mismatched addresses and zip codes, the next steps in this project may involve a deep learning approach to improve automated error correction. Other institutions may have more consistent address data, either due to limits on how the data is entered, or a systematic approach to catching and correcting errors before they are entered in the official medical record.

Conclusions

Accurate representations of addresses in the EHR are essential for linking patients to data related to their community-level social determinants of health. Though relatively straightforward processing led to matches for 84% of patients, the remaining address errors occurred across several categories, and were often difficult to address using automated tools and because of their random nature. Standardizing addresses upon their entry, using approaches that are common across many ecommerce sites, would eliminate most of the errors encountered.

References

Data-driven Risk Characterization and Prediction of Renal Failure among Diabetic Type 2 Patients using Electronic Medical Records

Prithwish Chakraborty, Vishrawas Gopalakrishnan, Sharon M.H. Alford, Faisal Farooq

1 IBM Watson Health, Cambridge, MA; 2 SUNY Buffalo, Buffalo, NY; 3 IBM Watson Health, Southfield, MI; 4 IBM Watson Health, Yorktown, NY

Introduction

The ready availability of Electronic Medical Record (EMR) systems has allowed providers and researchers to obtain a longitudinal view of disease and treatment progression from patient encounter histories. In recent years, EMR data from online and cloud-based data marts that collect and normalize data across a wide range of institutions, both acute and ambulatory, have further helped in reducing temporal blind-spots in patient medical history. Intelligent analysis of such meta-records can often lead to a revision of recommended treatment plans. Typically, most studies on disease outcome characterization using EMR records focus only on factors prescribed by experts. However, EMR records may also contain hidden/indirect factors that are correlated with higher risk of disease outcomes. For example, an increased number of diabetic related patient encounters can indicate an increased risk of extreme outcomes. In this study, we aim to identify potentially correlated factors from EMR records using data-driven methods. Specifically, we analyzed EMR records to identify relevant data-driven factors that are correlated with higher risk of renal failure in diabetic type 2 patients - one of the most prevalent afflictions worldwide. We used a L1 regularized cox-proportional hazard model to identify the factors and validated these factors by using these as covariates in a ‘generalized linear model’ based method to predict for renal failures in such patients. Note that we do not posit causality, only, correlated factors, which could be investigated with experts to better understand and possibly to predict disease progression.

Methods

Experimental Setup: We used a de-identified EMR corpus of approximately 47 million unique patients covering 1.6 billion medical encounters. We constructed a diabetic type 2 cohort spanning the time period 1990 to 2016 and covering 4 million patients, following a three step definition: (a) patients must have been diagnosed with diabetes type 2 anytime within this period (presence of Dx codes), (b) they must have had abnormal HbA1C levels (HbA1C ≥ 5.7) from lab observations corresponding to such diagnosis, and (c) had at least 3 diabetes related encounters within this period. A patient satisfying all the three criteria was considered to be diabetic and included in the cohort and, the first time-point when these criteria were satisfied was considered to be the first diagnosis date for the patient. This definition ensures that patients have been identified as diabetic from both procedural and diagnostic methods and hence reduces false positives. From the EMR records, we create normalized medical histories of patients using a number of steps such as encounter-observation and encounter-outcome co-reference resolution. Furthermore, we annotate encounters with renal failures (both acute kidney failure and chronic kidney disease) - both positive and negative cases.

Feature Identification: Typically, experts look at a fixed set of factors such as age, weight and HbA1C level in blood while analyzing renal failures in diabetic patients. In this study, we aim to identify other possible indicators, as supported by the data, from the corpus of near-complete patient medical histories. Following classical biostatistical approaches, risk factors for extreme outcomes like renal failure can be characterized using survival models such as Cox Proportional Hazard (CoxPH) models. In these models,

<table>
<thead>
<tr>
<th>LOINC</th>
<th>Description</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>8277-6</td>
<td>Body surface Area</td>
<td>2345-7 Glucose</td>
</tr>
<tr>
<td>28542-9</td>
<td>Platelet mean volume</td>
<td>26844-6 Monocytes</td>
</tr>
<tr>
<td>30180-4</td>
<td>Basophils/100 leukocytes</td>
<td>8480-6 Intravascular systolic</td>
</tr>
<tr>
<td>26450-7</td>
<td>Eosinophils/100 leukocytes</td>
<td>2571-8 Triglyceride</td>
</tr>
<tr>
<td>3094-0</td>
<td>Urea nitrogen [Mass/volume]</td>
<td>8310-5 Body temperature</td>
</tr>
<tr>
<td>ET3123-9</td>
<td>Neutrophils/100 leukocytes</td>
<td>1920-8 Aspartate aminotransferase</td>
</tr>
<tr>
<td>26511-6</td>
<td>Neutrophils</td>
<td>1742-6 Alanine aminotransferase</td>
</tr>
<tr>
<td>26444-0</td>
<td>Basophils [#volume] in Blood</td>
<td>26499-4 Neutrophils</td>
</tr>
</tbody>
</table>

*This work was done when the author was an intern at IBM Watson Health*
conditional event rates or hazards \( \lambda(t) \) associated with factors \( X_i = \{X_{i,1}, \ldots, X_{i,p}\} \) characterizes the corresponding risk of outcomes which, for a patient \( i \), can be given as:

\[
\lambda(t | X_i) = \lambda_0(t) \exp(\sum_{j} \beta_j X_{i,j}) = \lambda_0(t) \exp(X_i \cdot \beta)
\]

where, \( \lambda_0(t) \) denotes the baseline hazard at time \( t \) (independent of patient covariates) and \( \beta_j \) denotes the regression weight of the factor \( j \). To identify and characterize such factors from the data, we use L1 regularized survival models which have been proven to be effective in other domains\(^3\). Specifically, we find these parameters by fitting the regularized log-likelihood ratio (eq. 1) against normalized EMR history of the patients in the cohort.

\[
\ell(\beta) = \sum_{j} \left( \sum_{i \in H_j} X_i \cdot \beta - \sum_{t=0}^{m-1} \log \left( \sum_{i:Y_i \geq t_j} \theta_i - \frac{\ell}{m} \sum_{i \in H_j} \theta_i \right) \right) + \gamma \|\beta\|_1
\]

where, \( \gamma \) is the L1 penalty factor and we identify the factors which are (a) statistically significant for survival outcomes and (b) associated with the largest regression weights (\( \beta(> 0) \)).

**Renal Failure Prediction:** We validate the data-driven factors by using these as covariates to predict for renal failure in diabetic type 2 patients based on their medical histories. Although survival models are not particularly suited for prediction on time-intervals outside the training data, we note that the partial log-likelihood in CoxPH models have close ties to logistic GLM regressions\(^4\). We use a regularized logistic regression (eq. 2) to use these factors to predict the occurrence of renal failure in the prediction period (the next 3 months).

\[
\ell_{logistic} = \min_{w,d} \|w\|_1 + D \sum_{i=1}^{n} \log \left( \exp(-y_i(X_i^T w + d)) + 1 \right)
\]

### Results

Table 1 tabulates the top statistically significant data-driven factors w.r.t. survival as identified via eq. 1 from our diabetic type 2 patient corpus. Table 1 shows that our method uncovers several observational attributes, *from the data* and *without expert input*, that are potentially correlated with an increased risk of renal failures. We validate the importance of these factors using our predictive model as shown in Table 2 which shows that risk prediction performance increases when expert factors are supplemented by data-driven factors - thus indicating that the data-driven factors may be uncovering hidden correlations and thus, of potential interest to experts. Here, we found the L1 penalty using 5 fold cross validation and for each observation period, we used 10 fold cross-validation to quantify accuracy, sensitivity, and specificity.

### Discussion

Our experiments indicate that we are able to identify relevant factors for renal failures from EMR attributes that are often collected as part of regular patient encounters. These factors can be investigated by experts in a more clinical setting and potentially lead to the early prediction of renal failures and/or implement preventive measures.

### References


Using Electronic Health Records Data for Comparative Effectiveness Research

David Cheng, AM1, Ashwin Ananthakrishnan, MPH, MD2, Tianxi Cai, ScD1,3
1Harvard T.H. Chan School of Public Health, Boston MA; 2Massachusetts General Hospital, Boston, MA; 3Harvard Medical School, Boston MA

Introduction

The development of large electronic health record (EHR) research platforms has expanded opportunities for biomedical research. EHRs are becoming a valuable source for conducting comparative effectiveness research to assess treatment strategies and identify predictors that can explain treatment effect heterogeneity (TEH) for precision medicine. Conducting such analyses in EHR data, however, is challenging. One well-known issue is that the treatment assignment may depend on many confounders. Another primary challenge is the lack of direct observation on a pre-specified treatment response outcome. Ascertaining the outcome requires labor-intensive manual chart review, prompting the need for phenotyping algorithms that impute the outcome given available data. However, naïve use of imputed outcomes may lead to biased results especially when the quality of imputation is unclear. To address these issues, we developed robust methods that can infer the average causal treatment effect (ATE) and TEH in the EHR data setting, leveraging natural language processing (NLP) and semi-supervised learning (SSL) methods.

Methods

Algorithm: We frame the problem of estimating the ATE and identifying predictors of TEH with EHR data in an SSL setting, where a small set of patients are labeled with the true treatment response outcome and a large set of patients do not have their true outcome directly observed. Through NLP we extracted clinical terms from physicians’ notes that are relevant to the description of the treatment response outcome and developed a flexible imputation model for the outcome. Additionally, we estimated the propensity score to adjust for treatment by indication bias through a novel method based on modeling both the treatment assignment and the mean outcome given patient characteristics prior to treatment initiation. This novel propensity score estimator is more efficient and robust to mis-specification of its underlying models than traditional methods. Throughout we estimate model parameters using efficient shrinkage estimators that can accommodate high-dimensional covariates. Finally, we developed a semi-supervised response informed calibrated propensity score (RiCaPS-SS) estimator of the ATE and TEH using these imputations and propensity score estimates. Our algorithm results in estimators that are robust to mis-specification of the imputation model so that information from surrogate variables can be safely leveraged to improve the efficiency in estimating the ATE and TEH. The RiCaPS-SS estimators are also doubly-robust in that they will be valid provided that either the initial propensity score model or baseline outcome model is correctly specified. They are expected to be substantially more efficient than both complete-case (CC) estimators that neglect the unlabeled data and related missing data/causal inference estimators that we adapt to this setting to make use of the unlabeled data.

Data and Metrics for Evaluation: We evaluate the performance of our RiCaPS-SS algorithm based on simulation studies and an application to an EHR cohort of patients with inflammatory bowel disease (IBD) from Partner’s Healthcare. We designed the simulations to allow for mis-specification of either the propensity score model or baseline outcome model. The imputation model is mis-specified throughout. The relative efficiency, defined as the ratio of the mean-square error for the standard CC doubly-robust estimator (DR-CC) relative to each estimator, was used to compare the performance of our estimator to existing methods.

The EHR cohort included eligible patients with at least one ICD-9 code for Crohn’s disease (555.x) or ulcerative colitis (556.x) who were classified as IBD patients by a validated algorithm. Patients were also required to have at least 1 codified mention and 3 narrative mentions of adalimumab (ADA) or infliximab (IFX), with an interval of more than 0 days between the first and last narrative mentions. The final cohort consisted of 1,243 IBD patients, including 200 patients initiating treatment with ADA and 1,043 with IFX. The study aimed to compare ADA and IFX with respect to treatment response within 1 year of initiation. Through chart review of EHR records, the true treatment response status (responder vs none) was annotated for 117 randomly selected patients. We considered 35 post-treatment surrogates comprised of NLP mentions of clinically relevant terms that predict treatment response and adjusted for 12 baseline covariates including demographics, comorbidities, and prior utilization in the propensity score to mitigate confounding. We estimated the ATE using various estimators and their standard errors, relative efficiency compared to DR-CC, and confidence intervals (CI). We also performed analysis to identify predictors of TEH.
Results

Simulation: Simulations demonstrate that our proposed RiCaPS-SS algorithm produces an estimator that is substantially more efficient than existing estimators (Figure 1). We find that the efficiency gains hold regardless of the model mis-specification scenarios considered. RiCaPS-SS was found to be more than 50% more efficient than other estimators when both the propensity score and baseline outcome models are correct. When either the propensity score or baseline outcome models is mis-specified, the advantage of the RiCaPS-SS estimator is even more dramatic since the performance of the other estimators deteriorate due to mis-specification.

EHR IBD Study: The estimated ATE, defined as the difference in treatment response rate between ADA and IFX, from various approaches for the IBD study are shown in Table 1. The ATE comparing ADA vs. IFX is estimated to be -6.7% (95% CI -14.5% to -1.0%, p=0.039) using our approach. In this data our approach is estimated to achieve more than 220% efficiency gain over CC estimators 320% gain over missing data/causal inference estimators we adapt to leverage the unlabeled data. Our analyses also suggest that the Charlson comorbidity score is predictive of TEH between ADA vs. IFX. The difference in the adjusted log-odds ratio for ADA vs. IFX for those scoring higher by 1 Charlson score was -0.24 (95% CI -0.47 to 0.00), indicating that ADA should be recommended to patients with a lower score.

Table 1. Estimates of the ATE of ADA vs. IFX on treatment response in the EHR IBD cohort.

<table>
<thead>
<tr>
<th>Estimator</th>
<th>Estimate</th>
<th>SE</th>
<th>95% CI</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>None-CC</td>
<td>0.014</td>
<td>0.099</td>
<td>(-0.182, 0.196)</td>
<td>0.889</td>
</tr>
<tr>
<td>IPW-CC</td>
<td>-0.227</td>
<td>0.324</td>
<td>(-0.546, 0.178)</td>
<td>0.483</td>
</tr>
<tr>
<td>REG-CC</td>
<td>-0.067</td>
<td>0.132</td>
<td>(-0.274, 0.171)</td>
<td>0.609</td>
</tr>
<tr>
<td>DR-CC</td>
<td>-0.125</td>
<td>0.166</td>
<td>(-0.394, 0.179)</td>
<td>0.449</td>
</tr>
<tr>
<td>Naïve-SS</td>
<td>-0.051</td>
<td>0.088</td>
<td>(-0.313, 0.065)</td>
<td>0.565</td>
</tr>
<tr>
<td>PrePost-SS</td>
<td>0.033</td>
<td>0.110</td>
<td>(-0.244, 0.183)</td>
<td>0.762</td>
</tr>
<tr>
<td>RiCaPS-SS</td>
<td>-0.067</td>
<td>0.033</td>
<td>(-0.145, -0.01)</td>
<td>0.039</td>
</tr>
</tbody>
</table>

Discussion

In EHR data when the treatment response outcome is available on a small fraction of the observations, it is possible to leverage surrogate variables from unlabeled data to obtain more efficient estimators of the ATE and TEH. We develop an imputation approach coupled with a novel propensity score estimate that yields estimators of the causal treatment effect that are robust to mis-specification of the models imposed for the imputation and propensity score estimates. We find in simulations and data analysis that substantial efficiency gains are achieved over standard approaches. Even if both propensity score and baseline outcome model are not exactly correct, previous simulations indicated that similar results to hold as long as both models are not grossly mis-specified. This efficiency gain translates to being able to estimate treatment effects with much greater precision, or, alternatively, requiring potentially far fewer labeled observations to achieve a desired precision. Such a method could help facilitate comparative effectiveness research in EHR data when labeling patient records with a gold standard outcome is resource-intensive.

References

Patient and physician predictors of patient receipt of empiric therapies recommended by a computerized decision support system: a cohort study

Angela Chow, MBBS, MPH, MS, PhD1, David C.B. Lye, MBBS, FRACP1, Onyebuchi A. Arah, MD, MSc, DSc, MPH, PhD2

1Tan Tock Seng Hospital, Singapore; 2University of California Los Angeles, Los Angeles, California

Citation of work published in peer-reviewed journals in the last year:

Introduction

Antimicrobial resistance is a serious threat to public health and antibiotic use is the key driver. Antibiotic use in hospitals is high and on the rise, with 41-91% of all antibiotics prescribed in hospitals worldwide considered inappropriate. Antibiotic computerized decision support systems (CDSS) have been developed to improve antibiotic decision making through the accessibility of patient-specific clinical data and local antibiotic guidelines, at the point of prescribing.

Antibiotic CDSS are particularly useful for antibiotic selection for empiric therapy, as optimal selection is complex when the causative pathogen is unknown. Antibiotic CDSSs have been shown to improve antibiotic prescribing and patient clinical outcomes including the reduction of mortality. However, there is limited information on factors influencing physicians' acceptance or patients' receipt of CDSS-recommended antibiotics. Understanding these factors can guide strategies to improve patients' receipt of antibiotic therapies recommended by CDSSs and enhance clinical care.

Objective

We conducted a prospective cohort study to evaluate the extent to which hospitalized patients received antibiotics as recommended by our in-house antibiotic CDSS, Antimicrobial Resistance Utilization and Surveillance Control (ARUSC), and to identify patient and physician factors associated with patients’ receipt of empiric antibiotic therapies recommended by ARUSC when the prescribing physician had an initial preference for using broad-spectrum antibiotics and targets for improvement.

Methods

The study was conducted in a 1500-bed tertiary-care academic medical center in Singapore.

In 2009, the hospital launched its in-house antibiotic CDSS, ARUSC, which integrates antimicrobial stewardship with the hospital's computerized physician order entry system (CPOE) and provides patient-specific evidence-based antibiotic recommendations at the point of prescribing (Figures 1A-D). All medication orders in the hospital are made via the CPOE (Figure 1A). From September 12, 2011, whenever a physician makes an electronic prescription of piperacillin-tazobactam or a carbapenem for an inpatient, the prescription automatically triggers the launch of ARUSC (Figure 1B). Piperacillin-tazobactam and carbapenems are antibiotics of last resort for many bacterial infections, particularly those caused by multidrug-resistant pathogens. Hence, it is crucial to ensure the judicious use of these antibiotics.

Using a rules-based algorithm, ARUSC provides guidance on antibiotic selection and dosing, based on guidelines developed by the hospital's antimicrobial stewardship committee, which recommends the narrowest-spectrum antibiotic appropriate for common organisms responsible for the diagnosed infection taking into account the local epidemiology of infectious diseases, local microbiology and antibiotic susceptibility patterns in the hospital in the prior 5 years, and incorporating evidence-based international guidelines including the Infectious Diseases Society of America’s Practice Guidelines and Australia’s Therapeutic Guidelines: Antibiotics. Data from individual patients’ electronic medical records including medication history and drug allergies, as well as laboratory results such as creatinine levels are also pulled into ARUSC and included in the algorithm. ARUSC recommends the most appropriate antibiotic for the patient, taking into account the patient’s antibiotic allergies and renal function (Figure 1C). The prescribing physician can either accept or reject ARUSC-recommended antibiotics, which are assumed to be always appropriate (Figure 1D).

All patients admitted to the hospital, from October 1, 2011 through September 30, 2012, who were prescribed piperacillin-tazobactam or a carbapenem for empiric therapy and auto-triggered to receive antibiotic therapies recommended by ARUSC were included in the prospective observational cohort study. Prescriptions for prophylactic or definitive therapy were excluded. We chose to focus our study on empiric therapy, as empiric antibiotic prescriptions have been found to be the least concordant with recommended antibiotic guidelines. Furthermore, empiric antibiotics are usually the first antibiotics received by a patient in an infective episode; appropriate empiric antibiotics is a critical determinant of clinical outcome. Patients were followed up prospectively from the automatically-triggered launch of ARUSC at the point of antibiotic prescribing up to 30 days post-discharge from hospital.
Results

One-quarter of the 1886 patients received CDSS-recommended antibiotics. More patients treated for pneumonia (33.2%) than sepsis (12.1%) and urinary tract infection (7.1%) received CDSS-recommended antibiotic therapies. The prescribing physician—but not the attending physician or clinical specialty—accounted for some (13.3%) of the variation (Table 1). Prior hospitalization (OR 1.32, 95% CI 1.01-1.71), presumed pneumonia (OR 6.77, 95% CI 3.28-13.99), intensive care unit (ICU) admission (OR 0.38, 95% CI 0.21-0.66), and renal impairment (OR 0.70, 95% CI 0.52-0.93) were factors associated with patients’ receipt of CDSS-recommended antibiotic therapies (Table 1). 60% of patients who were auto-triggered into the CDSS due to a prescription of piperacillin-tazobactam were consequently recommended piperacillin-tazobactam by the CDSS, but not all received antibiotic therapy in the dose and frequency recommended by the CDSS. In the subgroup analysis of 1213 pneumonia patients, prior hospitalization (OR 1.37, 95% CI 1.01-1.86), ICU admission (OR 0.39, 95%CI 0.20-0.74), renal impairment (OR 0.73, 95% CI 0.52-1.02), and prescription made at night (OR 1.35, 95% CI 1.02-1.79) were associated with receipt of CDSS-recommendations. For the 673 non-pneumonia patients, ICU admission (OR 0.27, 95%CI 0.06-1.22, P=0.088) and renal impairment (OR 0.54, 95%CI 0.29-1.04, P=0.065) were marginally associated.

Conclusion: This study gave insights into predictors of patients’ receipt of empiric antibiotic therapies recommended by a CDSS when the prescribing physician had an initial preference for using broad-spectrum antibiotics. Patients admitted to the ICU or who had renal impairment were less likely to receive CDSS-recommended antibiotics. The clinical complexity of ICU and renal-impaired patients had required physicians to override CDSS-recommendations which were more tailored for general inpatients. Enhancements to the antibiotic CDSS can help address some of the unique patient needs, but the more complex clinical conditions and antibiotic needs of such patients may still require a physician’s assessment in addition to the CDSS recommendations.
Opportunities and Challenges for an Interdisciplinary Team to Guide Adoption of Technology to Dissipate Threats to Patient Safety in Real-Time

Anuj K Dalal, MD,1,2 Theresa E Fuller,1 Pamela Neri, MS,1 Dominic J Breuer, MS,3 David W Bates, MD, MSc1,2 and James C Benneyan, PhD3

1Brigham and Women’s Hospital, Boston, MA, 2Harvard Medical School, Boston, MA 3Northeastern Healthcare Systems Engineering Institute

Introduction: Minimizing preventable harm in the acute care setting is an ongoing challenge, but one that is necessary to create a safe healthcare system.1 Implementing health information technology (HIT) with the goal of improving patient safety is especially challenging in complex healthcare settings.2 Collaboration across disciplines, departments, and institutions is necessary to ensure successful implementation and sustainability but is often lacking. We assembled an interdisciplinary team with expertise in systems engineering, human factors and usability, and data analytics as part of a large AHRQ-funded Patient Safety Learning Laboratory (PSLL) initiative between Brigham and Women’s Hospital and Northeastern University. An important goal of our team was to facilitate institutional adoption of the HIT components (Figure 1: patient portal, screen-savers, safety dashboard) as a unified system integrated with our vendor-based electronic health record (EHR). In this study, we describe opportunities and challenges encountered for implementing novel HIT to identify, assess, and mitigate threats to patient safety in real-time.

Methods: This study took place at a large academic medical center in Boston, Massachusetts. We employed a sociotechnical approach to identify opportunities and challenges for facilitating adoption and sustainability of our technology in four categories: stakeholder engagement, organizational priorities, user experience, and technical and data integration.2 Institutional stakeholders were engaged in a series of meetings and presentations to ensure alignment with organizational priorities and initiatives. Our team worked with information system and quality and safety leaders to identify requirements for technical and data integration. We conducted interviews and observations of patients and clinicians to identify usability challenges. We identified key themes using a 2-person consensus approach.

Results: Key opportunities and challenges are presented in Table 1. In general, we found high degree of support for using screen-savers and dashboards for real-time visualization of data from the EHR as this aligned with the overarching institutional goal of reducing preventable harms due to hospital-acquired conditions (catheter-associated urinary tract infection). Specifically, these tools were seen as potent interventions for improving quality and safety process measures (nursing-driven Foley protocol ordering) reported to clinical dyads as part of a hospital-wide initiative when dashboard use data were provided simultaneously. Ensuring consistent interpretation of data displayed by various clinicians (“red” indicating action is needed) and appropriate follow-up action (updating documentation, entering orders in EHR) were important challenges identified by usability and human factors experts. Identifying potential failure modes (failure modes and effects analysis) and actual safety events (failure-based root-cause analysis) were methods employed by systems engineers in collaboration with safety leadership, and represent a crucial opportunity to encourage learning that would be used to iteratively refine HIT components. The user experience was enhanced by technical integration of safety icons into screensavers displayed on bedside workstations to engage both patients and clinicians in identifying and understanding safety risks. For clinicians, the user experience was also enhanced by integrating a research and innovations portal into the EHR on which information from the EHR (CAM score, sedatives, QTc) and appropriate action pertinent to mitigate a specific harm (delirium) were displayed. Finally, use of the acute care patient portal by patients and caregivers was improved considerably by enabling a “bring-your-own-device” strategy, but could be further enhanced by integration with the enterprise patient portal and by optimizing the enrollment process.

Conclusion: Overall, the experience of our team underscores the key opportunities and real-life challenges when implementing novel HIT to identify, assess, and mitigate threats to patient safety. Clinicians and institutional stakeholders were generally enthusiastic about using several HIT components (safety screen-savers, safety dashboard) because of clear alignment with institutional priorities and initiatives to improve patient safety.

Acknowledgements: The PSLL team is supported by the AHRQ (P30-HS23535): B. Couture, J. Espares, M. Duckworth, K. Schnock, A. Businger, E. Gershanik, S. Collins, R. Rozenblum, P. Dykes, and J. Schnipper.

Table 1. Opportunities and Challenges for Leveraging Expertise in Systems Engineering, Human Factors and Usability, and Data Analytics to Implement Novel Technology to Mitigate Safety Threats

<table>
<thead>
<tr>
<th>Category</th>
<th>Opportunities</th>
<th>Challenges</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Stakeholder Engagement</strong></td>
<td>• Support clinicians through quick visualization of “siloed” EHR data</td>
<td>• Ensuring consistent interpretation of information displayed visually on dashboard and screen savers</td>
<td>• Use simple color scheme to identify actionable items on safety dashboard (red, yellow, green)</td>
</tr>
<tr>
<td></td>
<td>• Help administrative and clinical leadership realize institutional quality and safety goals</td>
<td>• Selecting metrics that mutually benefit stakeholders; identifying safety events and near-misses in real-time</td>
<td>• Provide clinical-unit leadership with reports that demonstrate how compliance with HIT use improves quality safety process measures</td>
</tr>
<tr>
<td></td>
<td>• Align with mission of patient and family advisory councils to promote patient-centered care</td>
<td>• Encouraging use of patient portal by care partners during hospitalization</td>
<td>• Provide tools for patients to report safety events via patient portal</td>
</tr>
<tr>
<td></td>
<td>• Facilitate vision of information system leadership to support digital health research and innovation</td>
<td>• Navigating governance for digital health research and innovation</td>
<td>• Perform failure-based RCAs when safety events are identified</td>
</tr>
<tr>
<td><strong>Organizational Priorities and Initiatives</strong></td>
<td>• Align with hospital quality and safety reporting initiatives for hospital acquired conditions (HACs) and opioid response/education teams</td>
<td>• Assessing accuracy of quality and safety reports from new EHR</td>
<td>• Enable user-friendly displays to quickly identify patients at risk for harm (red = “actionable”)</td>
</tr>
<tr>
<td></td>
<td>• Address quality and safety gaps after vendor EHR implementation</td>
<td>• Identifying HACs with validated and available process measures</td>
<td>• Demonstrate improvement in process measures for falls, CAUTI, CLABSI, VTE prophylaxis, etc.*</td>
</tr>
<tr>
<td></td>
<td>• Foster culture of patient safety reporting and transparency</td>
<td>• Ensuring new safety initiatives do not adversely impact efforts at delivering patient-centered care and improving the patient experience</td>
<td>• Perform FMEA with patient safety leadership during roll-out to identify potential failures</td>
</tr>
<tr>
<td></td>
<td>• Address clinician burn-out while moving to all-electronic workflows</td>
<td>• Assessing cognitive load in actual vs simulated clinical settings</td>
<td>• Align with mission of patient and safety goals</td>
</tr>
<tr>
<td><strong>User Experience</strong></td>
<td>• Allow patients to use own device</td>
<td>• Enabling access to acute care patient portal from enterprise patient portal</td>
<td>• Show any harm (red = “actionable”)</td>
</tr>
<tr>
<td></td>
<td>• Enable tools from frequently used EHR access points (start menu, patient-chart)</td>
<td>• Ensuring clinicians are aware how to access tools from key starting points</td>
<td>• Optimize loading times for rapid access to dashboard on rounds</td>
</tr>
<tr>
<td></td>
<td>• Improve fidelity of EHR data entry via visual feedback to providers of data displayed on patient portal, safety screensavers, dashboard</td>
<td>• Ensuring information is displayed accurately by encouraging proper data entry in EHR</td>
<td>• Pin dashboard application to start menu for quick access</td>
</tr>
<tr>
<td></td>
<td>• Address cognitive burden by reducing number of clicks to access information necessary to assess risk</td>
<td>• Tailoring dashboard logic to different workflows and clinical services, addressing color blindness when using color coding schemes</td>
<td>• Glasses icon is displayed on screen saver when item is appropriately updated in nursing documentation</td>
</tr>
<tr>
<td><strong>Technical and Data Integration</strong></td>
<td>• Access to EHR data and services</td>
<td>• Obtaining timely approvals for data feeds, services, and access points</td>
<td>• Obtain CAM score from EHR flowsheet; obtain sedative medication from eMAR service</td>
</tr>
<tr>
<td></td>
<td>• Integration of patient and provider-facing technologies with enterprise information systems (patient portal, EHR, safety reporting systems)</td>
<td>• Prioritizing requests for services, flowsheet items, reports in enterprise EHR demand queue</td>
<td>• Integration of a research and innovations portal into EHR to enable rapid piloting of digital health applications (dashboard)</td>
</tr>
<tr>
<td></td>
<td>• Access to hospital administrative databases: nursing data, quality and safety dashboards, ADT feeds to enhance understanding of unit-risk</td>
<td>• Navigating governance to access institutional data and integrate digital health applications with EHR</td>
<td>• Timing and sensitivity of requesting nursing data due to impending strike</td>
</tr>
<tr>
<td></td>
<td>• Development of use and process database for dashboard</td>
<td>• Ensuring that institutional data are available at desired frequency</td>
<td>• Access to HAC dashboard to identify harms reported by hospital</td>
</tr>
</tbody>
</table>

*CAUTI = Catheter-associated urinary tract infection; CLABSI = Central-line associated blood stream infection; VTE = Veno-thromboembolic disease; FMEA = Failure modes and effects analysis; RCA = Root-cause analysis; CAM = Confusion assessment method

Figure 1. Technological Components of Patient Safety Learning Laboratory

The Systems Engineering, Usability and Human Factors, and Integrative Data-Analytics (SEU)(3) core was involved with a continuous safety improvement process including surveillance to identify at-risk patients, individualized risk assessment of threats to identified patients, mitigation strategies to reduce harm likelihood, and systematic analysis of threats and harms identified in the first 3 phases in order to narratively refine the individual HIT components and overall “system” over time.
Use of Clinical Phenotypes and Non-negative Tensor Factorization for Heart Failure Prediction

Yu Deng, BS1; Al’ona Furmanchuk, PhD1; Robert Chen, BS2; Faraz S. Ahmad, MD1; Jimeng Sun, PhD2; Abel Kho, MD1

1Feinberg School of Medicine, Northwestern University, Chicago, IL
2College of Computing, Georgia Institute of Technology, Atlanta, GA

Introduction
Heart failure (HF) is a diverse syndrome associated with multiple risk factors and diseases. Heart failure affec ts millions of adults in the United States annually, and for some subtypes of patients with HF, 5-year mortality is higher than 50%. In order to intervene earlier to reduce morbidity and mortality with targeted therapies it is important to identify patients at high risk for HF. Electronic health records (EHR) provide extensive information on patients that one can use as features for outcome prediction. Many machine learning (ML) methods are able to deal with large feature sets and complete the prediction task successfully. However, there are few efficient algorithms that also yield results that are easily interpretable. Furthermore, when applied to EHR data, such methods (e.g. logistic regression) will struggle from over-fitting due to low number of events per variable. Recent applications of non-negative tensor factorization (NTF) to EHR records show potential compromise between sparseness of EHR data and ease of comprehension of the model. Here, we assessed the performance of a novel ML method for detecting patients at risk of HF. In the first step, the NTF algorithm generates phenotypes of patients specific to patient diagnosis and pharmacological profiles prior to HF. Each patient is assigned to its own “fingerprint” label according to his membership to the different phenotypes. The fingerprints were then used as a feature set to predict HF. Ease of comprehension is achieved through transforming the high dimensional space (raw features) to a reduced clinical phenotype space (“fingerprints”).

Methods
We extracted medication (RxNorm codes) and diagnosis (ICD9 codes) from 1/1/2005 to 11/30/2016 on patients with possible HF from the Northwestern Medicine Enterprise Data Warehouse (EDW). We mapped diagnoses codes into a higher level of hierarchy based on PheWAS and Anatomical Therapeutic Chemical Classification System. We further reduced the study population to HF cases and controls only.

We identified a HF case if he/she has (i) at least 2 HF diagnoses from outpatient encounters; or (ii) at least 2 medications prescribed for a HF diagnosis; or (iii) at least 1 HF diagnosis from an outpatient encounter and at least 1 medication prescribed for a HF diagnosis. This definition was previously validated by Geisinger Clinic. For each case, we matched control patients on gender, year of birth and similar encounter information. We analyzed records of each patient within a 2-year time window prior to HF onset date to ensure that each patient had the same length of prediction window. Only cases and controls having a full 2-year period of records and at least 1 visit during these 2 years were kept in the final sample. Once a HF patient was identified, the first appearance of HF related medication or diagnosis was assigned as the onset date. A tensor was constructed with 3 modes including a patient mode, medication mode and diagnosis mode. Each tensor element represents the co-occurrence between medication and diagnosis for a certain patient during the available time window. NTF was applied on the tensor to generate patient phenotypes. The cutoff parameters (alpha: 1.0; gamma: 0.001, 0.08, 0.07) for NTF were set to reduce the noise of data and keep the phenotypes concise. A series of dummy variables were created for each patient based on phenotype membership. Those variables were then used as predictors for logistic regression (LR) and random forest (RF) prediction. The predictive performance of this method was compared to the performance of algorithm based on raw features transformed lower dimension space of principal components (PCA). The 3 main principal components were passed as predictors to RF algorithm. The area under receiver operator characteristic curves (AUC) were used to evaluate model discrimination ability (Fig 1).

Results and Discussion
We identified 518 subjects as qualified HF cases in the EHR. Patient onset age ranges from 20 to 89 years old with majority of HF onset age at 45-89 years old. For both RF and LR prediction, “fingerprint” labels outperformed raw features and principal components when the number of clinical phenotypes is greater than 20. The AUC score for raw feature/PCA based method is 0.628±0.145 (green solid and dashed lines in Fig. 1), while corresponding values for “fingerprint” –based methods are 0.733±0.05 (LR: red dots and lines in Fig. 1) and 0.760±0.136 (RF: blue crosses and lines in Fig.1). Based on how NTF handles noise of the data, number of cases and controls at the final
step (orange line in Fig.1) will be smaller than original number of patients selected for the study. Therefore, for small samples, accuracy of the final prediction will be significantly affected by the reduction of overall number of the patients. In our study, the case:control ratio is approximately 1:3. Even after adding weights to controls for imbalance adjustment, analysis will still suffer from reduction of HF case patients during NTF. We used the number of HF cases remained after NTF together with AUC value as criteria for selection of optimal number of phenotypes. It appears that 90-phenotypes is the optimal number (both classification performance and number of patients reach a plateau). We sorted phenotypes importance based on use of features in RF prediction. The 5 phenotypes out of top 12 were selected as the most significant risk phenotypes (table 1) for developing HF event. 66.73% of future HF case patients are in those five phenotypes, while contribution of these phenotypes to controls is only 16.24%.

Table 1. Contribution to HF cases and controls for each phenotype

<table>
<thead>
<tr>
<th>Phenotype label and description</th>
<th>Phenotype contribution to cases (case # in a phenotype/# of all cases)</th>
<th>Phenotype contribution to controls, %,(control # in each phenotype/# of all controls)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Atrial fibrillation/flutter &amp; edema</td>
<td>17.67</td>
<td>5.71</td>
</tr>
<tr>
<td>2. Lipoid disorders &amp; hypertension</td>
<td>16.63</td>
<td>6</td>
</tr>
<tr>
<td>3. Hypertension, ischemic heart disease</td>
<td>12.68</td>
<td>1.9</td>
</tr>
<tr>
<td>4. Type 2 diabetes mellitus</td>
<td>12.27</td>
<td>1.9</td>
</tr>
<tr>
<td>5. Hypertension, lipoid disorder, ischemic heart disease, ill-defined complications of HF</td>
<td>7.48</td>
<td>0.73</td>
</tr>
</tbody>
</table>

Physician evaluation shows clinical meaningfulness of the 5 phenotypes that contributed the most to the risk of subsequent HF. The prominent comorbidities in the five phenotypes in decreasing order of contribution were atrial fibrillation/flutter, lipoid disorders, hypertension, and type 2 diabetes mellitus, with the fifth being an amalgam of lipoid metabolism, hypertension and ischemic heart disease. Clinically, this order may suggest that some acute conditions (e.g. atrial fibrillation) may contribute more to subsequent heart failure risk in the short term (2-year window) than chronic conditions which may take years to lead to subsequent cardiovascular injury.

Conclusions

Our preliminary results show better performance of NTF membership linked classification over other popular dimension reduction method and suggested a plausible hierarchy of HF risk based on associated comorbidities. NTF also shows potential in capturing novel HF prediction phenotypes (i.e. features) with distinct clinical characteristics. Such phenotypes are likely to improve HF prediction and help discover mechanisms of HF pathogenesis.

Acknowledgements

This research is supported by the National Science Foundation, under Grant IIS-1417819.

References

3. Organization WH. ATC/DDD Index 2017
Meeting User Needs for a Data Discovery Index of Biomedical Big Data

Ram Dixit, BS1, Deevakar Rogith, MBBS, PhD1, Vidya Narayana, MS1, Mandana Salimi, MS1, Anupama Gururaj, PhD1, Lucila Ohno-Machado, MD, PhD2, Hua Xu, PhD1, Todd Johnson, PhD1

1The University of Texas Health Science Center at Houston (UTHealth) School of Biomedical Informatics, Houston, TX; 2University of California San Diego Health System Department of Biomedical Informatics, La Jolla, CA

Introduction

‘Big data’ in biomedicine is a reality and an opportunity for new forms of research and knowledge discovery1. However, the diverse locations, formats, and types of data combined with a lack of standards for describing datasets makes taking advantage of this opportunity challenging2. The NIH Big Data to Knowledge (BD2K) initiative issued a call to assemble these heterogeneous data from multiple sources into a single Data Discovery Index (DDI)1. A prototype DDI, DataMed (accessible at http://DataMed.org), was launched with the goal of enabling researchers to search for, access, and cite datasets3. Bringing together datasets from multiple sources poses challenges in information retrieval and user interface design to providing meaningful and useful search results to the users4, 5. Studies have been done to understand user needs and information seeking behavior in literature and citation indexes such as PubMed6, 7, but there is a gap in knowledge on user needs, expectations, and information seeking behavior regarding biomedical data discovery. We present the initial findings from our user needs and usability evaluation of DataMed. We discuss the challenges in biomedical data discovery to inform future DDI development efforts.

Methods

To understand biomedical researchers’ needs and guide the development of the DataMed, we conducted user analyses in two phases. In the first phase, we focused on why and how researchers look for online biomedical datasets through semi-structured interviews (13 participants) that informed the initial development of DataMed. In the second phase, we conducted iterative “think-aloud” usability evaluations of versions 0.5 (8 participants) and 2.0 (11 participants) of DataMed. We present results from the qualitative analysis of interviews and usability studies with these researchers, that focused on four main areas: searching for data, data formats, data visualization, and metadata.

Results

We present the summary of key findings from the user analyses:

Table 1. Summary of user analyses findings and design decisions for DataMed – a Data Discovery Index

<table>
<thead>
<tr>
<th>Phase</th>
<th>Key Findings</th>
<th>Design Decision</th>
</tr>
</thead>
</table>
| I – User Needs Assessment     | • Users search for online data to find similar datasets they can compare with their own analyses or to find complementary (cross-domain) datasets that can enrich their analyses  
• Key metadata such as source, type, contents, and availability, are necessary to assess dataset relevance and utility.  
• Users need to know appropriate tools and methods to process a chosen dataset.  
• Visualization is important for navigating and interacting with large-scale data. | • Link dataset to similar datasets  
• Link dataset to publications citing and similar publications  
• Natural Language Processing of query and UMLS concept mapping to improve breadth of search  
• Incorporate Data Tag Suite (DATS) as core metadata model for indexing datasets  
• Visualization of timeline and word cloud |
| II – Usability Testing        | • Users liked that the search interface matches PubMed interactions.  
• While they appreciated aggregated results from multiple repositories, users needed more information about search results and detailed descriptions of individual datasets. | • Metadata was improved to DATS 2.0 with additional nodes.  
• Process for ingestion of datasets into a uniform metadata descriptor |
While DataMed supported retrieval of highly specific datasets, exploring general topics datasets or discovering new datasets continues to be challenging. Assessing search results proved to be difficult due to incomplete or inconsistent metadata across datasets. Ongoing study – Potential design changes:

- Improvement to metadata model
- Improving indexing based on user preferences and interactions with datasets.

Discussion

The results presented here highlight key issues of metadata, data standards, and visualization in searching for and using biomedical data. Researchers attempting to find data online face a complex environment with diverse sources, formats, and types of data that can be difficult to navigate and require significant time and effort to utilize. Standardized formats, vocabularies, and documentation for datasets need to be developed for user to easily find and assess relevant data. Further, visualization techniques have the potential to support navigation and analysis of datasets, but current methods do not support this kind of interaction at scale.

To support user needs in finding relevant data, DataMed attempts to resolve the disparity in data sources by defining a common metadata model and applying tools for information retrieval. DataMed thus satisfies the basic definition of a DDI and provides an intuitive model for searching datasets. However, our initial findings show that searching for biomedical data poses a similar, but distinct problem from searching for websites or publications. Researchers' assessment of the usefulness or relevance of a dataset is not possible from only from summaries such as keywords, title, or abstract. Initial usability testing revealed that while aggregating metadata into a common search engine is an appealing solution to this problem, the inconsistency and incompleteness of most dataset metadata limits its usefulness for users. Ongoing usability evaluations indicate common and essential information such as the samples collected, collection technique, data processing methods, and statistical summaries are necessary to determine whether a dataset is suitable for a research topic.

Conclusion

Simply making data available is not enough: we emphasize the importance of conducting user needs analyses and usability evaluations in developing data infrastructures to support biomedical research. Promising research areas for improving the discoverability of data include: refining metadata models for describing datasets, developing standardized and analytic summaries of datasets, and improving user interfaces to large amounts of data through visualization. Future efforts in developing a DDI can leverage these insights to provide researchers with efficient and usable interfaces to datasets.

Acknowledgements

DataMed is developed as a part of the bioCADDIE project and supported by the NIH through the NIH Big Data to Knowledge, Grant 1U24AI117966-01.

References

Leveraging Electronic Health Record Data for Community Health Assessment and Surveillance

Brian E. Dixon1,2, P. Joseph Gibson1,3, Karen Frederickson Comer4, Jian Zou5, Jennifer Williams2, Marc Rosenman2,6
1IU Fairbanks School of Public Health, Indianapolis, IN; 2Regenstrief Institute, Indianapolis, IN; 3Marion County Public Health Department, Indianapolis, IN; 4Indiana University-Purdue University Indianapolis, Polis Center, Indianapolis, IN; 5Worcester Polytechnic Institute, Worcester, MA; 6Feinberg School of Medicine, Northwestern University, Chicago, IL

Abstract

Increased adoption of electronic health records presents an opportunity to enhance assessment, a core function of public health, by enabling access to up-to-date data on disease burden and outcomes. We will present how we’ve leveraged routinely collected and geospatially-enhanced EHR data available from a regional health information exchange to augment assessment at a county health department. We will highlight neighborhood level indicators developed for the project, data visualizations, and issues of representativeness or bias in the data.

Introduction

Public health authorities monitor population health to identify burden of disease, manage health assets, establish policy, and evaluate interventions. This assessment usually relies on information available through surveys, vital records, and paper-based disease reporting. These sources capture sparse data at the community level. Electronic health record (EHR) systems may provide more timely and dense data for geographic areas of interest. Yet there exist a number of challenges to routine use of EHR data, including health information exchange (HIE) between clinical and public health organizations as well as lack of mechanisms to routinely capture social determinants.

Methods

Data from multiple EHR systems within the five major health systems serving the residents of Marion County, Indiana were extracted from the Indiana Network for Patient Care (INPC), one of the oldest and most comprehensive health information exchange (HIE) networks. (1) Founded by the Regenstrief Institute and managed by the Indiana Health Information Exchange, the INPC receives data from over 100 hospitals, several physician practices, laboratories, and other healthcare facilities representing over 12 million unique patients and 6 billion pieces of clinical data.

This study included data on living individuals who resided in Marion County and had at least two clinical encounters within the health systems participating in INPC during a 3-year period beginning January 1, 2011 and ending December 31, 2013. The INPC represents 80% of Indiana citizens, but it contains data on >90% of citizens in Marion County. Representativeness is based on the 2010 decennial census data from the U.S. Census Bureau.

The study focused on population health indicators of highest interest and typically not available to health departments at levels below the county. The indicators included items measuring prevalence (e.g., prevalence of diabetes, prevalence of asthma) as well as chronic disease management (e.g., control of HbA1c, lipid control), and they were largely based on HEDIS (Healthcare Effectiveness Data and Information Set) measures that are routinely reported by health systems to payers and employers.

The denominator for prevalence measures represented the number of living individuals who received care from INPC providers during the three year time period. Health management denominators represented the number of individuals diagnosed with the disease as documented in the INPC. Numerators were defined using a combination of ICD9-CM clinical diagnostic codes as well as laboratory results and medication history data from longitudinal records in INPC.

Indicators were calculated at these geographic levels: local health department planning area (average population 40,000-50,000), ZIP code (average pop 8,000), geopolitical neighborhood (average pop 6,000), census tract (average pop 4,000), and census block group (average pop 1,500). Once calculated, indicator data were mapped using ArcGIS (ESRI, Redlands, CA) to visualize the data for local health department personnel.
Results
We generated 10 different population health indicators using data integrated from multiple EHR systems. Indicators ranged from the proportion of individuals with depression or a sexually transmitted infection to the proportion of patients with cardiovascular disease who had poor lipid control.

In Figure 1, we present a map of Marion County, Indiana, with the prevalence of diabetes by census tract. In Figure 2, we present a map of Marion County, Indiana, with the rates of diabetes control as measured by glycosylated hemoglobin (HbA1c) levels below 8.0% by those with a confirmed diagnosis of diabetes by census tract. Individuals at the Marion County Public Health Department, when viewing the maps and data depicted in Figures 1 and 2, stated that the availability of population data at geospatial levels smaller than a county is more insightful than examining the county wide rate of diabetes. Granularity was perceived as useful, because health department staff believe they may be able to target diabetes education or prevention efforts in the areas shaded in dark blue where there are either high rates of diabetes (Figure 1) or low rates of diabetes management (Figure 2) within a census tract.

During our project, we discovered that calculating rates at the census tract level or larger (e.g., average populations > 4,000 persons) is preferable. Rates at the census block level were more challenging to interpret as many blocks contained a count of zero or one. This makes comparing rates difficult, and it would be hard for health departments to release small numbered cells given concerns for privacy and confidentiality of health data.

Discussion
EHR systems include more people with more frequent contacts than do existing, standard public health data sources like surveys and death certificates. Whereas the Behavior Risk Factor Surveillance System’s estimated prevalence of diabetes is based on 500 individuals, the INPC contained information on 559,905 adults with an average of 2,500 persons per census tract. Therefore data from EHR systems could allow more frequent measurements of disease incidence and outcomes as well as more precise (tighter confidence intervals) identification of disparities.

Yet data from EHR systems also have limitations that need to be understood to fully interpret their meaning when influencing population health decisions. For example, data captured by EHR systems largely represents “sick care” as opposed to health as many people do not utilize health care facilities every year. Furthermore, some health care is provided outside of existing HIE networks, including but not limited to urgent care clinics independent of large health systems and public health clinics that are often not connected to the HIE. These additional data sources, as well as the Census Bureau and community organizations, may have unique, social determinant data that are often missing from EHR systems. Strategic partnerships and HIE models, as well as evaluation of these models, will be necessary to integrate and analyze data for population surveillance as well as community health assessment processes.

Acknowledgements
This project was supported by the Robert Wood Johnson Foundation (No. 71271). The podium presentation at AMIA is based on the following article published within the past year: Dixon BE, Zou J, Comer KF, Rosenman M, Craig JL, Gibson PJ. Using electronic health record data to improve community health assessment. Front Public Health Serv Sys Res. 2016;5(5):50-6. The presentation will feature updated results as well as next steps in our research.

References
Predictors of OpenNotes use by Veterans receiving Mental Health Care

Steven K. Dobscha\textsuperscript{1,2}, Lauren M. Denneson PhD\textsuperscript{1,2}, Maura K. Pisciotta MS\textsuperscript{1}, Donald S. Bourne MPH\textsuperscript{2}, David Philipps-Moses BS\textsuperscript{1}, Susan Woods MD, MPH\textsuperscript{3}

\textsuperscript{1}VA Center to Improve Veteran Involvement in Care, Portland, Oregon; \textsuperscript{2}Oregon Health & Science University; \textsuperscript{3}Center for Outcomes Research & Evaluation, Maine Medical Center, Portland, Maine

Introduction

Several healthcare systems across the U.S., including the Veterans Health Administration (VHA), offer patients online access to their clinical notes. Clinical notes are progress notes clinicians and other staff write in patients’ medical records to document healthcare visits. Online patient access to clinical notes, also referred to as OpenNotes, is facilitated by electronic patient health portals. Patient access to clinical notes has been shown to facilitate patient engagement and to strengthen the patient-clinician relationship (1,2). Some individuals have raised concerns about potential negative consequences of patients reading mental health clinical notes (3). However, recent research suggests that reading clinical notes online can strengthen patients’ trust in mental health clinicians and improve the therapeutic relationship (4). Despite these potential benefits, it is unknown to what extent, and which, patients with mental health conditions read their clinical notes online. The objective of this study was to identify demographic and clinical predictors of utilizing OpenNotes among veterans receiving mental health care.

Methods

This study was approved by the Institutional Review Board of the VHA Medical Center (VAMC) where the study took place. The VAMC is located in the Pacific Northwest. Approximately 200 mental health clinicians provide mental health services to 18,000 unique veterans across a range of programs and sites. We recruited and enrolled participants from February to July of 2016. Veterans were eligible to participate if they had 1) completed at least one in-person mental health appointment at the VAMC; and 2) logged into VA’s electronic health portal, MyHealtheVet, at least twice in the 6 months prior to recruitment. Exclusion criteria included: dementia or cognitive disorder diagnosis in the prior year, having a legal guardian, or recent discontinuation of mental health care at the VAMC. We mailed recruitment letters to 3,381 potentially eligible patients, and 598 indicated interest in participating. After describing the study and confirming eligibility, 407 participants were consented and enrolled. Fourteen participants withdrew prior to being sent questionnaires, resulting in 393 total participants. Participants were invited to complete a self-report questionnaire which assessed their use of, experiences with, and opinions regarding OpenNotes. The questionnaire included the eHealth Literacy Scale (eHEALS) (5) and several items adapted with permission from the Robert Wood Johnson Foundation’s OpenNotes project (1,6,7). We also extracted diagnostic and utilization data for the year prior to the study from VHA’s Patient Care Database. In analyses, multivariable logistic regression and tests of trend were used to examine relationships between self-report of having read OpenNotes and demographic and clinical characteristics; EHEALS score; patient-clinician communication about OpenNotes; confidence in understanding of the purposes and uses of Blue Button, and level of importance to the veteran of using OpenNotes.

Results

Of 393 participants who received questionnaires in August 2016, 359 (91%) returned questionnaires. Three hundred thirty-eight (94%) had complete sets of responses. There were no statistical differences in demographics when comparing respondents and those who failed to return questionnaires. The average age of respondents was 57.8 (sd=12.3); 68% were male; 55% married; 89% white; 56% not working; 9% with a high school education or less; and 32% lived in rural areas. In the prior year, 58% were given a diagnosis of post-traumatic stress disorder (PTSD), 60% were given a depression diagnosis, and 17% were given a bipolar disorder or schizophrenia diagnosis.

Fifty percent reported having ever read or downloaded their clinical notes through the My HealtheVet portal, and 17% recalled having spoken to their mental health clinicians about OpenNotes. The mean eHEALS score was 28.7 (s.d=6.7). In bivariate analyses, younger veterans (55.6 vs. 60.5; p=.001), veterans with PTSD diagnoses (63.9% vs. 52.7%; p=.047), and veterans who recalled having discussed OpenNotes with mental health clinicians (83.4% vs. 16.6%; p=.0001) were significantly more likely to report having read OpenNotes. In the final multivariate logistic regression model, PTSD diagnosis (OR=2.30 [1.31-4.07]), having spoken with a mental health clinician about the
availability of OpenNotes (OR=3.84 [1.69-8.72]), and patient confidence in understanding of the purpose and uses of OpenNotes were significantly associated with having read or downloaded OpenNotes. There was also a significant trend for the association of OpenNotes use with patient rating of the level of importance for health and recovery to have access to personal health information using OpenNotes (p<.001).

Discussion
To our knowledge this is the first study to examine predictors of use of OpenNotes in a population of patients receiving active mental health care who had access to them through an electronic health record portal. Not surprisingly, in our bivariate analyses, we found that younger age, education, and eHealth literacy were associated with OpenNotes use; these factors have been found to be associated with use of the internet and eHealth resources in prior studies of other patient populations (8,9). Recalling having spoken with their mental health clinician about OpenNotes was a strong predictor in the final model of having used OpenNotes.

We found that veterans with PTSD were more likely to be use OpenNotes that other patient groups; PTSD diagnosis was also a predictor of OpenNotes use in the final model. It is possible that anxiety or the challenges with trust (10) that have been associated with PTSD may drive desires to examine clinical notes. We also found in a prior study (4) that veterans with PTSD often have active claims for disability; such claims frequently depend on information in the record. Clearly, this finding needs to be explored in more depth. We also found that patients with schizophrenia or bipolar disorder were not less likely to use OpenNotes compared to other groups—findings regarding this population’s use of the internet have been mixed. Although one prior study showed that veterans receiving mental health care were not less likely to use MyHealthVet than veterans not receiving mental health care (8), other studies have found that veterans with serious mental illness report less internet and electronic health portal use relative to the general veteran population (8,11,12). Our study was limited by use of self-report of OpenNotes use.

Conclusion
Together, our findings demonstrate the value of mental health clinicians openly discussing OpenNotes with their patients if they wish to help them take advantage of the benefits OpenNotes may offer. We likely need to develop and disseminate tools to clinicians to help them engage in these proactive discussions, and further research should be conducted to better understand potential barriers to use of OpenNotes among patients with mental health conditions.

References
The Role of Chronic Inflammation as a Response Biomarker to Immune Checkpoint Inhibition in Cancer

Sherif M. El-Refai, PharmD, MBA¹, Val R. Adams, PharmD¹, Esther P. Black, PhD¹, Jeffery C. Talbert, PhD¹
¹University of Kentucky, Lexington, KY, USA

Abstract

Chronic inflammation provides cancer patients the ability to mobilize a swift immune response to tumor cells with checkpoint inhibition. Prior studies utilized animal models or human tumor samples, however, we have taken a novel approach to leverage national claims data to couple patient histories with response to therapy. We seek to identify a biomarker of response to immune checkpoint inhibitors that improves patient care using healthcare histories and molecular markers that correlate with therapeutic efficacy.

Introduction

Agents that unmask the anti-tumor immune response have provided some patients durable clinical benefit. Nivolumab, an immune checkpoint inhibitor (ICI), is effective in a small group (10-20%) of non-small cell lung cancer (NSCLC) patients when used as single-agent therapy. The approved biomarker of response is expression of the immune cell surface molecule, programmed death ligand 1 (PD-L1), on tumors measured by immunohistochemistry (IHC). The limitations of this assay include timing of tissue acquisition, tumor heterogeneity, and timing of therapy relative to the expression of PD-L1. Applying precision medicine to immuno-oncology requires the discovery of predictive biomarkers that can identify the patients most likely to benefit from this class of treatment.

Inflammatory diseases are, in part, caused by poor lifestyle choices and can contribute to cancer development. Paradoxically, the presence of inflammatory diseases may serve as biomarkers of durable response to immune checkpoint inhibitors. Our work provides experimental evidence that improves our understanding of the role of chronic inflammation in response to ICI. To position both wet lab and clinical studies that identify and validate components of the best biomarker of durable response, we initiated this project using claims data and informatics tools to ask whether ICI response and associated chronic inflammatory conditions are related. We identified a biomarker of response to PD-1 or PD-L1 inhibitors that improves patient care using a combination of healthcare outcomes and molecular markers that correlate with therapeutic efficacy.

Our preliminary data shows that chronic inflammation plays a role in ICI response. In an institutional IRB-approved study, 45 former and current NSCLC patients treated with a PD-1 inhibitor were assessed for a correlation between their history of chronic inflammation and ICI response. Patients with a history of at least one chronic inflammatory condition received more cycles of nivolumab therapy than those patients who did not have these conditions (Unpublished Data).

Study Design & Methodology

We utilized the Truven MarketScan claims database to identify cancer patients treated with ICIs, their responses, and their health histories to correlate ICI response with inflammatory disease pathologies. We characterized the response of cancer patients to ICI to determine whether durable response correlates with a history of chronic inflammation using health outcomes data. This study design takes advantage of our prior experience in the analysis of Truven MarketScan claims data 1, 2. We hold licenses for the use of the Truven MarketScan Claims database and this study represents ongoing collaboration with the director of the Institute of Pharmaceutical Outcomes and Policy (IPOP), Dr. Jeffery Talbert.

Adults aged 18 and older treated with an ICI between January 1, 2010 and December 31, 2015 were identified using J codes associated with the medications nivolumab, pembrolizumab & ipilimumab. The rate of all-cause death up to one year after diagnosis was compared by Cox proportional hazard regression. Patients were required to have at least a 24 month pre-index period. Chronic inflammation exposure groups were defined as patients with prior diagnosis of diabetes, hyperlipidemia, hypertension, obesity or chronic obstructive pulmonary disorder. Patient demographic characteristics included age, gender, geographic region and urban residence. Clinical characteristics measured during the 12-month pre-index period included previous chemotherapy use, previous radiation exposure, and the Charlson Comorbidity Index (CCI) 3. This includes 17 categories of comorbid conditions and is widely
used for risk adjustment with health outcomes data. Propensity score matching was conducted using the following variables: age, sex, region, CCI comorbidities, previous chemotherapy and previous radiation exposure. Cohorts were matched using a greedy, nearest neighbor algorithm with a caliper set at 0.2 times the standard deviation of the propensity scores in the sample, allowing for up to two matches. Propensity score matching mimics the randomization process of a clinical trial so that each matched pair has the same baseline probability (propensity) to respond \(^6\). Two-sample t-tests and chi-square tests are conducted to assess differences between cohorts before and after matching. Cox proportional hazard regression models are used to assess response and risk of death. Hazard ratios and their 95% confidence intervals are reported.

**Results**

Among 3,252 ICI-treated cancer patients identified in the Truven dataset, 2,339 had a history of chronic inflammation. Pre-existing chronic inflammation improved overall survival compared to no history of chronic inflammation (HR 1.23, 95% CI 1.01-1.50), specifically in patients with melanoma (HR 1.26, 95% CI 1.01-1.58). No differences were observed in patients diagnosed with lung cancer (HR 1.35, 95% CI 0.82 – 2.23). All results presented in Table 1. This study is subject to the limitations of all claims-based studies \(^7\), \(^8\). Notably, claims data lack detailed information on laboratory values or information on tumor staging. Additionally, studies with longer follow-up are needed to confirm whether chronic inflammation alters ICI response.

**Table 1.** One-year survival analysis. Hazard ratios (HR) and associated 95% confidence intervals (CI) for the effect of chronic inflammation on survival within the propensity-matched sample.

<table>
<thead>
<tr>
<th>N</th>
<th>Stratification</th>
<th>HR</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,218 vs. 730</td>
<td>All Selected Cancers</td>
<td>1.23</td>
<td>1.01 - 1.50</td>
</tr>
<tr>
<td></td>
<td>Lung Cancer</td>
<td>1.35</td>
<td>0.82 - 2.23</td>
</tr>
<tr>
<td></td>
<td>Melanoma</td>
<td>1.26</td>
<td>1.01 - 1.58</td>
</tr>
</tbody>
</table>

**Conclusions**

Our study suggests that a history of chronic inflammation improves response to immune checkpoint inhibitors. These findings concur with our prior 45-patient retrospective analysis. Importantly, this is the first study that has assessed the impact of chronic inflammation on response to ICI. Lastly, while propensity score matching is known to reduce selection bias in non-randomized studies, residual bias may remain present. The potential of precision immunology is vast and, once validated, can be used to inform and direct treatment decisions to improve outcomes.

**References**

Applying unsupervised learning to characterize rare observations in clinical data: the DQ\textsuperscript{e}-p tool

Hossein Estiri, PhD\textsuperscript{1,2,3}, Jeffrey G. Klann, PhD\textsuperscript{1,2,3}, Kavishwar Waghokit, MD, PhD\textsuperscript{1,2,3}, Shawn N. Murphy, MD, PhD\textsuperscript{1,2,3}

\textsuperscript{1}Harvard Medical School; \textsuperscript{2}Massachusetts General Hospital; \textsuperscript{3}Partners Healthcare, Boston, MA

Introduction
Data stored in Electronic Health Records (EHR) provide promising opportunities to advance health sciences research, and thereby, improve healthcare delivery and policy. However, data quality concerns have hampered secondary use of EHR data\textsuperscript{1,2}. Plausibility is an important dimension of data quality that represents whether EHR data values are “believable” or “truthful”\textsuperscript{3}. For example, a value of 1090 is almost unbelievable as the top number for systolic blood pressure in an EHR record. It is very likely that such a record presents an inherent data quality issue. Detecting plausibility issues in EHR data values currently requires extensive hard-coded rules – e.g., for each clinical observation, we need at least two rules, to identify minimum and maximum plausible values. DQ\textsuperscript{e}-p presents an alternative algorithmic solution to identifying implausible clinical data values by applying unsupervised learning algorithms. Here we describe the tool’s working assumptions, algorithms, and initial results, and discuss our ongoing efforts to improve its algorithms and architecture.

Methods
Development of DQ\textsuperscript{e}-p is based on two working assumptions. The first assumption is that unsystematic errors are small in size. That is, we do not expect to frequently observe, for example, a blood pressure record of 1090/80 in the EHR. Under this assumption and in the context of growing EHR data networks, we can use data from multiple EHRs to compare an individual data value to a large set of similar values and identify rare records. The second assumption is that clinical data are clusterable, meaning that they can be partitioned into groups of similar observations. DQ\textsuperscript{e}-p applies unsupervised clustering to identify rare observations in clinical observation data. To develop and test DQ\textsuperscript{e}-p, we use a sample of a million laboratory records from the Research Patient Data Registry (RPDR) from Partners HealthCare\textsuperscript{4} that are utilized by two large research networks, Scalable Collaborative Infrastructure for a Learning Health System (SCILHS) and Accrual to Clinical Trials (ACT). This dataset contains laboratory results from 86 Logical Observation Identifiers Names and Codes (LOINC).

Results
Before developing the tool, we tested the second working assumption/hypothesis (that clinical data are clusterable) by calculating Hopkins statistic for each group of laboratory values. Hopkins statistic is a reliable measure of spatial randomness\textsuperscript{5}. Our results showed that 73 percent of the 86 distinct laboratory tests are clusterable. We then applied a data transformation procedure that extended the distance between data points based on their z-score, multiplying each raw lab value to a product of its z-score (as a measure of deviation from the mean). Hopkins statistic on the modified values showed that 100 percent of laboratory tests were clusterable. We later used clustering results from the raw and modified values to create a flagging system. To cluster the data, we tested a number of unsupervised clustering algorithms including variants of the k-means and hierarchical clustering algorithms. Each method had its benefits and constraints. We found that the hybrid hierarchical-k-means (H-K-means) clustering\textsuperscript{6} provided the best performance. DQ\textsuperscript{e}-p categorizes data into groups and applies H-K-means algorithm to cluster each group. DQ\textsuperscript{e}-p uses a bootstrap procedure, \textit{kluster}, which we have developed and tested to identify the optimum number of clusters for unsupervised clustering. It then characterizes rare observations as clusters to which less than one percent of observations belongs.

After clustering each laboratory test result into the pre-identified number of clusters, DQ\textsuperscript{e}-p generates three flags that represent the potential importance of implausible records and produces an html report of the flagged records for further evaluation. The most important flag type is when DQ\textsuperscript{e}-p identifies a record as rare from both the raw and modified observation values. From the 1 million observation records from RPDR data, DQ\textsuperscript{e}-p flagged 48 records under type one. These records included, for example, three Urine Protein test values above 11,300, and four Creatine Kinase test values above 104,100. Type two flags are for observation values that are characterized as rare from the raw values, but not from the modified values. DQ\textsuperscript{e}-p flagged 238 observations under this flag type, including, for example, multiple Absolute Lymphocyte Count (ALC) values of above 2,585. Type three flags represent rare events...
(records from clusters with less than one percent data points) that are only identified from the modified observation values, which present the most liberal case between the three flag types. For example, in Figure 1, DQ²-p classification on modified observation values has characterized a single observation with a value of around 579 as a potential rare observation for Low Density Lipoprotein Cholesterol (CLDL) – the record belongs to cluster 3. The three clusters for CLDL were identified using Bayesian Information Criterion with kluster procedure. DQ²-p flagged a total of 2,160 observations as type three flag.

Figure 1. DQ²-p classification of Low Density Lipoprotein Cholesterol (CLDL).

Discussion
DQ²-p offers an algorithmic solution to approximate implausible observations in clinical repositories as potential data quality issues. This approach is an alternative automated solution to the common manual procedure that require at least two hard-coded rules (minimum and maximum acceptable values) for each observation. In our development database, DQ²-p produced a report with a total of 2,398 flagged records (out of 1 million observation records, or 0.24 percent) from 86 laboratory tests – which would have required at least 172 (86×2) hard-coded rules to identify. We are evaluating the specificity and sensitivity of DQ²-p results by comparing them with results from hard-coded rules. To improve specificity and sensitivity of the tool, we are experimenting further refinements that include implementing autoencoders from Deep Learning algorithms family to complement the H-K-means algorithm, as well as distinguishing observations by gender, age, and race/ethnicity, in addition to the current of categorization by LOINC codes. We are also exploring DQ²-p as a patient characterization cell for informatics for integrating biology and the bedside (i2b2)7 to perform automated test of implausible values and characterize patients into groups with similar health status.

Acknowledgements
This work has been partially supported by PCORI CDRN-1306-04608 (282364.5077585.0007) and NLM training grant T15LM007092.

References
Decision Support to Efficiently Identify Patients for Advanced Heart Failure Therapies.

R. Scott Evans\textsuperscript{1,2}, Abdallah G. Kfoury\textsuperscript{2}, Benjamin D. Horne\textsuperscript{2}, James F. Lloyd\textsuperscript{1}, Jose Benuzillo\textsuperscript{1,2}, Kismet Rasmusson\textsuperscript{2}, Colleen Roberts\textsuperscript{2}, Donald Lappe\textsuperscript{2}

\textsuperscript{1}Medical Informatics, Intermountain Healthcare, \textsuperscript{2}Cardiovascular Clinical Program Intermountain Healthcare. Salt Lake City, UT

Introduction: Heart failure (HF) is associated with significant morbidity and mortality [1-2]. Of nearly 6 million HF patients in the U.S., 5-10\% have advanced disease that can result in poor quality of life, frequent hospitalizations and rising costs that pose a significant burden for the different health care systems in the U.S. and worldwide [3-5]. Due to a limited number of donor hearts, cardiac transplantation remains an option for only a few advanced HF patients [6, 7]. Device and advanced medical therapy can significantly impact symptoms, improve function, and reduce mortality [8-13]. Previous studies report that patients who receive early advanced therapy have better long-term survival [14-16]. Thus, it is important for the patient’s primary care physician and/or cardiologist to refer patients at the appropriate time who meet established guidelines for these advanced therapies [18, 19]. However, many HF patients who could benefit from advanced therapies are not referred to advanced heart failure facilities as soon as possible [17]. We developed a clinical decision support (CDS) application based on 2013 ACCF/AHA criteria for advanced HF guidelines [12] that promotes timely and appropriate referral for advanced HF therapies. This presentation will discuss the development, use and initial impact this CDS application has had on the care for patients identified with advanced HF.

Methods: The CDS application went into production in January 2015 and identified HF patients associated with specialized heart facilities in the Intermountain Healthcare (IH) central region when they transitioned to advanced disease. The CDS application processes patient information found during specific time periods and automatically sends patient specific email alerts to the patients’ physicians. The message in the email is dynamic, lists the different patient risk factors and HTML links are imbedded allowing the provider to easily obtain further detailed information about referring patients and management for advanced HF. Intervention patients were those identified as eligible for advanced therapies by the CDS application during 2015 and were compared to control patients in the same region who would have been identified during 2013-2014 by the same criteria and method. Patients who had previously received device therapy before they met the advanced HF criteria used by the CDS application were not included in the final analysis since they were already receiving therapy. The same IH independently developed EHR was used to monitor all patients within the IH Central region during 2013 through 2015. A before-and-after and intention-to-treat study design was used to measure the impact of the new email alerts to identify patients meeting criteria for advanced HF. Advanced HF eligibility dates were defined as the date the first email alert was sent in the intervention population or would have been sent in the control population.

Results: In 2015, email recommendations for advanced HF therapy were sent to physicians for 516 different intervention patients. An average of 1.22 (range 1-5) email alerts were sent per patient and physicians requested to no longer receive the email alerts for only 25 (4.8\%) intervention patients because the patients had been placed on palliative care or the physician was no longer treating that patient. We did not receive any response from physicians indicating the alerts were not valid or a nuisance Intervention patients were referred to specialized heart facilities significantly more often within 30 days (57\% vs 34\%; p < 0.001), 60 days (69\% vs 44\%; p< 0.0001), 90 days (73\% vs 49\%; p < 0.0001) and 180 days (79\% vs 58\%; p < 0.0001) than control patients. Age and sex were not associated with HF facility visits, but patients of non-white race were less likely to go to a specialized heart facility. Significantly more intervention patients were found to be alive at 30 days (95\% vs 92\%; p = 0.036) at 60 days (95\% vs 90\%; p = 0.0013), 90 days (94\% vs 87\%; p = 0.0002) and 180 days (92\% vs 84\%; p = 0.0001) than control patients. In addition to visiting a specialized heart facility, age and sex were predictors of mortality, but race was not. Moreover, patients in both the intervention and control populations who visited specialized heart facilities had significantly longer survival at 30, 60, 90 and 180 days than patients who did not.

Discussion: This study found that patients with advanced heart failure who visited a specialized heart facility had significantly lower mortality rates within 30, 60, 90 and 180 days of identification of advanced HF based on our criteria and supports previous studies [14-16]. However, physician familiarity with continuous advances in medical
and device therapy along with constant heart failure patient monitoring can be challenging. The CDS application that used patient specific information was found to be able to monitor HF patients each day and identify those eligible for advanced therapy and help get them referred to a specialized heart facility sooner along with a significant reduction in mortality. Cardiovascular clinicians report the CDS application and email alerts helps physicians identify patients eligible for advanced therapies sooner. Development of the application required a multidisciplinary team including cardiovascular experts specializing in advanced HF and medical informatics. As usual, the value of the CDS application could only be developed and improved through prospective use and evaluation from the cardiovascular experts.

References

Using Visual Analytics and Patient-Generated Data to Support Clinical Decision-Making in the Context of Nutritional Therapy for Individuals with Diabetes
Daniel J. Feller¹, Marissa Burgermaster¹, Ph.D, Lena Mamykina¹, Ph.D
¹Department of Biomedical Informatics, Columbia University

Background. As many as 9% of adults living in the United States have diabetes, resulting in $176 billion in direct annual medical costs.¹ Self-management such as dietary self-monitoring is critical to the successful management of diabetes.² Increasingly, persons living with diabetes use mobile health technology (mHealth) to capture data related to their activities and health, for example their meals and blood glucose levels.³ The growing availability of these tools has led to an explosion of patient-generated data, which can provide critical insights into health behavior and patient well-being outside of clinical encounters. However, vast amounts of patient-generated data have raised concerns regarding information overload among clinicians and previous research suggests that identifying trends in self-monitoring data is challenging.⁴ We hypothesize that visual analytics - the science of analytical reasoning facilitated by advanced interactive visual interfaces - can enable the use of patient-generated data for clinical decision making.⁵ This study sought to (i) use participatory design to develop a novel visual analytics tool capable of revealing meal patterns associated with blood glucose changes and (ii) evaluate the usability and efficacy of the interactive tool with registered dietitians.

Methods. Participatory design was used in the development of Glucolyzer, an interactive tool featuring hierarchical clustering and a heatmap visualization to help registered dietitians (RDs) identify associative patterns between blood glucose levels and nutrition in meals. 10 RDs were recruited for a within-subjects experiment to compare the interactive tool to a static HTML representation (that is consistent with a meal log structure typical for meal diaries). Participants had 25 minutes to examine 1 month of diabetes self-monitoring data including photos of meals and blood glucose readings captured by an individual with type 2 diabetes. A ‘think-aloud’ protocol was employed and trial transcripts were transcribed and analyzed to reveal thematic content.

Results. 10 certified diabetes educators (CDEs) participated in the study. Participants generated 101 observations using the interactive tool and 70 using the static representation. Participant statements were more accurate using the interactive tool (69% vs 62%). Glucolyzer enabled participants to consider large groups of meals and identify more than 2x as many person-specific trends in glycemic impact. Usability was high although participants were varied in their enthusiasm for hierarchical clustering, which identified groups of meals with similar macronutrients and glycemic impact.

Discussion. Our findings suggest that computational learning and visual analytics may encourage the meaningful use of patient-generated data in diabetes by clinicians. We evaluated an interactive tool that empowers registered dietitians to analyze large volumes of data and identify reveal groups of meals with similar nutritional profiles. Glucolyzer provided insight into patterns of health behavior and glycemic control which lies in contrast to contemporary approaches to tailored nutritional therapy that rely upon the subjective assessment of psychological traits.⁶ Our results suggest that visual analytics may ameliorate concerns of information overload and thus reduce barriers to the inclusion of patient-generated data in clinical decision-making.

References.


---

**Figure 1.** Heatmap representation with meals sorted by glycemic impact (low to high)

**Figure 2.** Tooltip with embedded nutritional information and photo

**Figure 3.** Summary of nutritional and blood glucose information over 1 month of self-monitoring data.
New Relevance Search Algorithm for PubMed

Nicolas Fiorini, PhD¹, Zhiyong Lu, PhD¹
¹National Center for Biotechnology Information, Bethesda, MD, USA

Introduction

With more than 27 million articles in MEDLINE, retrieving and ranking the most relevant papers for a given query is increasingly challenging. Starting in the 2000’s, the machine learning community have focused on document ranking and created learning-to-rank (L2R) algorithms, demonstrating that robust and accurate relevance models can be built by utilizing various relevance signals and large training datasets. Recently, this technology has matured enough to scale up to real-world applications.

The method we present in this paper is twofold and is partly inspired from the works by Liu¹ and Dang et al.² in Web search: (a) a search engine retrieves papers that match a query and order them using a classical weighting model, commonly BM25³; (b) it reorders the most relevant documents retrieved using L2R – in our setup, L2R reorders the top 500 documents returned by Solr, an open-source indexing tool supported by a large community. L2R relies on a robust and accurate model heavily trained on thousands of queries. The main differences from previous approaches lie in that we need to design the system – further referred to as Solr-L2R – to fit the biomedical literature, and that we need to optimize it to handle PubMed load peaks.

Methods

In order for L2R to learn a model, it needs a gold standard to target. While there are many general Web search datasets available due to popular challenge tasks, biomedical datasets are rarer, smaller and some of them are outdated. Moreover, they do not always represent all behaviors we observe in PubMed, where various profiles have different information needs, query formulations, etc. We therefore created a training dataset from actual PubMed queries, using the anonymized queries stored in the logs, as well as any actions users subsequently took. We collected about one year and a half worth of logs, from which we created the training dataset. There are two main user actions that we consider to indicate that the document relevant. One is the abstract click, when a user clicks on a document in the list of results matching their query. The other is full text click, which occurs when a user requests the full text, after having clicked on an abstract.

We collect log data by following three steps. (1) We count all clicks for each article for each query; (2) we store the query and the corresponding articles that have at least one abstract click, as well as the 20 first results that have been returned by the system to the user; (3) we keep queries that appeared at least ten times – about 46,000 remain – and sum up the clicks for each article, for each query. This last step limits noise in the data by focusing on popular queries.

Documents have to be ordered by relevance in the gold standard, thus we propose a way to extract relevance from click data. Let us denote a document by \( d \) and a query by \( q \). \( a(d,q) \) is the number of abstract requests for \( d \) after \( q \). \( f(d,q) \) is the number of full text requests for \( d \) after \( q \). Say \( NFT \) represents the set of articles in the corpus for which the full text is not available. \( 1_{NFT}(d) \) is an indicator function such that:

\[
1_{NFT}(d) :=
\begin{cases} 
1 & \text{if } d \in NFT \\
0 & \text{if } d \notin NFT
\end{cases}
\]

The relevance of \( d \) with regard to \( q \) is calculated as follows:

\[
relevance(q,d) = \mu \cdot a(q,d) + (1-\mu) \cdot f(q,d) + \frac{a(q,d)}{\lambda} \cdot 1_{NFT}(d).
\]

Where \( \mu \in ]0,1] \) is the weight of abstract and full text clicks, and \( \lambda \in \mathbb{R}^+ \) is the boost provided to papers for which full text is not available. The reason justifying this boost is that two papers with the same number of abstract clicks and zero full text clicks should not be scored the same if full text is available for one and not for the other. In the gold standard, documents are ordered by descending relevance. \( \mu \) and \( \lambda \) were empirically tuned in offline experiments.

Our objective is to maximize the quality at the top of the list, i.e. the first page of results. To this end, we create a training set where, for each query, we assign scores between 12 and 3 with the articles Solr returns that appear in the top ten in the gold standard, then 2 with the following 10 documents, 1 for the remaining relevant documents and 0
for the remaining documents. We chose the NDCG\textsuperscript{4} (Normalized Discounted Cumulative Gain) metric in order to assess the ranking quality and train the ranker. It is defined as follows:

\[
\text{NDCG}@k = \frac{\text{DCG}@k}{\text{IDCG}@k} = \sum_{i=1}^{k} \frac{2^{\text{rel}_i} - 1}{\log_2(i+1)},
\]

where IDCG@k is the ideal DCG@k, that is, for the gold standard relevance order. rel\textsubscript{i} is the relevance of the document at rank i in the list to be evaluated. Finally, we designed a set of more than 150 features that capture the relatedness between the query and the document (e.g., the number of matches), document specifications (e.g., its publication type) and query specifications (e.g., the query length). The features capturing the relatedness between the query and a document contributed the most to the ranking quality. Query coverage (how well a query is represented in the text) and term proximity (how close the query terms are to each other in the document) are particularly important.

Results

LambdaMART\textsuperscript{5} was the best performing L2R algorithm in our tests. We first evaluated its ability to learn from the training set by scoring the two layers of the system. The first layer is Solr and is our baseline, the second layer is the method we compare, Solr-L2R. Solr gets a score of \textbf{NDCG@20 = 0.46}, while Solr-L2R gets a score of \textbf{NDCG@20 = 0.46}. This result confirms that LambdaMART can learn from the training set and the features we designed. We manually analyzed the output of Solr and Solr-L2R by submitting them to experts in various domains. It was found that Solr-L2R is able to return documents that are both relevant and recent.

We optimized the pipeline, as it needed to comply with PubMed’s load requirements. Especially, we optimized the complexity of feature calculation algorithms for the top 500 documents returned by Solr, as it was the most computationally expensive step. It is now able to process about a thousand queries per second in parallel at an average of 100ms per query on our computation cluster.

Finally, we evaluated the new system using an A/B testing approach with real users. We measured the performance of Solr-L2R and then-current PubMed by calculating the click through rates for each, that is, the proportion of queries where users click at least once on the first page of results. Solr-L2R showed an improvement in terms of click through rates of \textbf{10.8\%} over PubMed.

Discussion and conclusion

The results show not only that Solr-L2R provides documents that are much closer to ordering of the gold standard’s order (shown by NDCG increase), but also that the gold standard we created accurately represents how to answer users’ information needs (shown by click through rate increase). The new PubMed relevance search algorithm has been deployed in PubMed production system and is used when ‘Best Match’ sort order is selected by the users. Over time, users will generate more click data and Solr-L2R will be re-trained on more accurate targets, thus improving the search quality over time. We will be able to check this hypothesis by monitoring click through rates as we improve the model.

Acknowledgements

We thank Won Kim, Rafis Ismagilov, Grisha Starchenko, Evgeny Kireev, Maxim Osipov, Vadim Miller, Kathi Canese, John Wilbur and David Lipman for their help on the project and Robert Leaman for proofreading this abstract. This research was supported by the NIH Intramural Research Program, National Library of Medicine.

References

Forecasting the Maturation of EHR Functions among US Hospitals

Eric Ford¹, Hadi Kharrazi¹
¹ Johns Hopkins School of Public Health, Baltimore, MD, USA

Abstract
This study forecasts the maturation of EHR functionality and applications among non-federal US hospitals until 2035. HIMSS’ EMRAM data is used to train the Bass technological diffusion model. According to the forecast, the majority of hospitals will not reach Stage Seven of EMRAM until 2031, provided there are no major policy changes.

Introduction
The Meaningful Use (MU) program has propelled the adoption of EHRs among US non-federal hospitals. In 2008 only 9.4% of eligible hospitals used a comprehensive EHR, while this figure increased to 44.4% in 2012, and reached 83.8% of all non-federal hospitals in the US by the end of 2015 (1). Studies have shown that the adoption of EHRs (attesting to MU stage-2) has been slow in certain types of hospitals but the overall adoption has increased in all types of hospitals (urban and rural; teaching and non-teaching) (2). These studies, however, neither evaluated the adoption of advanced functionalities of EHRs (i.e., higher levels of HIMSS’ EHR maturation level (3)) nor forecasted EHR maturation over an extended period of time. The purpose of this study is to explore U.S. hospitals’ EHR technology development and adoption patterns (following a gradual upgrade of functionality and applications).

Methods
The HIMSS Analytics’ Electronic Medical Record Adoption Model (EMRAM) dataset was used to track historic uptakes of various EHR technologies and functionalities considered critical to improving healthcare quality and efficiency in hospitals (3). The Bass ‘BB-01 Generations’ model was used to model the technological diffusion rates for repeated EHR adoptions where upgrades undergo rapid technological improvements (4). The forecast used EMRAM data from 2006 to 2014 to estimate adoption levels to the year 2035. In addition, the model produced EMR sophistication forecasts describing how many hospitals will have achieved what level of capability in the near future. Microsoft Excel’s Visual Basic and SAS were used to train and apply the model to EMRAM data.

Findings
In 2014, there were over 5,400 hospitals that completed HIMSS’ annual survey of EMRAM. In 2006 the majority of the US hospitals were in stages 0, 1 and 2 of EMRAM while in 2014 most hospitals achieved stage 3, 4 and 5 of EMRAM (Figure 1).

Figure 1. Distribution of EMRAM scores in US Hospitals from 2006 to 2014
The overall Bass model reached an adjusted R-squared of .91. The final forecast depicted various trends for each of the EMRAM stages (Figure 2). The first year of observation (2006) already shows peaks of stage 0 and 1 as EHR adoption predates HIMSS’ EMRAM. By 2007, stage 2 reaches its peak while stage 3 reaches its height by 2011 and stage 4 by 2014. The first three stages create a graph that exhibits an ‘S-curve’ with inflection point being the peak diffusion rate. The forecast indicates that stage 5 will reach its peak by 2019 and stage 6 by 2026. Thought this forecast extends to the year 2035, no peak was readily observed for stage 7. Overall, although by 2020 most hospitals will achieve either stage 5, 6 or 7 of EMRAM, a considerable number of hospitals will not achieve stage 7 by 2035 (Figure 2).

![Figure 2. EHR adoption forecast among US hospitals using the EMRAM maturation stages](image)

**Conclusion**

By using the HIMSS EMRAM dataset and Bass diffusion models, we were able to forecast adoption of EHR capabilities from a paper-based environment (stage 0) to an environment where only electronic information is used to document and direct care delivery (stage 7). According to the forecast, the majority of hospitals will not reach stage 7 after the year 2031, provided there are no major policy changes.

**Learning Objectives:**

1. Explain the adoption of EHRs among US hospitals using HIMSS’ EMRAM maturation model
2. Understand the differences between meaningful use adoption of EHR levels and EMRAM’s forecasting of advanced EHR functionalities and applications

**References**

Creating a Computable Phenotype for Pregnancy for Clinical Research

Bret J. Gardner¹, Jay G. Pedersen, M.S.¹, James R. Campbell, M.D.¹, James C. McClay, M.D., M.S.¹

¹University of Nebraska Medical Center, Omaha, Nebraska

Introduction:

The simple question of whether a patient is pregnant is of critical importance in the clinic and for research. There are many tests and techniques to identify this status with confidence; however, these data are stored in various forms in diverse locations by unique electronic health record (EHR) systems limiting interoperability. With this diversity, determining an accurate answer to this query from EHR data has proven challenging. In an analysis of computerized physician order entry (CPOE) and decision support systems, Metzger et al noted 85% of adverse drug events related to pregnancy were not detected. With the increasing number of distributed research networks and federated queries for cohort identification, an accurate and interoperable approach to assessing pregnancy status is essential.

While many research protocols require knowledge of pregnancy status, the identification of pregnancy from EHR data is often made ad hoc and is not documented to be sensitive or specific. Our approach is to identify a validated pregnant population and age-matched control group, assess the frequency and predictive power of available facts from the EHR indicative of pregnancy, and create a valid logistic discriminant model for pregnancy status.

Method:

We proceeded to identify the cohort of all pregnant patients seen for any care at the University of Nebraska Medical Center (UNMC) during 2015 using a two-step process. We first employed our de-identified i2b2 data warehouse (IRB #132-14-EP) to retrieve obstetric episodes and employed the associated patient set to explore the discriminant value of data elements listed in table 2. Contingency tables were created to compare the pregnant cohort to a control population using Pearson Chi-Square tests for significance for each variable. For a control group we restricted our population to females between ages 15 and 50 years without a pregnancy episode noted. When IRB approval for exemption status for chart review is obtained, we will supplement case identification with record review of patients noted to have suggestive findings. NLP methods will be employed to search for episodes with scant EHR findings. The recorded date of delivery or termination of pregnancy coupled with estimated dates of conception based on gestational age will bound each pregnancy episode in the study cohort. We will confirm a study population of all patients with a pregnancy episode anytime in 2015. An age-matched control group will be randomly selected from women of the appropriate age who were seen for primary care services at UNMC during the same year.

Refining the Predictive Model:

Once the study population is confirmed, we will refine the criteria for a positive finding by review of temporal frequency of results in the study group. For example, in the case of a pregnancy test, we will choose random sampling times during the study year and explore the temporal relationship of positive findings to sample time. Since pregnancy tests are typically done as early as two weeks into a pregnancy and reliably become positive at 8 weeks, we would expect that random sampling of records would yield positive findings for pregnant patients with events in the previous 8 months. We might then conclude for optimal sensitivity and specificity that a positive pregnancy test within the last 7 months would be our best definition.

Once the variable set has been defined for categorical results, observation data sets will be retrieved for the study and control populations at several time intervals during the study year, probably quarterly. These data sets will be analyzed using multiple logistic regression techniques to determine the most parsimonious and valid model for a finding of pregnancy status. Varying the temporal definition of the independent variables, resampling at multiple times of year and varying the logistic threshold for assignment, we will perform sensitivity analyses and construct a receiver operating characteristic (ROC) curve for validation of the model. From the ROC we will determine the optimum discriminant strategy.

Preliminary Results:

The demographic characterization of the pregnant and control populations from the de-identified record review of step one are described in Table 1. The majority of both the pregnant cohort and control group were white (64% and 74%, respectively). The median age for this female population was 29 for the pregnant cohort and 33 for the control group. The frequencies and p-values for covariates are shown in Table 2. All variables demonstrated a statistically significant difference with the control population. The serum pregnancy test was more prevalent in the control group.
Table 1. Description of Population Demographics (prevalence of pregnancy episodes in one calendar year).

<table>
<thead>
<tr>
<th>Race (count (%))</th>
<th>Pregnant (n = 3,422)</th>
<th>Control (n = 55,361)</th>
<th>Age (years)</th>
<th>Pregnant (n = 3,422)</th>
<th>Control (n = 55,361)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Black or African American</td>
<td>592 (17.3)</td>
<td>6,339 (11.45)</td>
<td>Minimum age</td>
<td>14</td>
<td>15</td>
</tr>
<tr>
<td>Asian</td>
<td>104 (3.04)</td>
<td>1,272 (2.30)</td>
<td>Maximum age</td>
<td>47</td>
<td>50</td>
</tr>
<tr>
<td>White or Caucasian</td>
<td>2,220 (64.87)</td>
<td>41,269 (74.55)</td>
<td>Median ± SD</td>
<td>29 ± 5.70</td>
<td>33 ± 9.64</td>
</tr>
<tr>
<td>Other or Unknown</td>
<td>561 (16.39)</td>
<td>5,503 (9.94)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2. Variables and p-values for potential covariates (total number with population frequency (%)).

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
<th>Pregnant (n=3,384)</th>
<th>Control (n=52,315)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Encounter diagnoses</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ICD10CM:Z33-34</td>
<td>Supervision of pregnancy</td>
<td>2,599 (76.8)</td>
<td>1,259 (2.4)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>ICD10CM:O60-O94</td>
<td>Complications of delivery</td>
<td>768 (22.7)</td>
<td>508 (1.0)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>ICD10CM:O009-O29, O98, O99</td>
<td>Disorders during pregnancy</td>
<td>2,016 (59.6)</td>
<td>1,417 (2.7)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>ICD10CM:O31-O41</td>
<td>Pregnancy complications</td>
<td>762 (22.5)</td>
<td>428 (0.8)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Procedure performed</td>
<td>Obstetric ultrasound</td>
<td>2,158 (63.8)</td>
<td>1,389 (2.7)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>LOINC:822-1</td>
<td>ABO and Rh group [Type]</td>
<td>2,009 (59.4)</td>
<td>1,710 (3.3)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>LOINC:20415-6</td>
<td>B-hCG in serum by immunoassay</td>
<td>463 (13.7)</td>
<td>577 (1.1)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>LOINC:2106-3</td>
<td>hCG in urine</td>
<td>175 (5.2)</td>
<td>4,482 (8.6)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>LOINC:2118-8</td>
<td>hCG in serum/plasma</td>
<td>48 (1.4)</td>
<td>1,416 (2.7)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Problem list entry</td>
<td>Disorder of pregnancy</td>
<td>860 (25.4)</td>
<td>557 (1.1)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>SNOMEDCT:173300003</td>
<td>Patient currently pregnant</td>
<td>2,071 (61.2)</td>
<td>1,088 (2.1)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>SNOMEDCT:16356006</td>
<td>Multiple pregnancy</td>
<td>55 (1.6)</td>
<td>19 (0.0)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>OB clinical encounter and measures</td>
<td>Visit to obstetric clinic</td>
<td>47 (1.4)</td>
<td>270 (0.5)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>LOINC:11881-0</td>
<td>Uterus Fundal height Tape measure</td>
<td>1,720 (50.8)</td>
<td>750 (1.4)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>LOINC: 55283-6</td>
<td>Fetal heart Rate</td>
<td>1,848 (54.6)</td>
<td>760 (1.5)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>LOINC: 57088-7</td>
<td>Fetal movement - Reported</td>
<td>1,610 (47.6)</td>
<td>755 (1.4)</td>
<td>&lt;0.0001</td>
</tr>
</tbody>
</table>

Discussion:
An accurate predictive model for identifying the pregnancy status of a patient is paramount for clinical research. We have described a set of discriminant variables, readily available within an EHR which may be employed to compute probability of a concurrent pregnancy. Based on these covariates, using a pregnant cohort and a matched control population, a predictive model will be developed, tested and reported at AMIA. Plans for validation and characterization of the model for use in research study design and execution are discussed.

References
INTRODUCTION

Infectious diseases and antimicrobial resistance constitute one of the most pressing challenges for healthcare systems. Current responses to such challenges acknowledge the vital role played by citizens and aim to benefit from increased population self-efficacy. Clinicians, patients and potential patients can benefit from timely, easily accessible, and region-specific information about communicable infectious disease activity in their communities. Access to such information has the potential to improve decision making in the context of extremely common conditions that account for a majority of acute, ambulatory healthcare utilization and are associated with threats to patient safety (e.g., unnecessary antibiotics, adverse drug events) and patient satisfaction (e.g., complaints relating not getting an antibiotic).

Advances in infectious disease molecular diagnostics, data warehousing and analytics, and data visualization and dissemination platforms have made it possible to provide front line clinicians and health care consumers with access to timely, population-level and regionally relevant information about infectious disease activity in the populations being served by healthcare delivery organizations. These technologies combine to provide unprecedented access to information that can aid and inform clinical decisions for both caregivers (effective diagnostic test utilization, improved antibiotic stewardship, improved patient education) and care seekers (aid with self diagnosis and management, knowing when to seek care, overall more informed consumers).

Utilizing clinical data routinely generated during healthcare delivered in an integrated delivery system including 23 hospitals and >150 clinics, GermWatch provides Utah’s citizens with timely information about regional infectious disease activity including common respiratory tract infections (e.g., Influenza, RSV, Pertussis).

METHODS

We used focus groups with clinicians and parents to elucidate design objectives and iterative design principles to guide development of multiple approaches for disseminating information about regional infectious disease activity in Utah. The focus groups with parents addressed five topic areas: response to symptoms, frustrations with healthcare encounters, information sources and needs, parent overall reaction to the GermWatch concept and suggestions for how to improve the system.

RESULTS

We identified several salient themes regarding acute respiratory tract infections that helped to inform the design of our consumer facing information resource including: 1) a need for information to help understand when to seek care (red flags, warning signs) and what care to expect (symptom management, role of diagnostic testing and antimicrobial therapy), 2) a need for information that helps empower them as a consumer and facilitate efficient diagnosis (what symptoms and concerns to discuss, what questions to ask), 3) information needs about diagnosis and expectant management (more than “it’s just a virus”), 4) consensus that just searching the internet is inadequate, often overwhelming and misleading and surveillance information is not relevant to my community, 5) the desire to consult a trusted local source (friend or family member in healthcare, their clinician or other medical experts) and have information presented in an easy to interpret form.

The high-level system architecture is shown in Figure 1 and the web site is available at www.GermWatch.org.
tables in the enterprise data warehouse, transforming and loading (ETL) it into a dedicated multidimensional schema. From there, we run additional ETLs to further process, analyze and aggregate the data for use and display in the various components of the system. We use COTS business intelligence tools (IBM Cognos and Tableau) to build, display and distribute data visualizations including dashboards, epidemic curves and HIPPA-compliant (via expert determination methodology) maps that display surveillance information aggregated at a sufficiently small geographic area so as to both protect privacy and provide information that is meaningfully granular in a spatial sense (Figure 2).

Surveillance information – “What’s going around” – is presented in an easy to interpret, activity-sorted list on the landing page. Clicking on an individual pathogen takes you to a dedicated subpage in our GermSchool, where the user can see a current epidemic curve and map for that pathogen (Figure 2) along with static content addressing consumer information needs including: typical seasonality, common signs and symptoms (also presented in a novel information visualization we call the ‘Infection Timeline’, as shown in Figure 3), method of transmission, diagnosis and treatment and when to seek care.

Web site and app traffic: between 9/1/2015 and 3/1/2017, the site has seen over 450,000 page views, 90,000 visits and 60,000 unique visitors. The site averages over 2,500 page views per week during summer and over 10,000 page views per week during the peak of cold and flu season.

DISCUSSION

This free service helps Utah’s healthcare providers and potential healthcare consumers stay informed about what germs are circulating our communities. Providers having access to information about what germs are going around helps them stay alert for the illnesses caused by these germs. This alertness can improve provider diagnostic accuracy (getting the right diagnosis), which, in turn, makes it easier for them to provide the right treatment and the right information to the right patient at the right time. Potential healthcare consumers having access to a responsively designed web site that connects them to this information resource along with tools that help them identify and access related healthcare resources has opened up exciting new opportunities for radically transforming extant acute healthcare delivery models to support a consumer-centric approach focused on delivering the right care to the right person in the right place. In these new models of care delivery, a potential healthcare consumer with an acute illness is able to see what illnesses are active in their region and if other people in their region are having similar symptoms. To address unmet information needs or concerns with a clinician, they can view nearby urgent care facilities and schedule an appointment at the clinic showing the shortest wait time with a provider knowledgeable about “what is going around”. Better yet, they might avoid the clinic visit altogether by learning enough to effectively self-manage their illness at home or by participating in a telemedicine consultation right from their device with a provider that is similarly knowledgeable about what illnesses are active in their community.

CONCLUSION

Sustained use of the consumer facing GermWatch system suggests significant and ongoing interest in this kind of information resource. Additional study is required to understand how this information is being used, how it is affecting health literacy in this domain how it is impacting (if at all) self-management and care seeking decision making, and care delivery quality and experience.

REFERENCES

Measuring the Value of EHR’s Free-text in Identifying Geriatric Risk Factors

Fardad Gharghabi¹, Laura Anzaldi¹, Tom Richards¹, Jonathan Weiner¹, Hadi Kharrazi¹
¹Johns Hopkins University, Baltimore, MD, USA

Abstract

This study evaluates the added-value of EHR’s unstructured text in identifying geriatric risk factors. The study applied pragmatic pattern matching methods with sensitivity and specificity rates averaging around 84%. The denominator consisted of 18,341 elderly patients. When compared to patients identified using structured fields, free-text added a significant number of cases to each geriatric risk construct ranging from social support (99.4%) and malnutrition (87.5%), to pressure ulcer (23.5%) and dementia (17.3%).

Introduction

The elderly population is rapidly growing in the US (1), and aging adults are particularly vulnerable to health conditions that carry significant risk of morbidity and other adverse outcomes (2). Geriatric healthcare policy primarily draws from sources such as insurance claims data or structured fields in electronic health records (EHR), using diagnostic and pharmacy data to characterize patients (3). Generally, the richness of EHR free-text is overlooked due to the difficulty of extracting data (4). We examined the added-value of free-text compared to claims and/or structured EHR data in identifying 10 clinical constructs associated with increased geriatric risk: social support, malnutrition, vision impairment, falls, walking difficulty, urinary control, fecal control, weight loss, pressure ulcer, and dementia.

Methods and Study Design

The denominator consisted of 18,341 elderly patients with an average age of 75.9 years, 58.9% female, who had at least 12 months of continuous insurance enrollment and for whom EHR free-text notes existed (Table 1).

<table>
<thead>
<tr>
<th>Demographics/Utilization</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population (N)</td>
<td>18,341</td>
</tr>
<tr>
<td>Age in years (SD)</td>
<td>75.9 (7.5)</td>
</tr>
<tr>
<td>Sex (%)</td>
<td>F 10,806 (58.9) &amp; M 7,535 (41.1)</td>
</tr>
<tr>
<td>Average comorbidity count (SD)</td>
<td>11.6 (5.35)</td>
</tr>
<tr>
<td>Average months enrollment (SD)</td>
<td>33.3 (5.84)</td>
</tr>
<tr>
<td>Average number of notes (SD)</td>
<td>132 (107.9)</td>
</tr>
<tr>
<td>Average number of characters (SD)</td>
<td>165,766 (141,687)</td>
</tr>
</tbody>
</table>

We retrospectively analyzed one year of free-text notes from the population using a regular-expression driven pattern-matching algorithm. The patterns were created from a corpus of phrases generated by geriatric experts’ manual tagging of one year of free-text notes for 185 selected patients. To maximize specificity, the algorithm identified a patient as having a construct if there were at least two positive mentions of that construct without any negative mentions of that construct. These constructs were compared to geriatric risk constructs extracted from claims and structured EHR data of the same patient population.

Findings

When compared to the manual tagging of the 185 patients, sensitivity of the proposed text-extraction method ranged from 26.8% to 89.5% and specificity ranged from 97.4% to 100% across all 10 constructs (Table 2). We further reviewed a random sample of 100+ patients from the larger population for each construct to evaluate the true positive rate for the whole population. The true positive rate ranged from 66% for weight loss to 100% for fecal incontinence, walking difficulty, malnutrition, and social support (not shown in tables).

We then compared the patients identified by our algorithm to those identified using claims and/or structured EHR data for each of the 10 constructs (Table 3). We found that free-text added a significant number of cases to each construct (% added value computed as the number of cases found only by the algorithm divided by the number of
cases found by any source): social support (99.4%), malnutrition (87.5%), vision impairment (65.9%), falls (50.2%) and walking difficulty (48.8%), urinary control (35.3%), fecal control (32.6%), weight loss (29.5%), pressure ulcer (23.5%), dementia (17.3%).

Table 2. Performance of the pattern matching method in identifying manually tagged geriatric risk factors

<table>
<thead>
<tr>
<th>Construct</th>
<th>Sensitivity %</th>
<th>Specificity %</th>
<th>PPV %</th>
<th>NPV %</th>
</tr>
</thead>
<tbody>
<tr>
<td>AFC</td>
<td>89.47</td>
<td>100.00</td>
<td>100.00</td>
<td>98.81</td>
</tr>
<tr>
<td>DEC</td>
<td>84.21</td>
<td>100.00</td>
<td>100.00</td>
<td>98.22</td>
</tr>
<tr>
<td>DEM</td>
<td>84.85</td>
<td>97.37</td>
<td>87.50</td>
<td>96.73</td>
</tr>
<tr>
<td>FAL</td>
<td>55.13</td>
<td>100.00</td>
<td>100.00</td>
<td>75.35</td>
</tr>
<tr>
<td>MAL</td>
<td>75.00</td>
<td>100.00</td>
<td>100.00</td>
<td>96.41</td>
</tr>
<tr>
<td>SSN</td>
<td>64.42</td>
<td>98.77</td>
<td>98.53</td>
<td>68.38</td>
</tr>
<tr>
<td>URC</td>
<td>62.96</td>
<td>100.00</td>
<td>100.00</td>
<td>94.05</td>
</tr>
<tr>
<td>VIS</td>
<td>82.35</td>
<td>100.00</td>
<td>100.00</td>
<td>96.18</td>
</tr>
<tr>
<td>WEI</td>
<td>26.79</td>
<td>99.22</td>
<td>93.75</td>
<td>75.74</td>
</tr>
<tr>
<td>WLK</td>
<td>66.09</td>
<td>100.00</td>
<td>100.00</td>
<td>64.22</td>
</tr>
</tbody>
</table>

Table 3. Comparing rates of geriatric risk factors using claims and EHR data (structured and free text) within the larger population reference

<table>
<thead>
<tr>
<th>Construct</th>
<th>C%</th>
<th>E%</th>
<th>NLP%</th>
<th>C+E%</th>
<th>C+E+NLP%</th>
<th>E+NLP%</th>
</tr>
</thead>
<tbody>
<tr>
<td>AFC</td>
<td>0.52</td>
<td>0.87</td>
<td>2.04</td>
<td>0.93</td>
<td>2.22</td>
<td>2.18</td>
</tr>
<tr>
<td>DEC</td>
<td>0.65</td>
<td>0.45</td>
<td>1.11</td>
<td>0.84</td>
<td>1.47</td>
<td>1.19</td>
</tr>
<tr>
<td>DEM</td>
<td>3.10</td>
<td>3.94</td>
<td>6.36</td>
<td>4.67</td>
<td>7.21</td>
<td>6.82</td>
</tr>
<tr>
<td>FAL</td>
<td>5.48</td>
<td>5.73</td>
<td>22.10</td>
<td>7.32</td>
<td>23.16</td>
<td>22.77</td>
</tr>
<tr>
<td>MAL</td>
<td>0.12</td>
<td>0.03</td>
<td>2.37</td>
<td>0.14</td>
<td>2.44</td>
<td>2.37</td>
</tr>
<tr>
<td>SSN</td>
<td>0.03</td>
<td>0.06</td>
<td>28.02</td>
<td>0.07</td>
<td>28.03</td>
<td>28.03</td>
</tr>
<tr>
<td>URC</td>
<td>0.55</td>
<td>0.97</td>
<td>3.10</td>
<td>1.09</td>
<td>3.38</td>
<td>3.30</td>
</tr>
<tr>
<td>VIS</td>
<td>0.53</td>
<td>0.76</td>
<td>4.81</td>
<td>0.85</td>
<td>4.94</td>
<td>4.92</td>
</tr>
<tr>
<td>WEI</td>
<td>4.68</td>
<td>5.60</td>
<td>9.11</td>
<td>6.65</td>
<td>11.96</td>
<td>11.18</td>
</tr>
<tr>
<td>WLK</td>
<td>8.39</td>
<td>7.88</td>
<td>32.45</td>
<td>10.80</td>
<td>34.93</td>
<td>34.02</td>
</tr>
</tbody>
</table>

Acronyms: AFC: Absence of Fecal Control; C: Claims [structured data]; DEC: Decubitus Ulcer; DEM: Dementia;  
E: EHR [structured data]; FAL: Falls; NLP: Natural Language Processing [free text data]; MAL: Malnutrition;  
NPV: Negative Predictive Value; PPV: Positive Predictive Value; SSN: Social Support Network; URC: Urinary Incontinence;  

Conclusion

It is feasible to achieve satisfactory sensitivity and specificity using a pragmatic pattern-matching algorithm to parse the free-text of EHRs for geriatric risk markers. Claims and structured EHR data give an incomplete picture about several constructs that influence geriatric risk. There is a high possibility to miss patients with geriatric risk markers when excluding free-text from an analysis

Learning Objectives:

1. Compare the value of structured data in administrative claims and EHRs to information extracted from EHR’s free-text data in identifying geriatric risk factors and its effect on geriatric health policy
2. Explain simple methods to make EHR’s free-text useful in identifying geriatric risk factors

References

A Conceptual Representation of Exposome in Translational Research

Ramkiran Gouripeddi, MBBS, MS1,2, Nicole Burnett1,2, Mollie Cummins, PhD, RN, FAAN, FACMI1,2,4, Julio Facelli, PhD, FACMI1,2, Katherine Sward, PhD, RN1,2,4

1Department of Biomedical Informatics, 2Center for Clinical and Translational Science, 3Department of Chemical Engineering, 4College of Nursing; University of Utah, Salt Lake City, Utah, USA

Introduction

Recent definitions of the concept of the exposome include endogenous processes within the body, biological responses of adaptation to environment, and socio-behavioral factors beyond assessment of exposures1,2. Research in generating and utilizing exposomes is therefore translational in nature as it includes direct biological pathway alterations as well as mutagenic and epigenetic mechanisms of environmental influences on the phenome. In this presentation, we discuss an initial conceptual representation of these multi-model and multi-scale exposomic data for its use in translational research and its informatics implications.

Methods

In order to develop a conceptual representation, we reviewed 96 environmental health articles describing air quality studies. We manually extracted the type of study performed, and the description of the data used for analysis. In addition, we elicited use-cases from translational researchers for studies that could foreseeably use the Pediatric Research using Integrated Sensor Monitoring Systems (PRISMS) infrastructure we are developing3, and extracted the same information pertaining to the type of study and data description. We then classified the type of study4, categorized and matched the data based on their description to the overlapping domains of exposome1, and linked the data categories to study types based on the nature of use of the data for study analysis.

Results and Discussion

Based on the methods used for data collection (DC) (measurements or observations), we grouped exposomic data into six broad categories [Figure 1]: (1) Sensor, (2) Clinical, (3) Biospecimen-derived, (4) Participant reported, (5) Aggregates, and (6) Computational Models. We classified types of studies into (1) Basic Science and Engineering, (2) Person-centric clinical, and (3) Population studies4. When matching these data categories with existing exposomal domains1, we considered the possibility of these DCs being used for any of the above study types. Also use of these DCs need to consider differences and uncertainties in environmental measurements and true exposures. In addition, we modified the exposomic domains to reflect the proximity of DC to the subject under consideration: (1) Personal Environment: DCs from within, on or of the subject’s bodily environment, (2) Immediate Environment: DCs from close surroundings such as home, school or workplace, and (3) General Environment: DCs from larger geographic areas such as weather, environmental monitors or satellites.

Research using exposomic data is diverse and can fall anywhere in the translational spectrum5. These include those related to the research and development of sensor devices, chemistry of environmental species, atmospheric science, exposure pathways, mechanistic understanding of environmental species on the genomic for precision medicine6, pharmacodynamics studies, clinical trials, observational, comparative effectiveness, epidemiological and public health studies. When using exposomic data, limitations of the data, and in particular, uncertainties associated with using the data as exact quantifications of exposure should be important considerations (depicted in Figure 1). To our knowledge, there are no previous systematic efforts to characterize uncertainties in exposomic data for use in translational research. Discovering metadata associated with DC is a first step in characterizing these uncertainties. Integration of exposomic data requires a metadata-centric approach that stores data in its primitive form maintaining spatio-temporal integrity and minimizing transformational information loss3. Such a platform can support an ecosystem of analytic platforms and data models, environmental wide association studies2 and the eventual development of personal exposome records.

References

Figure 1. A conceptual representation of exposome for use in translational research. At the center of the figure are the three modified domains of exposome. The middle ring consists of six broad categories of exposomic data with examples. Arrows (purple or orange) between the data categories and exposome domains indicate the presence (or absence) of approximations when used to represent the latter. The outer ring represents types of translational research. Dashed arrows with +/- signs indicate uncertainties associated with using data from these domains in different types of studies. Integration of these data (e.g. with OpenFurther) requires capturing associated metadata and uncertainties.
Efficient Remediation of Terms Inactivated by Dictionary Updates

Benjamin J. Gross, MD, MMS,1 David H. Dubois, MBA, PMP,1 Mark H. Twelves MSW, LICSW,1 Roberto A. Rocha, MD, PhD1,2,3
1Clinical Informatics, Partners eCare, Partners Healthcare System, 2Brigham and Women’s Hospital, 3Harvard Medical School, Boston, MA

Introduction

Partners Healthcare System (PHS) uses a commercial EHR and a third party diagnosis dictionary, which serves as a clinical interface terminology. Providers document patients’ conditions by selecting from over 800,000 clinical terms, which are mapped to ICD-10-CM codes and SNOMED CT concepts. The diagnosis dictionary is updated periodically by the vendor, including the addition of new terms and inactivation of existing terms. PHS further customizes the dictionary, and with each update a varying number of additional terms are inactivated.

Most dictionary updates do not impede providers’ ability to enter new visit diagnoses or problems because the dictionary contains significant redundancy. Only a minority of terms has ever been selected since the EHR was first implemented over two years ago (21% of terms) and even fewer are selected frequently (99% of usage is covered by only 7% of terms). However, certain inactivated terms require replacement with semantically equivalent active terms because they have dependencies. When terms appear in patients’ records and can be used downstream we say they have “process dependencies” and when terms are linked to other system elements, we say they have “configuration dependencies”.

Diagnosis terms with process dependencies must be active to allow subsequent downstream workflows such as billing. For example, terms on a patient’s problem list may be copied as new visit diagnoses, and terms may be associated with orders that are yet to be performed. To ensure proper billing and ordering processes, these associated terms must be active. When providers attempt to access or reuse inactive terms, the EHR system offers alternatives when replacement terms are available. Providers can either accept the suggested diagnosis term, or search for alternative terms. Conversely, diagnosis terms with configuration dependencies must be active in order to prevent runtime errors or functionality gaps. Examples include terms linked to questions on a history questionnaire, and terms that trigger a decision support rule.

Inactivated diagnosis terms can be replaced in advance. For terms with process dependencies, replacements help reduce provider workflow interruptions and save them time. Automated replacements can be made using available system tools. For terms with configuration dependencies, system tools facilitate replacements in some areas, but in other areas system analysts need to make replacements manually. In rare cases, instead of replacing inactive terms, system analysts can remove or change the affected configuration items. We will describe how we identify the most appropriate replacement terms, and how replacements are made in configuration and data records.

Methods

Mapping inactive terms to semantically equivalent terms requires clinical expertise and each replacement must be approved individually. PHS has developed reports to suggest multiple options for each inactivated term. The suggestions are ranked by the assumed likelihood of equivalence, so likely matches are ranked at the top. The criteria for suggesting alternative terms and their ranking order are listed in Table 1. Suggestions with a given rank are only those that do not meet the criteria of a higher rank. For instance, a suggested term that has “similar text + matching SNOMED CT concepts + matching ICD-10 codes” to the source term (rank 3) will rank higher than a term that has “similar text + matching SNOMED CT concepts” but does not have “matching ICD-10 codes” (rank 4). For each inactive term, terminology engineers select a single optimal replacement term from the list of suggestions. Proposed mappings are then reviewed by clinical subject matter experts (SMEs) who are practicing physicians in various specialties.

Results

During the most recent dictionary update, the sum of unique inactive terms queued for replacement in each area was 880. The sum of unique inactive terms replaced in each area was 778. Replacements were made for inactive diagnosis terms associated with the elements listed in Table 2.

In order to limit the work effort, terms associated with fewer than 50 orders and fewer than 50 problems are not replaced. Providers must replace these inactive terms themselves. Currently, term replacements also do not occur in preference lists defined by individual users, or value sets containing manually selected diagnosis terms.
### Table 1 – Criteria for suggesting alternative terms and their ranking order

<table>
<thead>
<tr>
<th>Rank</th>
<th>Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Previously approved mapping</td>
</tr>
<tr>
<td>2</td>
<td>Suggested by the dictionary vendor</td>
</tr>
<tr>
<td>3</td>
<td>Textual match (normalized terms)* + SNOMED CT concepts + ICD-10-CM codes</td>
</tr>
<tr>
<td>4</td>
<td>Textual match (normalized terms)* + SNOMED CT concepts only</td>
</tr>
<tr>
<td>5</td>
<td>Textual match (normalized terms)* + ICD-10-CM codes only</td>
</tr>
<tr>
<td>6</td>
<td>Textual match (similarity function)* only</td>
</tr>
<tr>
<td>7</td>
<td>Textual match (normalized terms)* only</td>
</tr>
<tr>
<td>8</td>
<td>Matching SNOMED CT concepts + ICD-10-CM codes only</td>
</tr>
<tr>
<td>9</td>
<td>Matching SNOMED CT concepts only</td>
</tr>
<tr>
<td>10</td>
<td>Matching ICD-10-CM codes only</td>
</tr>
</tbody>
</table>

* Textual similarity is assessed using two separate methods. “Textual match (normalized terms)” are terms that share identical normalized terms, after removal of punctuation and conjunctions, standardizing the word order and substituting words using preferred synonyms, acronyms, and morphological derivations. For example, the terms “Open Fracture of Right Tibia and Fibula, Sequela” and “Tibia and Fibula Open Fracture, Right, Sequela” both normalize to “fibula fracture open right sequela tibia”, so they are considered textual matches. The second method of textual matching determines similarity using a similarity function [1].

### Table 2 – Replacements for inactive diagnosis terms made during the most recent dictionary update

<table>
<thead>
<tr>
<th>Element</th>
<th>Count</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Configuration dependencies</td>
<td>402</td>
<td>Every term must be replaced, or configuration dependency must otherwise be changed or removed</td>
</tr>
<tr>
<td>Future orders</td>
<td>165,159</td>
<td>Terms associated with &gt; 50 orders are replaced</td>
</tr>
<tr>
<td>Active treatment protocols</td>
<td>682</td>
<td>All inactive terms are replaced</td>
</tr>
<tr>
<td>Problem lists</td>
<td>70,620</td>
<td>Terms associated with &gt; 50 problems are replaced</td>
</tr>
</tbody>
</table>

### Discussion

The ongoing maintenance of a diagnosis dictionary is not limited to ensuring the appropriateness of terms available for selection. Reference terminologies and code systems associated with diagnosis terms evolve over time, requiring periodic content updates that create important challenges to those responsible for system configuration and data integrity. Diagnosis terms are used in multiple ways within an EHR system, requiring an efficient process to remediate terms and codes that are inactivated and/or superseded by new ones. Inactivated terms with configuration dependencies and inactivated terms that have process dependencies at the time of the update require replacements with semantically equivalent active terms. PHS has developed a process that relies on detailed reports to suggest active replacement terms. Clinical SMEs then approve the semantic equivalence of each suggested term before replacements are made. This process must be repeated with every dictionary update. In the most recent update, for example, the effort required to find replacement terms, obtain SME approval, and implement the replacements was approximately 75 hours, involving 13 people with complementary roles and expertise. Considering 880 inactivated terms were evaluated and 236,863 replacements were made, the relatively modest effort is the result of the efficient process created by PHS, particularly the custom reports that expedite the identification of semantically equivalent terms.

PHS is considering building an application to coordinate and further streamline the process of suggesting alternative terms, obtaining SME approval, and preparing mappings for replacement in various areas. However, replacing terms inactivated by periodic updates of the diagnosis dictionary is just one example of the overall effort to maintain the system configuration and data integrity of an EHR. The costs and resources required for maintenance and support, as well as the ability and expertise needed to handle frequent changes, may be the “Achilles heel” of complex EHR systems.

### References

A Six-Year Follow-Up of Clinical Decision Support System User Acceptability: Practice Makes Progress

Randall W. Grout, MD1, Erika R. Cheng, PhD, MPA1, Aaron E. Carroll, MD, MS2, Stephen M. Downs, MD, MS1, Nerissa S. Bauer, MD, MPH1
1Children’s Health Services Research, Indiana University, Indiana; 2Pediatric and Adolescent Comparative Effectiveness Research, Indiana University, Indiana

Introduction
Clinical decision support systems (CDSS) can improve care, but their efficacy is impacted by clinician acceptance. Prior analyses of user opinions in a pediatric CDSS demonstrated general acceptance. To our knowledge, long-term satisfaction results have not been reported among CDSS users in pediatrics. The current study examines user acceptance patterns over six years of continuous computerized CDSS integration and updates.

Methods
Child Health Improvement through Computer Automation (CHICA) is a CDSS in use for over 12 years at several urban clinics in Indianapolis, IN. CHICA gathers patient pre-visit screening data and computes a prioritized list of recommendations for providers using Arden Syntax medical logic modules (Figure 1). Providers can record responses to prompts to inform future prompts and research. CHICA was used in one health system (HS-E) with 5 clinics. Another health system (HS-U) joined in 2015. The CHICA team implements regular feature upgrades.

We administered anonymous surveys covering user characteristics and 12 core acceptability questions (Table 1). Respondents rated core questions on 5-point Likert scales. Surveys were distributed annually in HS-E between 2011 and 2016, except in 2013 (no funding). HS-U received the survey in 2016 only. We grouped core questions into function and usage domains. We collapsed responses into favorable vs. unfavorable responses with respect to CHICA (e.g., disagreeing with “CHICA makes lots of errors” is favorable). Neutral responses were coded as unfavorable. Providers were physicians or advanced practitioners; all others were non-providers. Frequent users were those who disagreed with “I rarely, if ever, use CHICA.” Intention to use CHICA was reported with “I would rather not use CHICA.” We used survey year, clinic role, part-time status, and CHICA use frequency as independent variables in multiple logistic regressions to predict a favorable response for each core question. Separately, we compared the first and last years of HS-E surveys with the HS-U survey to assess the effects of user experience and system maturity.

Table 1. Core survey questions.

<table>
<thead>
<tr>
<th>Domain</th>
<th>Questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Function</td>
<td>CHICA sometimes reminds me of things I otherwise would have forgotten. CHICA makes documentation easier. CHICA often makes mistakes. CHICA has uncovered issues with patients that I might not otherwise have found out about. I often disagree with the advice CHICA gives. The handouts CHICA produces are useful. CHICA makes lots of errors.</td>
</tr>
<tr>
<td>Usage</td>
<td>I rarely, if ever, use CHICA. CHICA tends to slow down the clinic. I would rather not use CHICA. CHICA has too many technical problems. Technical support for CHICA is very good.</td>
</tr>
</tbody>
</table>
Results

Of 354 surveys over five waves, 352 were returned. Full data were available for 93.8% (SD 0.8%) of core question responses. Compared to 2011, users responded more favorably for nearly all measures by 2016 (Figure 2); this was more heterogeneous in later years. Frequent CHICA users and providers reported favorable responses more often across several questions. Intention to use CHICA was significantly associated with frequent CHICA use.

Figure 2. Odds ratios of responding favorably to a core question. Transparent overlapping confidence intervals are below values. Reference year is 2011.

Twenty-eight survey responses from HS-U were compared to 71 (2011) and 75 (2016) from HS-E. Users with similar experience levels but a more mature CHICA implementation were more likely to report easier documentation and CHICA identifying missed issues. Users with the same maturity of software but less experience showed more unfavorable opinions in function and usage domains (e.g., worse technical support, rather not use CHICA).

Discussion

CHICA users were more favorable to both function and usage of the system over time. There was heterogeneity of user perspectives and patterns. Favorable opinions were more likely in frequent users, providers, and full-time employees. Dissatisfaction in 2015 may be attributable to the conversion from paper to electronic CHICA interfaces. Analysis with HS-U demonstrated that increasing system maturity was associated with improved CHICA function acceptance, and more user experience was associated with more favorable CDSS function and usage opinions.

These findings are consistent with acceptance models' where perceived usefulness (function domain) and perceived ease-of-use (usage domain) are key factors of behavioral intention and later use. A systematic review reported high short-term provider satisfaction for CDSS that were ambulatory, voluntary, locally developed, and included point of care recommendations – all of which CHICA fulfills. Experience with the CDSS and system quality improvement are both beneficial interventions associated with increased acceptance of clinical decision support systems.

References

Impact of VA Health Information Exchange upon the Overuse of Laboratory and Imaging Tests

David A. Haggstrom, MD, MAS; Laura Myers, PhD; Brian Dixon, PhD; Marc Rosenman, MD; Dustin French, PhD; Susan Ofner, MS; Susan Perkins, PhD; Michael Weiner, MD, MPH

1Center for Health Information & Communication, Department of Veterans Affairs, Health Services Research and Development Service CIN 13-416, Richard L. Roudebush VA, Indianapolis, IN; 2Department of Medicine, Indiana University School of Medicine, Indianapolis, IN; 3Regenstrief Institute, Indianapolis, IN; 4Fairbanks School of Public Health, Indiana University, Indianapolis, IN; 5Northwestern University Feinberg School of Medicine, Chicago, IL; 6Department of Biostatistics, Indiana University School of Medicine, Indianapolis, IN; 7Purdue University College of Pharmacy, Indianapolis, IN;

Introduction

Like most health care systems, the VA often has patients who receive some portion of their care outside the VA. Due to the lack of interoperability among electronic health record systems, providers caring for veterans who are receiving services outside the VA may not have timely access to important information about their patients. Health information exchange (HIE) advocates believe that greater interoperability will reduce redundant testing. In this study, measures of overuse were chosen which experts suggest may be sensitive to the effect of HIE.

Methods

Intervention: A regional VA-HIE demonstration program was conducted through a national initiative called the Virtual Lifetime Electronic Record (VLER). The VA-HIE performed secure, bi-directional health information exchange between the VA and community partners organized together through the Indiana Network for Patient Care. Health care providers inside and outside the VA were able to access exchanged data. Patients were enrolled in the VA-HIE program on-site at the Indianapolis VA in outpatient clinics or through the release-of-information office.

Study design: A pre-post cohort evaluation of participants in the VA-HIE program was performed, with a concurrent control group. For the evaluation, data on care received by patients inside and outside the VA were obtained for 1 year before, and 1 year after, the index date of patient enrollment. To allow for program implementation lag time and enable providers to become accustomed to using HIE data, the first study patients were enrolled in the cohort 3 months after VA-HIE implementation. All control patients were enrolled 3 months after VA-HIE implementation.

Population: Patients were included if they had at least 1 clinical encounter at the Indianapolis VA over a 1-year period prior to the index date. Overall, 57,050 patients were included in the cohort: 6,726 who were enrolled in the VA-HIE demonstration program and 50,324 in the concurrent control group. Further, subjects were required to have at least one test record in either the year before or the year after the index enrollment date.

Measures: For each laboratory test ordered over a 1-year period, a test was considered to represent potential overuse if the same test was already performed once within a short time interval before the second test according to the following criteria: 10 days (complete blood count (CBC) and renal profile) and 6 weeks (lipid profile and liver function tests). These time intervals were chosen based upon the prior deliberations of an expert panel.

For each imaging test ordered over a 1-year period, a test was considered to represent potential overuse if the same test was already performed once within 60 days before the second test for the following imaging procedures: computerized tomography (CT) scans, magnetic resonance imaging (MRI), positron emission tomography (PET) scans, and nuclear medicine studies.

Analyses: Multivariable regression models were constructed as a function of VA-HIE intervention status, time, and baseline covariates (age, gender, race/ethnicity, marital status, insurance, service-connected disability, distance traveled to the VA, Charlson comorbidity index, and the number of provider visits in the year before HIE enrollment).
Results

There were statistically significant difference-in-differences between the VA-HIE intervention and control groups in the change in rates of potential overuse, before and after the index enrollment date, across all categories of testing (Table 1). Before and after implementation of the VA-HIE program, there was greater reduction in the rate of overuse in the enrolled versus the control group for CBC and renal profile tests (-1.00% vs. 0.98%), lipid and liver function tests (-3.03% vs. 0.16%), and imaging (-0.77% vs. 0.55%).

Table 1: Rates of overuse by test type, multivariable regression model

<table>
<thead>
<tr>
<th>Test</th>
<th>Period</th>
<th>Control</th>
<th>Enrolled in VA-HIE</th>
<th>Difference between Control and Enrolled (95% Confidence Interval)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Rate (95% CI)</td>
<td>Rate (95% CI)</td>
<td></td>
</tr>
<tr>
<td>CBC &amp; Renal profile (10 days)</td>
<td>Pre</td>
<td>9.74% (9.47%, 10.01%)</td>
<td>12.02% (11.32%, 12.76%)</td>
<td>-1.98% (-2.93%, -1.05%)</td>
</tr>
<tr>
<td></td>
<td>Post</td>
<td>10.68% (10.40%, 10.97%)</td>
<td>11.05% (10.38%, 11.75%)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Δ</td>
<td>0.98% (0.64%, 1.34%)</td>
<td>-1.00% (-1.85%, -0.10%)</td>
<td>-1.98% (-2.93%, -1.05%)</td>
</tr>
<tr>
<td>Lipids &amp; Liver Tests (6 weeks)</td>
<td>Pre</td>
<td>12.70% (12.40%, 13.01%)</td>
<td>16.61% (15.76%, 17.49%)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Post</td>
<td>12.86% (12.56%, 13.17%)</td>
<td>13.64% (12.88%, 14.43%)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Δ</td>
<td>0.16% (-0.22%, 0.57%)</td>
<td>-3.03% (-4.02%, -2.03%)</td>
<td>-3.19% (-4.30%, -2.18%)</td>
</tr>
<tr>
<td>Imaging (60 days)</td>
<td>Pre</td>
<td>6.32% (5.96%, 6.69%)</td>
<td>6.81% (6.01%, 7.70%)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Post</td>
<td>6.79% (6.42%, 7.18%)</td>
<td>6.11% (5.36%, 6.96%)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Δ</td>
<td>0.55% (0.03%, 1.05%)</td>
<td>-0.77% (-1.92%, 0.33%)</td>
<td>-1.31% (-2.56%, -0.12%)</td>
</tr>
</tbody>
</table>

Discussion

Participation in VA-HIE reduced the ordering of laboratory and imaging tests at inappropriately short intervals in the ambulatory care setting. This finding is consistent with previous literature, suggesting that one of the positive impacts of health information exchange may be in the reduction of inappropriate or redundant testing. Those enrolled in VA-HIE had higher baseline rates of overuse, perhaps due to higher levels of illness than the control group. Nonetheless, the adjusted models accounted for both comorbidity and degree of outpatient utilization. The effect upon potential overuse was realized early, within the first year of implementation of the VA-HIE. Limitations of this study include our inability to assess the clinical indication for any single lab test. HIE appears to influence the ordering behavior of health care organizations, and future study can explore at whether this influence is exerted at the patient, provider, or system level.

References

Interactive Machine Learning for Medical Research:  
A Framework to Enhance the Engagement of Clinical Researchers

Bibo Hao, Yiqin Yu, Wen Sun, PhD, Yingxue Li, Guotong Xie, PhD  
IBM Research China, Beijing, China

Introduction

Machine learning (ML) and other big data analytics techniques have evolved increasingly in medical research. By applying ML techniques on various EHR data, data scientists and clinical researchers collaborated to build models to improve prognosis, improve diagnosis accuracy, improve the accuracy of predicting outcomes, etc.  

In the research of ML, considerable attention focused on automatic machine learning (aML), in which researchers try to automatize the whole ML life cycle and build systems that take data as input and generate classification or prediction model as output, with minimum user interaction. Pioneers in medical informatics research developed systems that take advantage of parallel computation to try and to select from different combinations of ML algorithms, strategies, and parameters. Despite the convenience of using these “black-box” aML systems, the ML pipeline and intermediate results are almost inaccessible and invisible to non-ML experts including clinical researchers. As a result, clinical researchers are out-of-loop in these systems and the ML pipelines.  

However, the engagement of clinical researchers plays a key role in medical research. It is essential for clinical researchers to truly understand the data analytics pipelines as well as properly interpret the models built from specific algorithm and data before they develop new methods for diagnosis, treatment, and prevention based on their understanding. In this study, we aim at developing a framework along with a system to support interactive machine learning (iML) that can enhance clinical researchers’ engagement in the ML or data analytics pipelines, contribute their knowledge in the analytics pipeline, and understand the pipeline like a “glass box”.

Methods

First, we break typical ML tasks in medical research into several stages and develop independent and reusable modules as an SDK for each stage, e.g., data loading, data exploration, cohort construction, feature engineering, risk modeling. We develop these modules in Python with open data science packages and including our customized algorithms and strategies. The SDK can then be used in scenarios of both aML and iML.

Second, we introduce and extend Jupyter Notebook, the popular web-based interactive development environment for data scientists, to meet their needs. In this extended Data Scientist View, data scientists can interactively perform analytic tasks by finding and calling the desired APIs of our SDK, and visualize intermediate data and results with pre-defined charts and tables. As designed in our system, a pipeline is defined by a Python class, in which each stage of the pipeline is a method of the class, and the intermediate data and results are member variables of the class. Such design makes it convenient to reuse the pipelines in aML scenario, which can run in a parallel environment with different algorithm and parameter combinations to search the best combination. Furthermore, data scientists can compose iML pipelines based on the Python class, by easily adding customized interactive HTML widgets in the same view and same notebook file. In the iML pipeline, executions of stages are invoked by the interactive HTML widget events.

Third, apart from the extended Data Scientist View, we developed a Clinical Researcher View that allows clinical researchers to use the system with a user-friendly web application like a wizard and does not require users to write code. Given an iML pipeline notebook file developed by data scientists in Data Scientist View, our system will automatically display the pipeline in Clinical Researcher View as a wizard-like user interface. In this view, users can run each step of a pipeline by fill HTML forms and click buttons. What’s more, users can pause at any step of the pipeline execution and inspect the value of intermediate data and results, like inspect the pipeline as a “glass box”.

Results

The screenshots of the Data Scientist View and Clinical Researchers View of a risk prediction pipeline are illustrated in Figure 1 and Figure 2 respectively. After data scientists composing the risk prediction pipeline in a Jupyter Notebook file, our system converts it to an interactive risk prediction task without code in Clinical Researcher View.
Data Scientists View provides an environment for data scientists to compose ML pipelines with tremendous level of flexibility, which 1) gives the data scientists full control to the pipeline, including put the pipeline in an aML use case, 2) allow data scientists to make use of their expertise to define and develop interactions for the clinical researchers.

Clinical Researchers View enhances clinical researchers’ engagement in the iML pipelines, where they can interactively set desired parameters, select or drop desired variables, inspect intermediate results, and of course, go back to previous steps to make a different choice and see the difference in results.

By collaborating in two views of our system, data scientists and clinical researchers can formulate a framework to enhance clinical researchers’ engagement in the ML loop, without sacrificing the pipelines’ use in the aML scenario.

Discussion

We developed a framework along with a system that can enhance clinical researchers’ engagement in the interactive machine learning use case. Meanwhile, data scientists can also use the pipelines in an automatic machine learning scenario. We hope this framework and system can make machine learning more approachable in medical research.

References

Gene annotation bias impedes biomedical research

Winston A. Haynes¹, Aurelie Tomczak¹, and Purvesh Khatri¹
¹Stanford University, Stanford, California, USA

Introduction

After analyzing samples with a high throughput technology, the de facto first step is to perform pathway or network analysis to identify biological processes that are statistically enriched in the data[1]. Researchers typically form hypotheses for their follow up experiments based on the genes or proteins involved in the enriched processes. Since the annotation resources commonly used for identifying gene functions are created by curation of the scientific literature, they typically only contain functional annotations for genes with published experimental data. Consequently, researchers select those genes or proteins for further validation that have prior experimental evidence, which, in turn, leads to more functional annotations for those genes at the expense of under-studied genes.

We hypothesized that this experimental paradigm has led to a gene-centric disease research bias where hypotheses are confounded by the streetlight effect. To test this hypothesis, we examined the annotation inequality for the human genome across a number of biomedical databases using gini coefficient, a measure of inequality such that high coefficient value indicates higher inequality[2].

Methods


Gene expression data collection and meta-analysis. Gene expression meta-analysis data was compiled from the MetaSignature database [12], which includes over 41,000 samples, 619 studies, and 104 diseases. Cases and controls were manually labeled for each disease and meta-analysis was performed using the MetaIntegrator package [12].

Data collection for disease-gene publications and SNP data. We downloaded the number of publications for each disease-gene relationship from PubPular [5, 6]. From Gene Ontology, we calculated the counts of non-Inferred from Electronic Annotation annotations for each gene [4]. For all correlations, we only considered disease-gene associations with at least 10 publications to limit false positive associations.

Results

Despite the tremendous growth of Gene Ontology (GO) from 20,826 annotations in 2004 for 7,524 human genes to 122,926 annotations for 16,173 genes in 2017, annotation inequality in GO has increased from a gini coefficient of 0.34 in 2004 to 0.50 in 2017. The growth in inequality over time validates that genes with existing annotations continue to receive even more annotations. Pathway databases, including Reactome[7] and CTD[8], have a similarly high level of inequality. Indeed, every gene annotation resource we examined displayed a similarly high level of annotation inequality, including: CTD chemical-gene associations[8]; PDB 3D protein structures[9]; DrugBank druggene associations[10]; GeneRIF gene publication annotations; and Pubpular disease-gene publication associations[5, 6]. We observed that the inequality index for many of the gene resources is higher than income inequality in any nation in the Organisation for Economic Co-operation and Development (OECD)[11].
We explored whether disease research may be affected by the inequality in gene annotation databases. In our manually curated meta-analyses of 104 distinct human conditions, we have integrated transcriptome data from over 41,000 patients and 619 studies to calculate an effect size for disease-gene associations [12]. Our meta-analyses covered diverse classes of human conditions, such as cancer, autoimmune disease, and infectious diseases. Published disease-gene associations exhibited no significant correlation with gene expression false discovery rate (FDR) rank [Fig 2A, cor=-0.005, p=0.7].

Based on these results, we hypothesized that the lack of correlation with molecular evidence may have been an artifact of research bias towards well-characterized genes. Therefore, we examined correspondence between publications about a disease-gene pair and existing knowledge about that gene as indicated by the number of GO annotations. The number of GO annotations for a gene was significantly correlated with the published disease-gene associations [Figure 2B, cor = 0.100, p=8.7e-13], but not with gene expression effect size FDR rank in disease [cor = -0.010, p = 0.136][4].

Discussion

Collectively, our results provide evidence of a strong research bias in literature that focuses on well annotated genes instead of the genes with the most significant disease relationship in terms of both gene expression and genetic variation. Despite the rise of high throughput technologies, annotation inequality has continued to grow over time. While focusing research on the best characterized genes may be natural because it is easy to formulate a mechanistic hypothesis of the gene’s function in disease, we propose that omics-era researchers should instead allow data to drive their hypotheses. To enable researchers to pursue data-driven hypotheses, we have made our gene expression meta-analysis data publicly available at [http://metasignature.stanford.edu]. By focusing on genes with the strongest molecular evidence instead of the most annotations, researchers will break the self-perpetuating annotation inequality cycle that results in research bias.

References

Implementing Automated Data Reporting to Improve Early Hearing Care

Shan He, PhD\(^1\), Christine Perfili, MBA\(^2\), Darren K. Mann, BS\(^1\), Matt Hoffman, MD\(^3\), Viral Shah\(^1\), Gary Anderson\(^1\), Rishabh Bhatia\(^1\), Sidney N. Thornton, PhD\(^1,4\)

\(^1\)Intermountain Healthcare, Salt Lake City, UT; \(^2\)Utah Department of Health, Salt Lake City, UT; \(^3\)Utah Health Information Network, Salt Lake City, Utah; \(^4\)University of Utah, Salt Lake City, UT

Introduction

With the widespread implementation of Early Hearing Detection and Intervention (EHDI) programs across the United States, more than 97% of newborns were screened for hearing loss, usually before hospital discharge\(^1\). However, screening is only the first step. Diagnostic follow-up testing for infants not passing the screening before discharge is essential to confirm if an infant has hearing loss and to ensure early intervention. Generally, the initial hospital-based newborn hearing screening results are reported to the jurisdictional EHDI program on a consistent basis. However, the reporting of the follow-up diagnostic evaluations is inconsistent\(^2\). In Utah, diagnostic evaluation reporting from healthcare providers to the Utah EHDI program has been manual and paper-based. The process was time consuming with no guarantee that a diagnostic report was received by EHDI. National data indicate about 35% of infants failing the hearing screening were not documented to have received the recommended evaluation. These children are classified as loss to follow-up (LFU) if they did not receive the follow-up diagnostic services, or classified as loss to documentation (LTD) if they received services without the results being reported to the EHDI program\(^2\). The LFU/LTD rate for Utah is about 13% according to the 2013 data. Improving follow-up diagnostic data reporting is considered to have noticeable impact on reducing LFU/LTD rates and achieving optimal newborn hearing care.

Methods

With funding support from the Office of the National Coordinator of Health IT (ONC), the Utah Department of Health collaborated with Intermountain Healthcare and the Utah Health Information Network (UHIN) to implement a solution that enables real-time, automated diagnostic data reporting from Intermountain Healthcare to the state EHDI program.

As shown in Figure 1, three major components comprise our implemented solution. First, event-based triggers were implemented in Intermountain’s Electronic Health Record (EHR) system to monitor infants who received diagnostic hearing testing and evaluation. Two types of events automatically fire the trigger: 1) finalization and storage of any type of diagnostic audiology report; and 2) storage of a cytomegalovirus (CMV) laboratory result. Congenital CMV is the leading non-genetic cause of childhood hearing loss. By working collaboratively with the EHDI professionals and audiologists, a total of twelve diagnostic audiology report types including the most commonly used Audiology Auditory Brainstem Response (ABR) Report and seven CMV laboratory results were identified in the EHR and configured in the trigger code list. Second, when the trigger criteria are met, the system automatically generates a standard C-CDA (Consolidated-Clinical Document Architecture) document. We chose the Progress Note template for this use case because it best suits the clinical scenario and the information need by EHDI without requiring the irrelevant data elements in other templates. Since the audiological reports were stored as free-text notes or scanned
images in the EHR, the C-CDA Progress Note was generated by embedding the entire clinical note as a media object in the “Assessment and Plan” section. Patient demographics and provider information were included in the header section and CMV laboratory results were listed in the “Results” section mapped to standard LOINC codes. Finally, the C-CDA Progress Note was sent securely to the Direct email box for the state EHDI program. Only reports on children under 6 years old are considered relevant and transmitted to the state EHDI.

Results

The automated reporting of newborn hearing diagnostic evaluation from Intermountain to the state EHDI program went live on September 2016. Between October 2016 and February 2017, a total of 1,895 C-CDA Progress Notes on 1,021 distinct patients were sent to EHDI through Direct messages. Among these documents, 1,237 (65%) were triggered by a diagnostic audiology evaluation note such as an ABR report and 658 (35%) were triggered by a CMV laboratory result. About 6% of patients have documents triggered by both types of events.

Since the C-CDA documents were generated and sent in real-time without the need for manual fax, the EHDI professionals were able to review the evaluation results as soon as they were saved in the EHR. The LTD rate for Intermountain infants diminished to zero. With faster access to the diagnostic reports, the EHDI professionals are now able to coordinate with healthcare providers and make early interventions whenever appropriate. For example, during our study period, the EHDI program identified a case where the CMV laboratory test was not ordered with the correct specimen type by the pediatrician. The EHDI staff then contacted the ordering physician to suggest the appropriate test. As a result, the infant received the reliable diagnostic result and proper treatment within a 21-day time frame, which is significantly faster than the 1-3-6 benchmarks recommended by the Joint Committee on Infant Hearing4. The standard-based electronic Progress Note also enhanced the efficiency of the Utah EHDI staff by reducing the processing time of received diagnostic reports from 3 hours per week to 1 hour per week.

Discussion

Our electronic reporting solution provides the opportunity for EHDI professionals to have real-time access to diagnostic evaluation results at Intermountain, which is the largest integrated healthcare delivery system in Utah. The project expects to accurately and timely identify the loss to follow-up patients and to increase the percentage of infants enrolled in early interventions prior to six months of age from the current 77.9% to over 90%. However, getting the data flow automatically from healthcare facilities to the EHDI program is only the first step toward an integrated public health and healthcare system. The current phase of the project requires EHDI staff to manually upload the electronic documents to the newborn hearing surveillance and tracking system (Hi-Track), which manages initial newborn hearing screening results reported by hospitals. The next phase of the project is to collaborate with the CDC EHDI team to develop standard code sets and implement the HL7 CDA standard for Early Hearing Care Plan (EHCP)3. Our early experience working with the state EHDI program to understand their data needs will contribute to the development of the new industry standards. With structured and coded early hearing information, an integration service with Hi-Track will be developed to automatically link the follow-up hearing results and interventions to Hi-Track records. A comprehensive view of a child’s hearing care will be available in the statewide EHDI hearing registry.

Conclusion

Accurate and timely hearing screening and follow-up data reporting from hospitals, audiologists, and other providers to the state or territorial EHDI program is crucial to ensure early detection and intervention for deaf and hard of hearing (DHH) infants. Our automated and standard-based electronic transmission of audiology reports can greatly improve communication and coordination between multiple caregivers. When the solution is implemented statewide, the LTD rate will be significantly reduced to near zero and the LFU can be accurately calculated based on the number of infants who failed newborn hearing screening but did not have a diagnostic evaluation received by EHDI.

References

Integrating Patient Registries into Enterprise wide Patient-Discovery Strategies

Christopher Herrick MBA¹, Alyssa Goodson MS¹, Wayne Chan¹, Lori Phillips MS¹, Michael Mendis¹, Shawn Murphy MD, Ph.D.¹,²,³

¹Partners Healthcare, ²Massachusetts General Hospital, ³Harvard Medical School, Boston, MA

Abstract: Some of the most valuable patient data that is available for research exists in curated disease registries. Traditionally these registries are not included in enterprise research patient query systems such as Informatics for Integrating Biology and the Bedside (i2b2) (Murphy, Weber et al. 2010). Because the effort involved in systematizing their ontologies and formatting their data into a common information model is very labor intensive Given that a solution to this problem would be a valuable asset, we developed the ability to import registries in a mostly automated fashion such that a registry owner could integrate the registry using a self-serve mechanism, achieving the integration of big and little data.

Introduction: Considerable time is often devoted to creating a set of well characterized patients, characterized by data derived from the electronic medical record or newly collected during an encounter with the patient. When the data is collected in a highly systematized manner, often congruent across multiple sites, this collection is known as a “Registry”. Given the time and effort that go into creating these collections of well characterized patients, it is puzzling why they are usually not included in enterprise research query systems such as i2b2. There appears to be at least three reasons for this absence:

1) They are often not created with standard ontologies and so do not “fit in” to many enterprise query tools.
2) The individuals who create and manage them are often not savvy to how such a database would be integrated into a structured research database such as i2b2.
3) Even if the registry owners were aware of how they could be integrated into research databases, the process of integrating many externally managed, custom collections of patient data would overwhelm the team of engineers who need to extract, transform, and load data into the enterprise research database.

To overcome these obstacles, we developed software in the BD2K Center of Excellence “PIC-SURE” which can manage the integration of patient data registries into a larger i2b2-based enterprise research database, allowing the registry data to be completely interoperable with the data obtained from the electronic medical record data (http://www.pic-sure.org).

Methods: The software solution for integrating patient registries into an enterprise query system was based upon a tool named “BRISSKit” (Butters, Issa et al, 2016) which was developed by a group in the United Kingdom. The BRISSKit software allowed a table of external data to be formatted into the star schema of an i2b2 database and the ontology to be extracted from the data. Integration of the data was then possible through the PIC-SURE distributed query system. Briefly, the overall system is architected to have the software from i2b2 act as a supervisory layer on top of various systems that can be formulated into presenting a structured observation-fact table. A workflow is set up to query and return sets of patients and structured data from the various systems through the i2b2 ontology-driven query system. Administrative and regulatory details regarding use of the data in each Data Repository is governed by the i2b2 privacy layer that constrains the use of the data to comply with HIPAA and HHS regulations (Murphy, Gainer et al. 2011). Permissions for each user are defined as part of a project which governs the access to data across the systems.

Results: A data format that is common to many data registries, or which many registries have ready-made tools to create, is the n x m common table, with rows representing each patient, and columns representing features of a patient. A common variation is to have a time-dependent encounter represented in the rows of the table, and features that the patient had during the encounter represented in the columns. These types of tables can be imported into the enterprise query system in three steps. The first step is to create the proper regulatory environment to publish the data. The data is uploaded with the IRB determination that data was collected with a broad consent or a waiver that allows use within
the enterprise. Although identifiers on the patients must be uploaded with the data to allow proper linkage to enterprise medical record numbers, these identifiers are not exposed to researchers in aggregate queries performed using the i2b2 query tool. With a level of permission in the i2b2 query tool that allows a matrix of patient data to be downloaded, this can be done under a data use agreement (for a limited data set), or an IRB protocol (for an identified data set).

The second step is to allow an ontology to be automatically generated for the registry. This is done by uploading and then analyzing the columns of the registry. The elements in the columns are determined to be values, and they are compiled from each column and determined to be numeric or enumerated. The title of the column is used in the ontology to specify the name of the concept, and if the concept is numeric a summary of high, low, average, etc. is retained to guide a query. If the concept is enumerated the values are determined to be an existing coding system (SNOMED for example), or an invention. Inventions are compiled as new sets of enumerations, and the opportunity is given to map them to existing coding systems. However, they are not forced to map them and they can still be used in queries.

The third step is to create formatted set of data automatically that integrates with enterprise queries. Each transformed table exists independently and may have several different versions, but these are formulated into the query results such that even an enterprise system with 5 billion rows can integrate an almost unlimited number of registries.

Discussion: Data from a healthcare system includes that from the medical record, but also from carefully maintained patient registries. We developed the ability to import registries in a mostly automated fashion such that a registry owner could integrate a registry using a self-serve mechanism into an enterprise wide system that supports research queries for patient cohorts and complex analytics. A significant regulatory problem was that many patient registries have their patients consented such that the data collected in the course of creating the registry cannot be made discoverable across the enterprise. The solution to this problem was to allow registry owners to create projects such that only they could see the registry data but could perform queries across all of the enterprise data. Although more limited in scope, for the registry owners this was extremely valuable because they could easily combine the registry data with all of the Enterprise electronic health care data in this fashion. Furthermore, with specific types of consents, several registries could be combined into one project and queries could be performed across registries. Overall, this allows academic medical centers that implement the open source PIC-SURE i2b2 system to rapidly and flexibly combine data from patient registries with electronic health care data and present it to their investigators in self-serve analytics.

This work was conducted with support from National Institutes of Health Centers of Excellence for Big Data Computing in the Biomedical Sciences (U54), Grant Number 1U54HG007963-01


Interpretable Clustering for Prototypical Patient Understanding: A Case Study of Hypertension and Depression Subgroup Behavioral Profiling in National Health and Nutrition Examination Survey Data

Pei-Yun S. Hsueh, PhD1, Subhro Das, PhD2
1IBM Watson Research Center, Yorktown Heights, NY, USA

Introduction

Behavioral factors are the key contributors to morbidity and mental health risk [1]. For cardiovascular diseases alone, behavioral factors have contributed to 41 percent of its global disease burden [2]. In recent years, clinical guidelines (e.g., the 2017 Type II Diabetes management guideline [3]) have established the need to better understand patient behavior and identify patient-centered goals for care plan personalization. However, such guidelines for personalization are usually not formalized and will not grow organically with the increasing knowledge of patients. The key question is, therefore, how to further inform personalized care plan by accurately and reliably quantifying the behavioral patterns from target user-generated health data?

The recent emergence of health consumers and person health IT technologies made available an unprecedented amount of patient-generated health data (PGHD) [4]. To augment the current guideline-based care plan with PGHD insights, we introduce an interpretable clustering method for uncovering outcome-differential behavioral patterns in distinctive patient subcohorts. In addition, to further enable case-based reasoning for subcohort behavioral profiling, the proposed method also explores adding the identification of prototypical users for sense-making explanation and persuasion. Curating the learned subcohort profiles has been expected to facilitate a hybrid approach of PGHD-augmented guideline deployment.

Methods

Using the self-reported outcome survey and sensor measurement data in National Health and Nutrition Examination Survey (NHANES) during 2005 and 2006, we deployed an interpretable clustering method to identify distinctive behavioral profiles related to two types of proxy outcome indicator: blood pressure control (BP) and depression. For the first BP study, we follow [5]'s footstep to use a set of 64 features generated from the accelerometer measurements, blood pressure readings and questionnaire items of 5,695 subjects. The set of features include physical activity intensity, duration, variance of activity, systolic and diastolic BP, and so on. Similarly, for the second Depression study, we use only the set of 38 features generated from the accelerometer sensor readings.

We evaluate the proposed Interpretable Clustering (IC) method that segments subjects into different behavioral segments based on their behavioral patterns and proxy outcomes. Our hypothesis is that although each feature contains only weak behavioral signals for the whole population, when considered collectively in certain sub-cohorts, these features can explain rich behavioral patterns that matter to some proxy health outcome indicator, e.g., whether a user scores less than 4 (minimal depression or none) using the Patient Health Questionnaire (PHQ-9).

In particular, we first apply the Locally Supervised Metric Learner (LSML) [6] of patient similarity analytics to estimate the outcome-adjusted behavioral distances between the users. Then, based on the adjusted behavioral distances, hierarchical clustering is employed to generate sub-cohorts and learn the key features (which contain behavioral signals about implicit user preferences and barriers) that drive the differential outcomes. An automatic tuning algorithm is applied to determine the optimal number of segments. In addition, as we expect it to be easier to interpret patient need from behavior profiles by examples, the proposed method also includes a component to identify prototypical examples (i.e., a set of top 10 subjects who are the closest to the centroid of each behavioral segment in the outcome-adjusted distance space).

Results

For both the BP and Depression datasets, five distinctive sub-cohorts are identified, each appearing in more than 1% of the subjects. Statistical and information theoretic measures are applied to infer cross- and within-subcohort behavioral patterns among subjects and identify 3-10 distinctively informative patterns for each subcohort.

In order to allow for evaluation, both global and personalized predictive models are constructed. While the former is trained on the whole population, the latter is trained for each sub-cohort. In addition, we also compare the outcome-
differentiating Interpretable Clustering (IC) method with K-means clustering and Agglomerative clustering method that do not directly account for outcome differentiation during the clustering process. The evaluation includes clustering quality, i.e. Silhouette score that indicate internal consistency by measuring how similar an object is to its own cluster (cohesion) compared to other clusters (separation), and outcome-differentiating F-statistics and z-scores. Results show that the IC-based personalized prediction model yields significantly better outcome differentiation results in terms of the sub-cohorts it produced. Although the personalized IC models did not produce more accurate risk scores than the general model, it captures more intuitively understandable patterns. Behavioral profile evaluation metrics show that further zooming into the prototypical examples helps identify distinctively more different profiles across the sub-cohorts, as demonstrated in the differences in the top behavioral patterns and those between the individual and global risk factors.

Discussion

The consumer and pervasive health informatics community is increasingly handicapped by the problem of not being able to interpret patient need. Previously, a family of patient phenotyping approaches (including the use of deep learning) has been developed to extract distinctive phenotypes from medical records [7,8] and “digital phenotypes” from digital data (including device/sensor data and self-reported outcomes) [9,10]. One missing key is a behavioral learning mechanism that can sift through user-generated health data to identify outcome-differential patient behavioral patterns.

To address this important problem, this presentation has conducted a 3-fold investigation: First, it has reviewed existing guidelines and data-driven methods of patient subgroup identification and digital phenotyping and identified gaps for realizing the personalization goal of guidelines. Secondly, this presentation also proposed an interpretable clustering approach that adds behavioral response pattern understanding into patient subgroup analysis and informs guideline deployment. Lastly, this presentation provided evaluation by scanning through NHANES data to identify behavioral profiles and prototypical examples related to blood pressure control and depression.

The result demonstrates a large potential for learning methods to derive outcome-differentiating patient behavioral insights, which in turn lend support to better decision tools for care team and facilitate positive changes in patient behavior. In addition, the proposed method also improves on the interpretability of the learning results by providing more distinctively different behavioral patterns and prototypical examples. Finally, by curating the detected sub-cohort and differential behavioral patterns, a natural next step would be to investigate whether we can better deploy guideline-based personalized care plans with better individualized interpretations.

References

How do physicians read electronic progress notes?

Gretchen M. Hultman, MPH\(^1\), Jenna L. Marquard, PhD\(^2\), Swaminathan Kandaswamy, MS\(^2\), Elizabeth Lindemann, BS\(^3\), Genevieve B. Melton, MD, PhD\(^1,3\)

\(^1\)Institute for Health Informatics, \(^3\)Department of Surgery, University of Minnesota, Minneapolis, MN, \(^2\)College of Engineering, University of Massachusetts, Amherst, MA

Introduction

Clinical notes are a vital part of the modern electronic health record (EHR). One important type of clinical note is the “progress note“, which typically contains four sections: subjective (S), objective (O), assessment (A), and plan (P)\(^1\). Electronic progress notes are often lengthy and contain information thought to be extraneous, irrelevant or inaccurate\(^2\). While the Assessment and Plan section is often considered most valuable to clinician, finding this information often requires users to scroll through lengthy supporting data\(^3\). There is currently a debate about whether notes should be ordered differently from the traditional SOAP format, placing narrative text earlier in the note (i.e., APSO). In practice, notes often vary in their order, due to the prevalence of customizable EHR note templates. The purpose of this study was to use eye tracking technology to determine the impact of note section order on how clinicians read electronic progress notes. If we can understand how presentation order affects reading, we can better identify strategies for improving clinical note design.

Methods

This study took place at the University of Minnesota and utilized a previously described EHR system prototype\(^4\). Second through fourth year internal medicine resident participants and fellows (n=8) were seated at a desktop computer and asked to review four patient cases as they normally would and provide a brief verbal summary of the case. Each participant reviewed the cases in the same order, but the order in which the sections were presented within the nine notes was randomized across the four patient cases, using the following four orders: all SOAP, all APSO, all SAPO and Mix (3 notes of each specified order). During the session, participants wore eye tracking glasses (SMI, Inc.), calibrated at the beginning of the session, that recorded both audio and video of what the participant said and looked at. We used screen capture software (TURF) to record interactions with the EHR. The eye tracker videos were coded by noting when and for how long participants look at each section (S, O, A, and P). Two researchers coded 25% of the eye tracker videos and resolved all discrepancies, to determine a standard coding process. One researcher coded the remaining videos. Using the coded data, we calculated three metrics for each of the four note orders: 1) the average amount of time into the case when participants first looked at each section (time to first fixation), 2) how long it took participants to read each section, normalized by the number of characters in the section (dwell duration), and 3) how many times participants looked at each section (number of glances).

Results

Figure 1 shows that for all four note orders, time to first fixation was lowest for S, followed by O, and highest for A and P, indicating that participants initially located and read information in the same order. When the sections were in all APSO, however, participants reached the A and P sections much more quickly than in the other note orders. Figure 2 shows the average time per character spent reading each section, noted as the average dwell duration. Patient cases with notes in SOAP format had the lowest dwell duration (in seconds/character), meaning there was more scanning of those notes, or that they were read the most quickly. Across all note orderings, participants read the S section the slowest (largest average dwell duration). Participants read the A section more slowly when the notes were all in APSO or all SAPO. Figure 3 shows the number of total glances at each section. When notes were all in SOAP order, participants had the fewest number of total glances, indicating that participants shifted between sections less frequently in this note order.

Discussion

We found several insights by using an eye tracker to assess how note section order impacted note reading. First, it appears that participants initially looked at the sections in the same general order (S, O, A/P),
regardless of the physical order of the note. This initial reading pattern may be due to participants’ experience reading notes in SOAP format. This implies that simply rearranging notes may not cause individuals to change their initial information search patterns, but that reading patterns may be driven by their existing practices. However, participants skipped to the A/P sections more quickly when they were presented physically at the top of the note. This suggests that APSO format may aid individuals in quickly finding the A/P sections. Participants had fewer overall glances at each section when all notes were presented in SOAP order, perhaps because this ordering is familiar and they do not have to scan back and forth across the note sections. Overall, this study provides preliminary data suggesting that note reading patterns are affected by both individuals’ existing practices, and the physical layout of the information. This study only included one user group at a single institution, and it is possible that other user groups, such as attending physicians, would be affected differently by changes in note section order. Future work should explore whether users from different specialties or with varying levels of experience are impacted differently by note ordering. Future work should also explore the impact of note order in real world settings, where clinicians may be familiar with the patient cases. Research should also explore if giving participants training or a brief “primer” on different note orders impacts their experience reading notes.

Acknowledgments

The Agency for Healthcare Quality and Research (R01HS022085) and the National Science Foundation (1150057) supported this work. The content is solely the responsibility of the authors and does not represent the official views of AHRQ and NSF.

References

(1) Weed LL. Medical records, medical education and patient care: the problem oriented record as a basic tool. 1971.

Figure 1: Average time to first fixation Figure 2: Average dwell duration

Figure 3: Number of glances
Using Social Networking Analysis to Understand the Importance of Patients and Caregivers in Online Teams of Care

Trevor Jamieson, MD MBI1 2 3, Michael Chiang, MD MA3, Teja Voruganti, MD/PhD (c)4, Allison Kurahashi, BSc Med5, Alyssa Bertram, BA5, Sara Subramaniam, MA5, Amna Husain, MD MPH1 5

1University of Toronto, Toronto, Ontario; 2Women's College Hospital Institute for Health System Solutions and Virtual Care (WIHV), Toronto, Ontario; 3Oregon Health & Science University, Portland, Oregon; 4Institute for Health Policy Management and Evaluation, University of Toronto, Toronto, Ontario; 5Temmy Latner Centre for Palliative Care, Mount Sinai Hospital, Toronto, Ontario

Introduction

Social Networking Analysis (SNA) is a discipline rooted in graph theory, sociology and anthropology that allows mathematical analysis of the relationships between actors in a network. SNA is advocated by the Agency for Healthcare Research and Quality (AHRQ) as a way to improve the understanding of behavior in healthcare coordination networks. 1

LOOP is a social networking application that brings together cross-professional, cross-institutional teams of care in the ambulatory environment; patients and caregivers (PtCG), where caregivers in this case are exclusively family members, exist as full participants. LOOP does not allow directed messaging; all messages are visible by all participants within that care team. A LOOP administrator, a research assistant in the study, participates in each team as a technical support resource (other participants are not members of the research team). We used SNA to assess the centrality and roles of PtCGs across 18 teams in the intervention arm of a recently completed, in press, randomized controlled trial of LOOP in patients with advanced malignancy. Teams participated for a maximum of 3 months from enrollment.

The “importance” of an actor can be assessed by centrality, represented by two metrics: degree (the number of outbound and inbound ties) and betweenness (the frequency with which an actor lies on the shortest path between other actors). 3 Actors with high degree are “hubs”, acting as a focal point for interaction, whereas those with high betweenness are “brokers”, acting as key players in the information transmission across the network. Communities of practice, like LOOP, can be defined by four social networking characteristics: density (the percentage of ties relative to the maximum number of ties), distance (the average distance along the shortest path between nodes), fragmentation (the degree to which the graph is a connected whole), and coreness (the extent to which a network has “core” actors who densely participate with each other versus “peripheral” actors who participate only with the core but not with each other). 4 The coreness of a particular actor in a network is the degree to which that actor acts as a member of the core versus the periphery. Based on prior exploratory work during LOOP beta-testing, we hypothesized that PtCGs would be highly central team members; in particular that they would show higher outdegree, betweenness and coreness than other participants, and that teams in which they participated would be more dense and connected than those in which they did not.

Methods

Audit trail messaging data from LOOP was used to create “care” networks for the 18 teams. For each message, a directed tie was considered to exist between actor A and actor B if A’s message either (a) was marked “attention to” B, (b) mentioned B by name in the text, or (c) was a reply to a message from B. Messages could generate ties to multiple actors and messages that were not about patient care (simple salutations, technical support requests from LOOP administrators) were excluded. Ties were weighted, and each additional message meeting criteria would increase the tie strength by one. External actors, i.e. those mentioned but not having an account in the system, were included; these actors were offered participation in the system but were either not contactable or refused. Impossible ties (across teams) were accounted for and not considered in the analysis.

Contingency table analysis is a standard SNA technique that assesses whether the observed patterns of interaction between block-reduced (aggregated) roles in a network differ from the mathematically expected patterns of interaction given the network density and number of actors in each role, assuming equal probabilities of interaction between any two nodes. 5 We conducted a contingency analysis across the aggregate 18-team network to assess
whether the patterns of interaction between PtCGs and healthcare providers (HCPs; predominantly medical/radiation oncologists, palliative care physicians, and home and clinic nursing staff) differed from what would be predicted; LOOP administrators were excluded. To assess the centrality of PtCGs, we compared their mean out-degree, in-degree, betweenness and coreness within their respective teams against all other team members. To assess the impact of PtCG participation on team function, we compared the density, distance, fragmentation and coreness of teams in which patients or caregivers did or did not participate.

All analyses were completed in UCINET 6.587, and comparisons used bootstrapped t-tests over 10000 iterations.

**Results**

Teams had an average of 4.6 non-administrator members, including 1.4 PtCGs and 3.2 HCPs. 74 messages were sent in total, 4.1 per team (range: 0-21). Contingency table analysis shows significant differences in the observed interaction patterns versus what would be expected (p=0.0097, Table 1). PtCGs interact with HCPs more than expected (observed/expected, O/E=1.88), and HCPs interact with patients (O/E=0.92) and with each other (O/E=0.71) less than expected. Patients have a significantly higher out degree (p=0.0071), betweenness (p=0.0037), and coreness (p=0.0004) than other actors in the care teams (Table 2). Teams in which PtCGs participate are more dense (p=0.0015), have lower average distance between actors (p=0.0023), are less fragmented (p=0.0009), and have a higher average coreness (p=0.0007) than those where PtCGs do not participate (Table 3).

**Discussion**

In LOOP, we see a pattern of asymmetric communication between PtCGs and HCPs. PtCG are central, acting in both hub and brokerage capacities, and their participation is significantly related to more tightly interacting and less fragmented teams. Tools like LOOP should consider PtCG to be integral participants. Key limitations of such an analysis are the inability to know from audit trial data whether communication is occurring outside the tool and how these patterns of communication relate to patient oriented outcomes. Further studies should assess the incentive structures for HCP participation in tools like LOOP.

**Tables/Figures**

Table 1. Observed/expected interaction between PtCG and HCPs (contingency analysis of entire network, p=0.0097)

<table>
<thead>
<tr>
<th></th>
<th>PtCG</th>
<th>HCP</th>
</tr>
</thead>
<tbody>
<tr>
<td>PtCG</td>
<td>0.42</td>
<td>1.88</td>
</tr>
<tr>
<td>HCP</td>
<td>0.92</td>
<td>0.71</td>
</tr>
</tbody>
</table>

Table 2. Mean centrality measures for PtCG within teams compared to other providers (Rest) across entire network.

<table>
<thead>
<tr>
<th></th>
<th>PtCG</th>
<th>Rest</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Out Degree</td>
<td>1.68</td>
<td>0.67</td>
<td>0.0071*</td>
</tr>
<tr>
<td>In Degree</td>
<td>0.93</td>
<td>0.96</td>
<td>0.89</td>
</tr>
<tr>
<td>Betweenness</td>
<td>0.79</td>
<td>0.18</td>
<td>0.0037*</td>
</tr>
<tr>
<td>Coreness</td>
<td>0.36</td>
<td>0.10</td>
<td>0.0004*</td>
</tr>
</tbody>
</table>

Table 3. Mean team metrics for teams where PtCG participate versus those where they do not.

<table>
<thead>
<tr>
<th></th>
<th>PtCG Participates</th>
<th>No PtCG Participation</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Density</td>
<td>0.27</td>
<td>0.048</td>
<td>0.0015*</td>
</tr>
<tr>
<td>Distance</td>
<td>1.27</td>
<td>3.00</td>
<td>0.0010*</td>
</tr>
<tr>
<td>Fragmentation</td>
<td>0.63</td>
<td>0.95</td>
<td>0.0009*</td>
</tr>
<tr>
<td>Coreness</td>
<td>0.25</td>
<td>0.052</td>
<td>0.0007*</td>
</tr>
</tbody>
</table>

**References**

Participatory Design of Probability-Based Decision Support Tools for In-Hospital Nurses

Alvin D. Jeffery, PhD, RN, 1 Laurie L. Novak, PhD,2 Betsy Kennedy, PhD, RN,3 Mary S. Dietrich, PhD,3 Lorraine C. Mion, PhD, RN4
1TN Valley Healthcare System, U.S. Department of Veterans Affairs, Nashville, TN, USA
2Department of Biomedical Informatics, Vanderbilt University, Nashville, TN, USA
3School of Nursing, Vanderbilt University, Nashville, TN, USA
4College of Nursing, The Ohio State University, Columbus, OH, USA

Introduction

The value of clinical decision support tools in general is increasingly recognized; however, gaps still exist in understanding interactions between users and these tools.1, 2 Notably, probability-based clinical decision support (PB-CDS) tools (referred to by some as predictive analytics), with an inherent focus on mathematical probabilities, are increasingly prevalent but have not received adequate attention regarding their influence on clinicians’ behaviors.3, 4 To facilitate the study of these tools, an initial approach would be to study a clinical situation where clinicians’ prompt decisions and actions are warranted, and using currently available data to predict events likely to occur within 24-48 hours would be ideal. In-hospital cardiopulmonary arrest meets these criteria and served as an exemplar by which to study PB-CDS phenomena. Nurses spend more time with hospitalized patients than any other clinician and became the focus of this study. The overall objective of this study was to obtain nurses’ recommendations for the design of a PB-CDS tool. In this presentation, we report our findings on the information preferences of bedside nurses, charge nurses, and rapid response team nurses for the design of a tool assisting cardiopulmonary arrest identification.

Methods

We conducted three separate participatory design sessions, each with different participants. Participatory design is a qualitative method that engages participants as co-investigators in the design process. We recruited a purposive sample (attempting equal representation of bedside nurses, charge nurses, and rapid response team nurses) from an adult teaching hospital, a pediatric teaching hospital, and an adult federal hospital in a large urban city in the mid-South region of the U.S. Each 2-hour session comprised 5-10 end-users currently working as either bedside nurses, charge nurses, or rapid response team nurses. Facilitated by at least two of the researchers, each session contained a priming activity (~20 minutes in length), a designing activity (~60 minutes), and a debriefing activity (~30 minutes).

During the priming activity, all participants watched an 8-minute video vignette in which a patient experienced clinical deterioration warranting activation of a rapid response team. The designing activity engaged nurses in the hands-on creation of a physical representation of an artificial, electronic CDS tool using paper, colored pencils, scissors, rulers, and adhesive note paper. Designing occurred in a simulation laboratory with high and low fidelity manikins and several patient rooms. This setting re-created a clinical environment that mimicked the context in which the PB-CDS tool would realistically be used. During the debriefing activity, researchers used a semi-structured guide to ask participants to share their sketches and provide rationale for their chosen visual and functional elements.

The research team leveraged theme-based content analysis, ongoing aggregation of results, as well as discussion and deliberation of nurse end-user comments and artifacts. After a preliminary analysis was conducted, we consulted with human-computer interaction and design experts to provide an informal evaluation of the tool’s proposed visual and functional elements. We synthesized all recommendations, developed a low-fidelity prototype, and shared the prototype with 14 (70%) of the nurse end-users who participated in the sessions.

Results

Six bedside nurses, eight charge nurses, and six rapid response team nurses (n=20) attended the sessions from 14 unique units. Three major themes emerged from the designing and debriefing activities and represent participants’ goals for the CDS tool: (1) communication of patient status, (2) empowerment, and (3) consistency with context. First, participants reported they wanted a CDS tool that “paints a picture” or “tells the story” of the patient condition over time. They requested the ability for individual users to select which variables become visible and layer those variables’ trends for hypothesis generation and succinct communication. The concepts of advocacy and autonomy surfaced in the second goal. If a CDS tool is designed well, the tool could empower nurses to advocate for the patient and
contribute to treatment decision-making. As an objective assessment of the patient’s condition, the CDS tool has the potential to provide participants with a structured method by which nurses can garner support for their recommendations. In the third goal, nurses agreed that the model had to make sense, and the general perception was that probability-based models are more helpful for confirming what one already thinks rather than identifying unrecognized patient conditions.

Regarding specific design elements, participants frequently expressed a desire to visualize the temporal trend of predicted outcome probabilities with user-selected overlapping depictions of vital signs, laboratory values, and outcome-related treatments and interventions. Charge nurses and rapid response team nurses had a strong request for only viewing a ranked order of the highest risk patients at first; however, when viewing individual patient information, all nurse roles expressed similar preferences. Less notable but fairly commonly-heard requests included alerts only for values exceeding an absolute threshold or high degree of change, a green/yellow/red color scheme, and the ability to view the tool on both a mobile device as well as a dashboard. Participants who reviewed the prototype did not suggest changes. The Figures are screenshots of the prototype produced from these goals and recommendations. The top 2 figures are full-size screenshots; the bottom 4 figures are partial screens demonstrating interactive features.

Discussion and Conclusion
We used the participatory design method to identify important design elements and create a prototype for a PB-CDS tool that predicts cardiopulmonary arrests. Co-creation of the tool via participatory design was beneficial because the active involvement of multiple stakeholders facilitated the identification of novel, integrative design concepts that groups of participants (i.e., researchers and end-users) might not have identified separately. Designers and developers can use and extend the goals and associated recommended design elements to create PB-CDS tools for in-hospital nurses in the future. The information we gained about the preferred design elements of predictive analytics tools that support, rather than interrupt, nurses’ cognitive workflows can benefit future studies in this field as well as nurses’ practice. As these themes and elements undergo additional testing and refinement, we anticipate they can eventually serve as standards for developing PB-CDS tools that are more likely to influence clinician behavior and ultimately patient outcomes.

References
Building an FHIR Ontology based Data Access Framework with the OHDSI Data Repositories

Guoqian Jiang, MD, PhD¹, Guohui Xiao, PhD², Richard C. Kiefer¹, Eric Prod’hommeaux³, Harold R. Solbrig, MS¹

¹Mayo Clinic, Rochester, MN, USA; ²Free University of Bozen-Bolzano, Italy; ³MIT/W3C, Cambridge, MA, USA

Introduction

Clinical and translational research increasingly relies on the existence of robust integrated data repositories (IDRs) to combine clinical and “-omics” data. A variety of data models have been developed to provide standardized interfaces to organize research data in a clinical data repository. The Observational Health Data Sciences and Informatics (OHDSI) Common Data Model (CDM) has been increasingly used to build a large-scale international data network in support of observational studies [1]. Other data models include the i2b2 star schema, and the PCORNet CDM. These data models serve well as a layer of standardization for clinical research data within their own research network; however, if investigators want to reuse and integrate applications and accompanying research datasets across different research networks they still face huge challenges. This situation demands a global data model as a reference standard to facilitate data model harmonization and integration.

The HL7 Fast Healthcare Interoperability Resources (FHIR) is emerging as a next generation standards framework for facilitating health care and electronic health records (EHR)-based data exchange [2]. There is a critical need to build the FHIR-based data access and query on existing relational data sources to facilitate standards-based semantic data integration, sharing and discovery in broader scientific research communities [3]. The objective of the study is to develop an FHIR-based data access framework to enable answering semantic queries over the OHDSI CDM-based data repositories. We leverage the FHIR ontology and an open-source Ontology-based Data Access (OBDA) system known as Ontop [4] to demonstrate the feasibility of our approach.

Methods

Figure 1 shows the system architecture, comprising four layers: an input layer, a transformation layer, a semantic query layer and an application layer.

OHDSI CDM and Virtual Machine (VM): The OHDSI CDM 5.0.1 version (https://github.com/OHDSI/CommonDataModel) contains 39 database tables in 6 categories: standardized clinical data, standardized health system data, standardized health economics, standardized metadata, standardized vocabularies and standardized derived elements. We installed an OHDSI VM that is conformant to the OHDSI CDM 5.0.1 version. The VM contains the full OHDSI technology stack and is loaded with both standard vocabularies and sample data. The VM uses a relational database known as PostgreSQL as a storage backend. Ontop and its Protégé 5 Plugin: Ontop is an open-source OBDA system developed for querying relational data sources using an ontology-based approach. We installed the Protégé 5 Ontop Plugin (http://ontop.inf.unibz.it/), comprising 1) the Ontop Mappings tab – managing data source (i.e., database connection), and mapping creation through building an OBDA model (i.e., a specification of how the data in a data source is mapped to the vocabulary in an ontology); and 2) the Ontop SPARQL tab – providing a query editor to allow for editing SPARQL queries and executing a query to test an OBDA model. We at the Mayo Clinic have been collaborating with the FHIR and W3C HCLS community to develop the FHIR Resource Description Framework (RDF) representation specification and associated transformation and validation tools [2]. One of such efforts is to produce the FHIR StructureDefinition resource to OWL transformation, known as the "FHIR Ontology" (http://build.fhir.org/fhir.ttl). The StructureDefinition resource is the metamodel for FHIR resource definitions, meaning that a FHIR resource such as Patient, is formally defined using an instance of StructureDefinition that declares elements like “Patient.name” and “Patient.birthDate” and associated metadata and

Figure 1. System Architecture
constraints (e.g., datatype and cardinality). The FHIR Ontology formally enumerates the classes, predicates, domains, ranges and specific datatypes that are used in describing the FHIR instance data.

We first loaded the FHIR Ontology into a Protégé 5 environment, and used the Datasource manager in the Ontop Mappings tab and established the database connection to the OHDSI VM. And then we used the Mapping manager and Mapping assistant in the same tab and created an OBDA model declaring mappings between the FHIR Ontology and the OHDSI CDM database schema. The consensus of the mappings was achieved through a group discussion among co-authors. We also extended the FHIR Ontology with a number of datatype properties to capture the values of primitive datatypes. Once the datasource is connected and the mappings are defined, the Ontop Core APIs can be invoked and the SPARQL endpoints with reasoning capability (in terms of OWL 2 QL and RDFS) can be established. In this study, we used the Ontop SPARQL plugin that encapsulates both transformation and semantic query layers for the testing purpose. We created a collection of the SPARQL queries representing clinical research data questions, and used the queries to test the system.

Results and Discussion

We created the mappings for 7 OHDSI CDM tables: Person, Condition Occurrence, Drug Exposure, Observation, Concept, Concept Relationship and Concept Ancestor. Out of 7 tables, 5 tables are mapped to the FHIR core resources, i.e., the Person table to fhir:Patient (with 2 properties), the Condition Occurrence table to fhir:Condition (5 properties), the Drug Exposure table to fhir:MedicationAdministration (6 properties), the Observation table to fhir:Observation (7 properties). The rest of tables are from the Vocabulary CDM. The Concept table is mapped to a complex FHIR datatype Coding (3 properties), and the Concept Relationship table and the Concept Ancestor table are mapped to the FHIR resource ConceptMap (3 properties). A total of 418,602,850 virtual triples were produced through the mappings. We successfully created the SPARQL query templates that can identify patient cohorts using a single concept code or its descendants (i.e., demonstrating inference capability) from different domains (i.e., Condition, Medication Administration, and Observation). Table 1 shows the query results of three example concept codes in three domains. We verified that the results are accurate using an OHDSI open source cohort identification tool known as ATLAS (https://github.com/OHDSI/Atlas).

Table 1. Query results of three example concept codes in three different domains.

<table>
<thead>
<tr>
<th>Domain</th>
<th>Concept Code</th>
<th>Descendants</th>
<th>Number of Patients Retrieved with a Single Code</th>
<th>Number of Patients Retrieved with Descendants</th>
</tr>
</thead>
<tbody>
<tr>
<td>Condition</td>
<td>Acute myocardial infarction/57054005/SNOMED</td>
<td>57</td>
<td>6757</td>
<td>14453</td>
</tr>
<tr>
<td>Medication Administration</td>
<td>Warfarin/11289/RxNorm</td>
<td>489</td>
<td>0</td>
<td>25602</td>
</tr>
<tr>
<td>Observation</td>
<td>past history of procedure/416940007/SNOMED</td>
<td>481</td>
<td>4678</td>
<td>56772</td>
</tr>
</tbody>
</table>

Discussion

In this study, we developed an ontology-based data access framework for enabling FHIR-compliant semantic queries over the OHDSI CDM-based data repositories. We have demonstrated the feasibility of the framework by implementing a prototype leveraging the FHIR ontology and an open-source Ontop system. Our ongoing next steps include: 1) building a set of robust mappings between FHIR and OHDSI CDM through a community-based consensus approach; 2) enhancing the Ontop OBDA model to handle blank nodes which are required for the FHIR canonical RDF representation; 3) producing a library of SPARQL query templates for complex cohort identification logic; and 4) enabling federated queries across the OHDSI data repositories and beyond.

Acknowledgement

This study is supported in part by NIH grants U01 HG009450, U01 CA180940, and R01 GM105688.

References

1. OHDSI URL: https://www.ohdsi.org/; last visited at March 6, 2017.
Performance of Algorithms Using Healthcare Claims and Electronic Health Record Clinical Notes to Identify Relapsing-Remitting Multiple Sclerosis in U.S. Integrated Delivery Network Healthcare Database

Aaron Kamauu, MD, MS, MPH; John Holmén, PhD, MS; Christopher L. Fillmore, MD, MS; Hoa Van Le, MD, PhD; Andy Wilson PhD MSTAT; Chi Thi Le Truong, MD, PhD; Canter Martin, MBA; Schiffon L. Wong, MPH

1Anolinx, Salt Lake City, UT; 2Intermountain Healthcare, Salt Lake City, UT; 3PAREXEL, Durham, NC; 4MedCodeWorld, Mississauga, Ontario, Canada; 5EMD Serono, Billerica, MA

Abstract
This study was designed to develop and validate operational algorithms using healthcare claims and electronic health record (EHR) clinical notes for Relapsing-Remitting Multiple Sclerosis (RRMS) patient identification. Using data from an U.S. Integrated Delivery Network (IDN), algorithms using codes from healthcare claims and unstructured text from EHR clinical notes were constructed and positive predictive values (PPV) were calculated. Both the claims- and EHR clinical notes-based algorithms, with high PPV for identifying RRMS, are promising methods for future research.

Introduction
As the demand for observational studies using real-world data has grown, algorithms to improve case ascertainment have become increasingly critical to ensure the integrity and rigor of outcomes research as well as to help guide clinicians in disease subtype treatment and resource utilization decision-making. This study aimed to develop and validate operational algorithms using healthcare claims and EHR clinical notes for RRMS patient identification in an U.S. IDN healthcare system. These methods may be used to improve RRMS patient identification and to leverage the development and validation of algorithms for case ascertainment of other medical conditions.

Methods
IDN data from 2010-2014 were used to create a cohort of patients aged 18 or older who had at least 1 year of baseline data before a recorded MS diagnosis and no other demyelinating disease diagnoses to address any potential confounding. The claims-based algorithms used only structured/coded data, which were queried for the following combinations: presence of (1) MS ICD-9 diagnosis codes, MS-specific symptom codes recorded as a part of a neurology visit, physician prescribed MS disease modifying therapy (DMT) codes, and codes signifying brain/spinal MRI(s) performed; and (2) ruling out progressive MS (P-MS) through: (option A) medications more commonly used in MS patients with progressive disease; (option B) MS severity/progression from Kurtzke Functional Systems Scores (KFSS) adapted from ICD-9 coding; and (option C) P-MS based on combinations of supportive therapy and resource utilization. Our EHR clinical notes-based algorithm made use of clinical text documents and natural language processing (NLP), where the narrative text of clinical notes was searched electronically using key terms and phrases that potentially indicated MS subtype or progressive MS disease. This provided a comparison of information available via structured claims records vs. narrative text of clinical notes. Terms indicating RRMS subtype such as, ["multiple, sclerosis" or "ms"] and ["relapsing" and/or "remitting"] were used to categorize RRMS patients; terms indicating progressive MS diseases were used to exclude RRMS classification.

Random samples of at least 100 RRMS patients who were identified by each algorithm were manually validated by a single physician review of all available clinical notes, including outpatient clinic notes, inpatient progress reports, and brain/spinal MRI reports. Among the random samples, there were a total of 145 overlapping distinct patients with 1,713 reviewed clinical documents. Through this manual validation, these patients were assigned as Positive, Negative, or Unknown for RRMS subtype based on the information available in the clinical notes. Unknown was assigned to cases where the clinician’s explicit impression of MS subtype was not included in the notes and therefore could not be determined. Positive predictive values with 95% confidence interval (CI), excluding Unknown cases, were calculated for both algorithms. Sensitivity analyses were also conducted.
Results

Of 3,111 MS patients identified from the IDN, the claims-based algorithm identified 2,960, and the EHR clinical notes-based algorithm identified 990 as having positive evidence of RRMS subtype. Of the 2,960 possible RRMS patients identified using the claims-based algorithm, 2,271 (77%) were identified as not having evidence of progressive disease based on the P-MS rule out options A, B or C (98%, 79%, and 98%, respectively), and therefore became our final RRMS patient sample. The PPV (95% CI) for the claims-based algorithm compared to the clinical note-based manual validation was 0.89 (0.75-0.95) overall, and 0.87 (0.78-0.93), 0.89 (0.79-0.95) and 0.88 (0.79-0.93), respectively for the three P-MS rule out options. One third of patients identified using the claims-based algorithms were classified as Unknown cases, which were included in sensitivity analyses. The Unknown determination in validation review most often reflects the lack of explicit indication of MS subtype in the physician’s clinical notes.

Of the 990 possible RRMS patients identified using the NLP-based EHR clinical notes algorithm (i.e. MS patients who had a term for RRMS in their EHR), 837 (85%) were identified as not having evidence of progressive MS mentioned in a clinical note until 90 days after the index date. These patients were characterized as our final RRMS patients from the NLP-based EHR clinical notes algorithm. The PPV (95% CI) for the EHR clinical notes-based algorithm compared to the clinical note-based manual validation was 0.99 (0.94-1.00) when excluding Unknown cases and 0.96 (0.91-0.99) when Unknown cases were assumed negative for RRMS.

Conclusion

Both the claims- and EHR clinical notes-based algorithms had high PPV for identifying RRMS among patients with documented MS. Thus, they produced a cohort of RRMS patients that can be studied in clinical research with confidence. High PPV may be expected due to the high prevalence of RRMS among MS patients. We chose an approach using multiple criteria and several options to assess the contributions of each factor towards positively distinguishing RRMS patients from primary progressive MS (PPMS) and secondary progressive MS (SPMS) patients. Traditional medical chart reviews can support the clinical note-based manual validation, particularly for patients without clinical notes of MS subtypes. Our study assigned a high number of patients as Unknown during the manual validation because there was insufficient evidence of MS subtype documented in the clinical notes. Our validation required clear or explicit documentation by a clinician of the subtype. Often symptoms and course of care were well documented, but no clear or explicit indication of RRMS subtype was found. At times, this occurred when the precise MS subtype had not been clinically determined or when more time was required for the patient’s pattern of MS symptoms to present. Another scenario encountered was when patients were in a relatively stable clinical state with their MS, and the MS subtype may not have seemed relevant to the patient’s continued clinical care; therefore, it was not documented. This particular scenario helps explain why a much smaller number of RRMS patients were identified by EHR clinical notes only compared to the claims-based algorithm.

With the growing number of available disease modifying therapies indicated for relapsing forms of MS, better documentation of subtype will aid future clinical research as well as medical decision making for improved patient care and health outcomes. In the meantime, these claims- and EHR clinical notes-based algorithms to identify RRMS are promising methods for future MS research using real-world data.

References

A Comparison of Using Full and Partial Information from Administrative Claims Data to Predict Future Health Care Costs

Hong J. Kan, PhD, MPP, MA\(^1\), Hadi Kharrazi, MHI, MD, PhD\(^1\), Hsien-Yen Chang, PhD\(^1\), Jonathan Weiner, DrPh\(^1\)

\(^1\)Center for Population Health IT, Department of Health Policy and Management, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD

Introduction
In the last several decades, research has focused on modeling health care cost and utilization using statistical and econometric approaches, with much effort spent on explaining their variation and adjusting for underlying differences in risk for payment purposes\(^1\). Those models were typically developed with substantial domain knowledge (eg, a good understanding of the distribution of health care cost and utilization). Recently, an emerging literature started to apply machine learning methods to predicting health care cost and utilization\(^2\,3\). This study used lasso regularized regression to focus on important variables from a large pool of variables in the past year and longitudinal patterns of disease development in the past 4 years in predicting future health care costs under the scenarios of having complete claims data, simulated complete electronic medical records (EMR) data, and simulated incomplete EMR data. Since the linear regression model is still widely used as the de facto standard modeling framework for risk adjustment, predictors selected by lasso regression were also evaluated with linear regression to isolate the effects of variable selection for cost prediction.

Methods
This retrospective cohort study used IMS LifeLink\(^\circledast\) database\(^4\) comprised of commercial health plan claims in the United States. Patients born between 1949 and 1953 from a large health plan with 5 years of continuous enrollment from 2009 to 2013 were included for the study. Standard linear regression and lasso regression models were used to predict 2013 total cost with predictors from 2012 and clusters of longitudinal disease development patterns from 2009 to 2012. Many predictors were constructed using the Johns Hopkins ACG\(^\circledast\) system\(^2\) and used in the following scenarios (see Table 1 for a full list of variables in each scenario):

- Scenario 1: A user (such as a health plan) has complete claims data including cost, utilization, diagnoses, procedures, and medications over multiple years.
- Scenario 2: Simulating a user (such as an integrated health care delivery system) with complete EMR data including inpatient and outpatient care over multiple years. We used the same claims data but excluded cost.
- Scenario 3: Simulating a user (such as a small primary care practice) with incomplete EMR data. We assumed that the only complete data available was diagnoses, and thus only included diagnoses from claims data and grouped them into Expanded Diagnosis Clusters (EDCs, aggregate disease groups based on diagnosis codes) and Aggregated Diagnosis Groups (ADGs, an even higher level aggregation of diagnosed conditions).

Lasso regression was used to select variables and predict cost in each scenario. In addition, the variables selected by lasso plus age, sex squared and gender were included as predictors in linear regression models. Model performance was assessed using cross-validated \(R^2\) and cross-validated root mean square error (RMSE). We also tested the importance of longitudinal disease development clusters based on yearly EDC counts and yearly major ADG counts from 2009 to 2012 formed by k-means clustering.

Results
A total of 64,258 patients were identified with mean (SD) age of 60.9 (1.4) years old and 49.9% being female. Mean (SD) total allowed cost in 2013 was $10,380 ($30,232). Twenty clusters of yearly major ADG counts and 21 clusters of yearly EDC counts from 2009 to 2012 were identified. Using linear regression models, we found: (1) Cross-validated \(R^2\) with lasso-selected predictors was only slightly lower than that with the full set of predictors in scenarios 2 and 3, and was even slightly higher in scenario 1; (2) Cross-validated RMSE were essentially the same between the full and reduced sets; (3) Some of the longitudinal major ADG count clusters and EDC count clusters are important predictors in scenario 3. Relative to linear regression, lasso regression performed better with slightly higher \(R^2\) and lower RMSE in all scenarios (Table 1).
Table 1: Cross-validated R² and RMSE of linear and lasso regression models with full and reduced sets of predictors

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Predictor Set</th>
<th>Full set of predictors 2012</th>
<th>Lasso selected predictors in 2012</th>
<th>Lasso selected predictors: adding 21 EDC clusters</th>
<th>Lasso selected predictors: adding 20 ADG clusters</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>cv R²</td>
<td>cv RMSE</td>
<td>cv R²</td>
<td>cv RMSE</td>
</tr>
<tr>
<td>Scenario 1: Complete claims data including cost</td>
<td>OLS</td>
<td>22.8%^a</td>
<td>26,474^a</td>
<td>23.1%^b</td>
<td>26,413^b</td>
</tr>
<tr>
<td>Scenario 2: Simulating complete EMR data with no cost</td>
<td>OLS</td>
<td>17.8%^c</td>
<td>27,354^c</td>
<td>17.4%^d</td>
<td>27,308^d</td>
</tr>
<tr>
<td>Scenario 3: Simulating incomplete EMR data with EDCs and ADGs</td>
<td>OLS</td>
<td>14.1%^*</td>
<td>27,958^*</td>
<td>13.8%^e</td>
<td>27,995^e</td>
</tr>
<tr>
<td></td>
<td>Lasso</td>
<td>-</td>
<td>-</td>
<td>14.3%</td>
<td>27,990</td>
</tr>
</tbody>
</table>

a. 408 predictors: Age, age squared, female; allowed cost: inpatient, outpatient and pharmacy; total number of claims: inpatient, outpatient and pharmacy; 34 ADG indicators; frailty, generalist, coordination_LCI, coordination_PCI, resource utilization band 1-5, hospital dominant morbidity, chronic condition count, major ADG count, active ingredient count, majority of source of care, unique provider count, specialty count, management visit count, 280 EDC indicators, 67 RxMG indicators. b. 11 predictors: Age, age squared, female; allowed cost: inpatient, outpatient and pharmacy; total number of claims: inpatient, outpatient and pharmacy; major ADG count, REN06. c. 405 predictors: same as except for no allowed cost. d. 11 predictors: Age, age squared, female; total number of claims: inpatient, outpatient and pharmacy; hospital dominant morbidity, major ADG count, REN06, MAL03, and rxmgGURx020. e. 317 predictors: Age, age squared, female, 280 EDC indicators, 34 ADG indicators. f. 35 predictors: Age, age squared, female, ADG3, ADG4, ADG7, ADG9, ADG10, ADG11, ADG16, ADG22, ADG26, ADG27, ADG28, ADG32, ADM02, TOX04, MUS01, CAR05, NUR21, REN02, CAR16, HEM02, REN01, REN06, RES01, ALL06, HEM08, RES12, MAL12, MAL03, NUR12, REN03, NUR08, NUR10, REN12, REN03, MAL03, MAL12, MAL03, NUR12, REN03, NUR08, MAL10, cluster0. g. 34 predictors: Age, age squared, female, ADG3, ADG4, ADG7, ADG9, ADG10, ADG11, ADG16, ADG22, ADG27, ADG28, ADG32, ADM02, TOX04, MUS01, CAR05, REN02, CAR16, HEM02, REN01, REN06, RES01, ALL06, HEM08, RES12, MAL12, MAL03, NUR12, REN03, NUR03, MAL08, MAL10, cluster3, cluster4, cluster5, cluster7.

Conclusion
Using high-dimensional data to develop a linear prediction model may risk overfitting and inclusion of irrelevant variables. This study demonstrated that reducing the number of input variables in a linear regression through a lasso-based variable selection procedure preserved most of the prediction power, potentially meaningful to users with limited EMR data. Longitudinal disease development patterns based on EDCs and major ADGs were important when only disease diagnoses were available. However, their contribution to model performance was minimal. In addition, lasso regression showed slightly better performance in terms of R² and RMSE relative to linear regression. Although complete cost and utilization data from claims data maximized prediction of future costs, limited EMR data with diagnosis information only could still explain ~14% of variation of future cost in a liner regression model.

References
4. IMS LifeLink® database, QuintilesIMS, Watertown, MA
Agile Clinical Decision Support Development

Vaishnavi Kannan, MS, Mujeeb Basit, MD, MMSc, DuWayne Willett, MD, MS
University of Texas Southwestern Health System, Dallas, Texas

Abstract

Designing effective Clinical Decision Support (CDS) tools in an Electronic Health Record (EHR) can prove challenging, due to complex real-world scenarios and newly-discovered requirements. As such, deploying new CDS EHR tools shares much in common with new product development, where “agile” principles and practices consistently prove more effective than traditional project management. Typical agile principles and practices can thus prove helpful on CDS projects, including time-boxed “sprints” and lightweight requirements gathering with User Stories and acceptance criteria. Modeling CDS behavior removes ambiguity and promotes shared understanding of desired behavior, but risks analysis paralysis: an Agile Modeling approach can foster effective rapid-cycle CDS design and optimization. The agile practice of automated testing for test-driven design and regression testing can be applied to CDS development in EHRs using open-source tools. Ongoing monitoring of CDS behavior once released to production can identify anomalies and prompt rapid-cycle redesign to further enhance CDS effectiveness.

Introduction

Despite years of seeking portable clinical decision support (CDS), most CDS development remains local. The concept of shareable CDS tools—designed, constructed, and tested once, then re-used at other sites—holds intuitive appeal. Adopting what's proven, avoiding re-inventing the wheel—the concept is compelling. But now that most healthcare organizations are on an EHR, what is actually happening "on the ground" with CDS? Today, most CDS tools (alerts, rules, banners, etc.) still are developed locally within a single institution. While successful novel designs may be presented, the volume of actual record-sharing appears low. Cycle times for delivery of a new CDS tool are generally felt to be long. Defects in CDS tools delivered to production are perceived to be common. How then can we deliver to our own production environments CDS tools of higher quality, faster, and with less wasted effort? Much foundational groundwork has been laid and progress made in CDS logic design, user interface design, workflow context evaluation, and evaluation: we deal here with methods for implementing CDS in a production EHR, employing these established concepts. New CDS development shares characteristics with new product development, including creating something that did not exist before, and a complex environment where solutions may not be initially obvious but emerge from frequent feedback. In new product development, agile principles and practices have proven superior. Our core tenet is that certain of these approaches (agile project management, agile modeling, test-driven development, and rapid-cycle monitoring) will improve the effectiveness of clinical decision support development.

Methods

From the Agile Project Management (APM) domain, we employed Scrum (iterative development), "user stories" for lightweight requirements gathering, and a use-case driven development approach. Agile modeling methods included principles such as "use the simplest tools possible", "model together", and employed a combination of standard, rigorously-defined diagram types, both Unified Modeling Language (UML) and non-UML diagrams. Automated acceptance testing was accomplished with FitNesse, an open-source wiki-based testing framework, with the dbFit extension allowing queries against the EMR database. Monitoring to-date has been semi-automated, with construction of data cubes (Microsoft Analysis Services) and interactive dashboards and reports (Excel, PowerBI, Microsoft) for multidimensional analysis of CDS behavior in production, looking for anomalies and opportunities for iterative improvement.

Results

Agile Project Management: User stories were employed in the typical format of "As a <role>, I want <some CDS tool>, so that <value to be achieved>. We found writing these in the voice of the user who would be exposed to the CDS tool—not the requestor—was instructive. This helped clarify the who, what, and why of the CDS "5 rights" before beginning detailed design. The use of time-boxed 2-week iterations with demonstration of working CDS product at end-iteration proved feasible, and provided the opportunity for early feedback and iterative refinement.
Agile Modeling: A subset of modeling techniques and diagrams proved most consistently useful:

- Use case diagrams to depict all the user activities in a system and actors/roles involved, including which roles would use which CDS tools.
- Swimlane workflow (UML Activity) diagrams, to show a defined sequence of activities across multiple roles.
- Decision trees, to unambiguously describe the logic for when an alert should and should not appear.
- User interface storyboarding and Use Case Text (the latter in the form of “user does this”, “system does that”): to visualize how the user would interact with the CDS user interface, and how exactly the system should respond to any given user action.
- Object diagrams: to depict both the business objects and their relationships in the "problem domain", and the EHR objects and their relationships in the "solution domain".

Test-Driven Development and Monitoring: Employing the open-source testing framework Fitnesse with an add-in (dbFit) that allows writing and running SQL database queries enabled writing of acceptance tests to specify the structure and behavior of a CDS rule/alert, even before it was built. The test would be "red" or fail until the CDS alert was constructed, then the test would turn green if the build matched the original specifications. Suites of these tests were also developed.

A multidimensional data model for user interaction with CDS alerts was extended and provided both time-series analyses and snapshot comparison graphs of subgroups to evaluate CDS behavior, in terms of both triggering events and user responses.

Conclusion

Typical agile principles and practices can prove helpful on CDS projects, including time-boxed “sprints” and lightweight requirements gathering with User Stories and acceptance criteria. Modeling CDS behavior removes ambiguity and promotes shared understanding of desired behavior, but risks analysis paralysis: an Agile Modeling approach can foster effective rapid-cycle CDS design and optimization. Modeling using standard rigorous model types (UML and non-UML) provides one mechanism for sharing designs between organizations for faster replicability, and potentially to evolution of re-usable design patterns. The agile practice of automated testing for test-driven design and regression testing can be applied to CDS development in EHRs using open-source tools. Ongoing monitoring of CDS behavior once released to production can identify anomalies and prompt rapid-cycle redesign to further enhance CDS effectiveness.

References

Analyzing Prevalence of Obesity among VHA Patients Using Clinical, Temporal, and Geographical Data Extracted from a Nationwide EHR

Hadi Kharrazi1, Tamara Box2, Stephen Fihn2, Jonathan Weiner1
1 Johns Hopkins University, Baltimore, MD, USA
2 U.S. Veterans Health Administration, Washington DC, USA

Abstract

This study evaluates the use of EHR data to measure the prevalence of obesity among 9,389,909 male veterans seeking care at one of the VHA facilities in the continental US between 2000 and 2015. Data is extracted from VHA’s national Clinical Data Warehouse representing VHA’s nationwide EHR. Obesity prevalence using EHR-extracted BMI, data quality issues, and various geographical and temporal analyses of data are discussed.

Introduction

Obesity is one of the most prevalent adverse health outcomes in the U.S. population (1). Obesity rates among veterans seeking care at the VHA is higher than the general population (2). Understanding obesity trends in the VHA population is important to prevent obesity and obesity related diseases and to control future costs (3). Similarly, geographic areas of excess obesity will be essential to target interventions and services to address obesity among the VHA population.

Methods and Study Design

The VHA’s Corporate Data Warehouse (CDW), a data repository of VHA’s VistA EHR that archives clinical records of 21+ million veterans daily (4), was utilized to identify and track the BMI (body mass index) level of individual veterans over an extended timespan (2000 to 2015) and on a national scale (50 US states; excluding territories). BMI levels were adjusted and/or stratified based on age (18-30, 30-40, 40-50, 50-60, 60-70, 70-80, 80+) and stratified based on gender. Patient’s addresses were translated to latitudes and longitudes and mapped with Census shape files (5). Geographical aggregate BMIs (e.g., county level BMIs) were calculated after adjusting/controlling for age and stratifying for gender using multi-level modelling approach.

The initial denominator included 21,470,625 patients. Patients with missing age/sex, invalid veteran status, or a death record before 2000 were excluded (4,392,183). In addition, patients who did not have at least one weight or height were excluded (6,712,870) as well as veterans without a US address (252,645). An analysis of the excluded patients did not reveal a significant difference in demographics, comorbidities, and/or geographical distribution from patients included in the study (10,112,927). Only male patients (9,389,909) are reported in this abstract.

Findings

Average BMI of male veteran patients has increasing across all age-groups from 2000 to 2015 (after adjusting for age). Smallest BMI increase from 2000 to 2015 has been for the 18 to 30 age group with .87 points. In the same timeframe, the largest BMI increase has been for the 40 to 50 age group with 2.45 points (Table 1). Other age groups have had a consistent increase in BMI ranging from 1.18 points to 1.65 (Table 1).

Table 1. BMI (average, SD) of male veteran population seeking care at VHA facilities between 2000 and 2015

<table>
<thead>
<tr>
<th></th>
<th>18-30</th>
<th>30-40</th>
<th>40-50</th>
<th>50-60</th>
<th>60-70</th>
<th>70-80</th>
<th>80+</th>
<th>patients</th>
</tr>
</thead>
<tbody>
<tr>
<td>2000</td>
<td>28.33 ± 5.07</td>
<td>28.66 ± 5.70</td>
<td>27.92 ± 6.10</td>
<td>28.13 ± 6.11</td>
<td>28.34 ± 5.54</td>
<td>28.22 ± 4.86</td>
<td>28.05 ± 4.35</td>
<td>1,980,484</td>
</tr>
<tr>
<td>2010</td>
<td>29.00 ± 5.17</td>
<td>29.76 ± 5.67</td>
<td>29.81 ± 6.13</td>
<td>29.83 ± 6.45</td>
<td>29.44 ± 6.12</td>
<td>29.47 ± 5.30</td>
<td>29.30 ± 4.51</td>
<td>4,343,447</td>
</tr>
<tr>
<td>2015</td>
<td>29.20 ± 5.45</td>
<td>29.84 ± 5.79</td>
<td>30.37 ± 6.17</td>
<td>29.49 ± 6.43</td>
<td>29.63 ± 6.22</td>
<td>29.83 ± 5.63</td>
<td>29.69 ± 4.64</td>
<td>4,584,068</td>
</tr>
</tbody>
</table>

BMI has increased in some geographical boundaries more than other (Figure 1). Certain counties showed a higher increase in adjusted BMI for the male patient population of VHA over time than others.
Figure 1. Geographical distribution of average BMI of male VHA patient population

Geographical unit of analysis: Census defined county; Years: 2000 [top-left] to 2015 [bottom-right];
BMI scale/ranges (same for all maps): 27.5 [green] to 32.5 [red]

Conclusion

EHR data can be utilized to track BMI levels on a temporal and spatial basis for population-health management purposes at the VHA. VHA uses considerable resources to tackle obesity and its consequences. Findings of this study can guide policy makers and local clinical program managers to allocate and guide existing resources, such as VHA’s national MOVE weight management program (6), to areas of excess obesity.

Learning Objectives:
1. Explain the use of EHR data (e.g., data quality) in analyzing geographically-bound population health outcomes
2. Compare the obesity rates in the VHA’s male population with the general population

References

Consumer Views of Electronic Health and Genetic Data Sharing: Findings of a National Survey

Katherine K. Kim, PhD, MPH, MBA1, Lucila Ohno-Machado, PhD, MD, MBA2
1University of California Davis, Sacramento, CA; 2University of California San Diego, La Jolla, CA

Abstract
Networks have emerged to provide infrastructure including technology and data access to support person-centered care and precision health. Consumers are willing to electronically share genetic and health data but factors such as trust and purpose may moderate their willingness. This national, random digit-dial survey (n=1504) showed greater trust in healthcare and educational organizations accessing health data than national health agencies or companies. Respondents are more willing to share health data for research than genetic data.

Introduction
Networked health systems have emerged to provide infrastructure including network technology, access to data, and governance mechanisms to support national priorities of enabling person-centered care and delivering precision health. Several national programs have recently been initiated to spur innovation in these areas. The Patient-Centered Outcomes Research Institute (PCORI) has made a major investment in distributed research networks (DRNs) to enable efficient, large-scale person-centered clinical research that leverages secondary use of clinical data.1,2 The All of Us Precision Medicine Research Program aims to build a cohort of 1 million Americans with collection of genetic and longitudinal clinical data from EHRs.3 Recent studies show that consumers are willing to participate in research, contribute genetic data4 and electronic health data.5 However, this willingness may be moderated by factors such as trust, the purpose for data access, and type of data accessed. The purpose of this study was to explore these factors related to sharing of electronic health data and genetic data.

Methods
A survey was developed that included new questions and existing questions from the literature. New questions were pretested for clarity with six individuals prior to fielding the survey. A random-digit dial telephone survey was conducted using a national sample of phone numbers, 68% landline and 32% mobile representing the national distribution.6 The survey was conducted in English and Spanish by bilingual interviewers and lasted an average of 20 min. Organizational trust was assessed with the question: I would be comfortable if [organization] had electronic access to my anonymous health information for research from1=strongly agree to 5=strongly disagree. Willingness to share electronic data was assessed via the following survey questions all rated from 1=very likely to 4=very unlikely. 1) If health researchers asked to use your anonymous health information and to send it electronically to other researchers, how likely or unlikely would you be to agree to that?; 2) If health researchers asked to use your genetic information and to send it electronically to other researchers, how likely or unlikely would you be to agree to that?; and 3) If you were offered the choice to have your health information automatically shared electronically (without requiring any action by you) with the different places where you receive health care, how likely would you be to agree to it?

Results
The survey concluded with 1504 responses, representing a 9.95% response rate. 28 interviews were conducted in Spanish. The mean age of respondents was 53.5 years (SD = 17.7, range 18-99). Respondents were representative of the US population in terms of gender, but the sample is less diverse (78.0% white, non-Hispanic/Latino vs. US proportion of 63.0%) and more highly educated (34.5% college educated vs. US proportion of 28.2%).7 Respondents’ trust differs by organization type although the differences are small. They were most comfortable with healthcare and educational organizations accessing their health information: hospital (M = 2.34, SD = 1.17), doctor’s office (M = 2.35, SD = 1.19), college/university (M = 2.89, SD = 1.31), and state/local government health agency (M = 2.97, SD = 1.37). They were less comfortable with national agencies and companies: national government health agency (M = 3.16, SD = 1.45), pharmaceutical company (M = 3.16, SD = 1.36), biotechnology company (M = 3.22, SD = 1.31), and insurance company (M = 3.40, SD = 1.33).
There were significant differences in willingness to share health and genetic data (paired sample t-tests, p < .001). Two-thirds of respondents were likely to share health data for research, while less than half were likely to share health data for healthcare or genetic data for research.

![Electronic Data Sharing*](image)

**Figure 1.** Willingness to share data electronically by type of data and purpose

**Discussion and Conclusion**

Consumers have different concerns and expectations for use of genetic data vs. health data, and for the purpose of healthcare vs. research. Since genetic data and health data used in delivery healthcare rely on identified data, the public may be more concerned about disclosure and lack of privacy. A recent study on consent for data sharing in California found that patients are more likely to share de-identified health information for research purposes than to share identified information for healthcare purposes. Person-centered care and precision health both rely on these data. Programs such as PCORI and *All of Us* must pay close attention to the views of consumers if they are to build the public trust necessary for success.

This study was limited by a low response rate which raises concerns about potential non-response bias. However, the response rate was in line with the declining rates seen in other national surveys which currently averages 9%. Respondents were also not fully representative of the US population.

Networks are emerging to fill the needs of large-scale data collection and secure data access for healthcare and research. Understanding of consumers' expectations can inform the construction of national and network-specific policies that support the laudable aims and protect public trust.

**Acknowledgements:** The study was funded by AHRQ 1R01HS019913-01.

**References**

Examining the Coverage of Nursing Content in the UMLS-CORE Problem List

Tae Youn Kim, PhD, RN¹, Nicholas Hardiker, PhD, RN, FACMI²
¹University of California Davis, Sacramento, CA, USA
²University of Salford, Salford, UK

Introduction
The meaningful use of electronic health records (EHR) requires maintenance of an up-to-date patient problem list which is coded in a principled way.¹ This problem list should include clinical diagnoses and patient problems (or findings and complaints) that are relevant to the current care of an individual regardless of specialties and disciplines.² Nursing problems identified by nurses through their clinical judgment should also be integrated into a patient problem list to ensure continuity of care and drive patient-centered treatment plans.³,⁴ The standardization of coded problem lists has many potential benefits, such as support for clinicians’ decision-making⁴ through patient-centered alerts or reminders that could be triggered for disease management, follow-up care, health promotion, and/or early detection and prevention of an adverse outcome in the EHR. Yet maintaining a standardized list of patient problems that are sharable within and across settings is challenging for a range of reasons, including variations in clinical work flow, documentation practice, and administrative and technical support at the organizational level. As part of an effort to promote the implementation of a standardized problem list, the U.S. National Library of Medicine (NLM) and the International Health Terminology Standards Development Organisation (IHTSDO) have released Systematized Nomenclature of Medicine—Clinical Terms (SNOMED CT) Subsets that cover selected patient problems.²,⁴,⁵ The purpose of this study was to examine the extent to which the Clinical Observations Recordings and Encoding (CORE) Problem List² can serve as a multi-disciplinary problem list, especially in coding patient problems addressed by the nursing profession. Specifically this study aimed to create a primary nursing problem subset coded with SNOMED CT and determine the coverage of nursing problems in the CORE.

Methods
This study involved a 4-step process from data acquisition to analysis. First, two existing sets of nursing problems were obtained from the NLM and IHTSDO.⁶,⁷ The NLM Nursing Problem List (201411 Version) is a set of nursing diagnoses created using the Unified Medical Language System (UMLS) Metathesaurus⁴,⁶ while the IHTSDO Nursing Reference Set (201607 release) refers to a product developed through a collaboration agreement between the International Council of Nurses and IHTSDO.⁷ Second, a dataset was then created by merging these two sets of nursing problems, which is named as the Nursing Core Problem List (or Nursing CPL) in this study. Third, the content of the Nursing CPL was compared with that of the CORE Problem List including 6,159 SNOMED CT concepts derived from the UMLS 2016AB.⁸ In other words, the coverage of nursing problems in the CORE Problem List was determined by examining the proportion of nursing content in the UMLS-CORE. Lastly, all SNOMED CT codes were verified against the SNOMED CT 20160731 release and then linked to concept unique identifiers in UMLS 2016AB for identifying semantic types or high-level categories in the semantic network.⁹,¹⁰

Results
The NLM Nursing Problem List (201411 Version) and the IHTSDO Nursing Reference Set (201607 release) comprised 506 and 492 SNOMED CT-encoded patient problems respectively. There was overlapping content between the two nursing subsets (i.e., 261 concepts were identical). After removing duplicates, a total of 723 unique problems made up the experimental Nursing CPL for this study. According to the SNOMED CT hierarchy, all but six problems were located in the Clinical Finding hierarchy. The outliers were classified as either Situation (e.g., family grieving) or Event (e.g., elderly person maltreatment) concepts in the hierarchy. When examining further the semantic types of the 723 nursing problems in UMLS, the majority were classified as Finding (³ 80%) and Biologic Function (³ 16%). The semantic type ‘Finding’ includes problems discovered by direct observations or measurements, such as signs and symptoms while ‘Biologic Function’ consists of concepts associated with mental or behavioral dysfunction, disease or syndrome, and mental process.⁹,¹⁰ When querying the 723 nursing problems against the CORE Problem List, approximately half of the nursing problems (343 problems or 47.4%) were found, accounting for 5.6% of the CORE concepts. Table 1 presents a summary of example concepts appearing in the Nursing CPL and CORE subsets.

Discussion and Conclusion
Nursing problems appear across a range of terminologies. The NLM Nursing Problem List has been developed through the identification of nursing problems that exist within the U.S. healthcare environment.⁴ In contrast, the IHTSDO Nursing Reference Set was developed through an exploration of nursing practice worldwide.⁵,⁷ In an effort
to create a primary set of nursing problems, this study combined the two nursing subsets into a Nursing CPL, yielding a total of 732 concepts relevant to the nursing profession. Although this list is not exhaustive, it provides a useful starting point and can serve as a foundation for future research. With regard to semantic locations of nursing content in the SNOMED CT and UMLS Metathesaurus, the discrepancy was somewhat expected as each organization employs a different modeling and integration process of clinical concepts. While the study findings indicate that the nursing CPL is only partially represented in the UMLS-CORE Problem List, this study illuminates potential for research to enhance this nursing CPL and the CORE. Interestingly, the UMLS-CORE Problem List contains additional concepts that could be seen as having relevance to nursing, as shown in Table 1. Further examination is needed to improve the semantic locality of nursing content in UMLS and explore the applicability of this nursing CPL to clinical practice in order to promote data collection, retrieval and exchange across healthcare settings where various EHR systems have been implemented along with different standards of care. While just over half of the nursing problems within the nursing CPL were not found in the CORE, with some enhancement to nursing content, the CORE has the potential to serve as a multi-disciplinary problem list.

Table 1. Example Concepts Appearing in the Nursing Core Problem List (n=723) and UMLS-CORE (n=6,169)

<table>
<thead>
<tr>
<th>SNOMED CT Fully Specified Name (20160731 Release)</th>
<th>Semantic Type (UMLS 2016AB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>162116003 Increased frequency of urination (finding)</td>
<td>A2.2 Finding</td>
</tr>
<tr>
<td>7058009 Noncompliance with treatment (finding)</td>
<td>A2.2 Finding</td>
</tr>
<tr>
<td>404640003 Dizziness (finding)</td>
<td>A2.2.2 Sign or Symptom</td>
</tr>
<tr>
<td>3253007 Discoloration of skin (finding)</td>
<td>B2.2.1.2 Pathologic Function</td>
</tr>
<tr>
<td>80394007 Hyperglycemia (disorder)</td>
<td>B2.2.1.2.1 Disease or Syndrome</td>
</tr>
<tr>
<td>15167005 Alcohol abuse (disorder)</td>
<td>B2.2.1.2.1.1 Mental or Behavioral Dysfunction</td>
</tr>
</tbody>
</table>

A. Problems appearing in both Nursing CPL and CORE

B. Problems appearing in the Nursing CPL only

C. Nursing-relevant problems appearing in the CORE only

Acknowledgements
The authors would like to thank the International Council of Nurses and SNOMED International for their support to the completion of this work.

References
Reuse of PCORnet Data to Support the Precision Medicine Initiative: Data Model Harmonization

Jeffrey G. Klann, PhD1,2,3; Matthew A. H. Joss2; Kevin Embree2; Shawn N. Murphy, MD, PhD1,2,3

1Harvard Medical School, Boston, MA; 2Partners Healthcare, Boston, MA; 3Massachusetts General Hospital, Boston, MA

Introduction
The proliferation of research repository systems, which are increasingly used for secondary analysis of electronic health record data, has brought data interoperability to the forefront of medical informatics. Several contrasting data models for clinical research are emerging, such as the PCORNet Common Data Model (CDM), Informatics for Integrating Biology in the Bedside (i2b2), and the Observational Medical Outcomes Partnership (OMOP) CDM. To maximize participation in clinical research initiatives, it is increasingly important for systems to be compatible with multiple data models.

The Precision Medicine Initiative, now known as the All Of Us Research Program, presents a current practical data interoperability challenge. AllOfUs is an NIH-sponsored, “participant-engaged, data-driven enterprise supporting research at the intersection of lifestyle, environment, and genetics to produce new knowledge with the goal of developing more effective ways to prolong health and treat disease.”[1] AllOfUs will be collecting electronic health record data from consented participants using the OMOP data model, and AllOfUs is presently requiring its partner organizations to implement OMOP data export. Our partnership, the New England Precision Medicine Consortium, comprises two sites which have already invested significantly in the PCORNet Common Data Model through the Scalable Collaborative Infrastructure for a Learning Health System (SCILHS) network: Boston Medical Center and Partners Healthcare. Therefore, we explored whether some of their investment could be reused to prepare data for AllOfUs.

We previously published a methodology that allows i2b2 data to be transformed into a PCORNet conformant data model using "information models", which are modeled as i2b2 “ontologies,” or hierarchies of terminology. [2] By developing a PCORNet information model in i2b2 and implementing a robust methodology for mapping data into it, we technologically enabled the transformation of i2b2 data into the PCORNet CDM. This was practically important because we presently run a PCORNet network that internally uses i2b2 but externally must support PCORNet's initiatives. We found that the PCORNet CDM supported a comprehensive information model for network interoperability.

We hypothesized that the similarities between the PCORNet common data model and the OMOP common data model would allow significant reuse of our existing investment. Partners Healthcare has data for 3.5 million patient lives mapped to the PCORNet information model and Boston Medical Center (BMC) has 1.5 million. Together, they could potentially bring 5 million patient lives to the OMOP CDM.

Here we describe our unique solution, which utilizes the PCORNet information model in i2b2 to transform directly into OMOP. This has also given us a detailed understanding of the differences and similarities in these data models.

Methods
Our existing methodology consists of two components: an information model which exactly mimics the fields and permissible values in the PCORNet CDM; and, a SQL program which transforms data using the mappings in the information model into the PCORNet CDM. We sought to modify only the latter component, so that sites with their data mapped to the PCORNet data model could produce an OMOP data set without any additional work.

We did ultimately make one change to our information model, but one that does not require remapping. OMOP version 5 assigns its own unique "OMOP numbers" to every standard term that is in its concept dictionary. Therefore, we added an OMOP code column to our information model that had a direct cross-reference to the OMOP number in the concept dictionary.

AllOfUs chose the tables in OMOP that the leadership believed were the most clinically relevant domains and excluded the fields that were deemed to be too obscure from the data model. The domains included: Person, Visit, Condition, Procedure, Measurement, and Drug. Our task was to provide data in OMOP format covering as many of the remaining fields as possible. To accomplish this, we analyzed differences in: a) the data model and b) terminology design between PCORNet and OMOP. We then implemented changes to the SQL transform to transform i2b2 data mapped to our PCORnet ontology into OMOP format.

Results
Data model analysis
Our analysis found five categories of field relationships between PCORnet and OMOP:

- **The field exists in OMOP and PCORnet** but with different names, and the field does not use a structured terminology. These cases were trivial renamings of the field names in the transformation. This included mostly dates and pseudo-identifiers. \((52/89 = 58\%)\)

- **The field does not exist in PCORnet.** (ex. drug lot #, procedure quantity) These were impossible to transform without modifying the information model. None of the fields in which this occurred were available in our source data, so at our site, this did not affect the output. \((14/89 = 16\%)\)

- **The field does not exist in OMOP.** (ex lab result priority, data source type) These could not be transformed into OMOP, which resulted in an implicit information loss in the transformed data. \((na/denominator\%\) is OMOP; comparison is PCORnet)

- **OMOP and PCORnet represent the same information in different ways.** (ex. drug type, normal ranges) Though PCORnet valuesets often mapped directly into OMOP, sometimes incompatibilities arose, yielding mapping with information loss. For example, PCORnet distinguishes between prescribed, dispensed, and administered drugs, but not subtler variations supported by OMOP, such as “dispensed through mail order” or “Physician administered drug, identified from referral record.” \((7/89 = 8\%)\)

- **The field exists in both OMOP and PCORnet, but with different structured terminologies.** (ex. condition & procedure codes) The subtleties of this process were the most challenging element of creating the new transform. OMOP’s concept dictionary defines terms’ domain in a nonstandard way. There are thousands of cases in which, e.g., a procedure code is assigned to the diagnosis table and vice-versa. Furthermore, OMOP’s mappings to their preferred terminologies split single source terms into multiple target terms, sometimes spanning multiple destination tables. Finally, although OMOP’s mappings do include retired codes, many codes in our ontology were missing from one or both mappings. For this version, we chose to support only the straightforward use case: the term is mapped and the target terms all fall into the same table as they appear in the PCORnet ontology. \((16/89 = 18\%)\)

(Percentages are % of fields in the OMOP data model selected for AllOfUs.)

### Pilot Transformation

We successfully transformed a subset of our PCORnet datamart at Partners Healthcare into OMOP format and ran it against a suite of validations provided to us by AllOfUs. These validations are based on standard OMOP analytical tools, Achilles and Achilles Heel. The validation tools are available for download at [https://github.com/cumc-dbmi/pmi_sprint_reporter](https://github.com/cumc-dbmi/pmi_sprint_reporter). Both a set of test data and our production data on real patients passed all tests.

### Discussion

With relative ease, we leveraged our previous work in transforming i2b2 data into PCORNet CDM. When the source data is i2b2, we now have a methodology and a common ontology to transform this data into two popular analytical data models, PCORNet and OMOP CDM.

We do not at present support terms that map to a different domain than their source domain, terms that bifurcate into terms in multiple domains, and several optional columns. However, the current transform is sufficient for initial data analysis in the AllOfUs research project. Our tool is open source and is being made available in the SCILHS GitHub repository as components are finalized. [3]

Our analysis of these two data models revealed that, among their common domains, there are far more similarities than differences. The most significant differences were in terminology; most of the data fields were represented similarly between the two models. Terminological differences are, in our experience, the most daunting and challenging aspect of data transformations in medical informatics, which has continued to prove true in this endeavor despite the similarity of preferred terminologies between PCORNet and OMOP. A relatively small number of OMOP fields are not present in PCORNet, and designing a harmonized PCORNet/OMOP ontology (terminology differences aside) would be relatively straightforward.

Our work will allow for considerably easier participation in a variety of initiatives even if they use different data models for their analytics.

### References

Evaluation of Text Mining Methods to Support Reporting Public Health Notifiable Diseases Using Real-World Clinical Data

Matthias Kochmann, MD1,2, Jiachen Wang, MS1, Brian E. Dixon, MPA, PhD1,3, Suranga N. Kasthurirathne, BEng4, Shaun J. Grannis MD, MS1,2
1Regenstrief Institute, Indianapolis, IN, USA; 2Indiana University School of Medicine, Indianapolis, IN, USA; 3Indiana University Richard M. Fairbanks School of Public Health, Indianapolis, IN, USA; 4Indiana University School of Informatics and Computing, Indianapolis, IN, USA

Abstract:
Accurately and automatically identifying conditions reportable to public health is an ongoing challenge. Machine learning approaches may be more effective and efficient at reporting of public health notifiable diseases when compared to existing rules-based approaches. Using real-world transactional clinical laboratory data from multiple health systems, we trained four machine learning algorithms, analyzed their performance, and compared the results. Machine learning models exhibited high accuracy suggesting they could be used for operational purposes.

Introduction
Accurately and automatically identifying conditions reportable to public health authorities is an ongoing challenge (1). Despite the increasingly ubiquitous use and promising published performance measures for machine learning in this decade, their potential for accurately and automatically identifying and reporting diseases to public health authorities has previously only been hypothesized (2). To assess whether broad adoption of these methods is practical, we sought to assess the feasibility of implementation and performance characteristics of these approaches. Such results have the potential to contribute to emerging evidence-based best-practice guidelines for using machine learning methods that support reporting of public health notifiable diseases. To generate such evidence, this study determines the performance of machine learning algorithms using routinely available clinical data.

Materials and Methods
We extracted 12 months of HL7 v2 laboratory transactions for four infectious diseases: hepatitis C, hepatitis B, chlamydia, and gonorrhea; from the Indiana Health Information Exchange (IHIE), the nation’s largest and longest-running health information exchange (3). These transactions contained both negative and positive notifiable disease cases. We created a gold standard training and test datasets for each disease by manually reviewing and labelling the HL7 laboratory transactions as positive (reportable) or negative (non-reportable). We preprocessed the transactions by removing formatting (e.g., punctuation and HL7 formatting), stop words, stemming, and identifying words in negated context using the Negex algorithm (4). Each laboratory transaction was then transformed into a feature vector containing counts for each token in the negated and non-negated context.

Using these feature vectors, we selected four machine learning models implemented in R, which included a 3-layer neural network, random forest, support vector machine, and logistic regression. Two of the algorithms were selected based on our prior experience with random forest and logistic regression (2), and the remaining two based on their wide use and prevalence in the literature. We trained the models using varying combinations of parameters: four feature set sizes (5, 10, 15, and 20 features); three ngram sizes (1, 2, and 3-grams); and four diseases (chlamydia, gonorrhea, hepatitis B, and hepatitis C). Therefore, we had (4 x 3 x 4 x 4) combinations of parameters for 192 analytical units. For each of the analytical units, we calculated sensitivity, specificity, positive predictive value, accuracy, F-score and area under the receiver operator characteristic curve (AUC).

Results
Manual review of 16,834 HL7 v2 messages identified N=2,441 Chlamydia reports (56% positive), N=3,432 gonorrhea reports (82% positive), N=2,500 hepatitis B reports (50% positive), and N=8,461 hepatitis C reports (11% positive). Across all four diseases and four algorithms, feature vectors using 1-grams and 20 features the range for the various performance characteristics revealed: accuracy 98-100% (p<0.0001), sensitivity 98-100% (p<0.0001), specificity 85-100% (p<0.0001), PPV 97-100% (p<0.0001), F- score 98-100% and AUC 97-100% (p<0.0001). There was no
significant difference between use of 1-grams and 2-grams. For hepatitis B using 3-grams and only 5 features, accuracy decreased to 80%. Results and specific trends for all 192 analytical units will be described during the presentation.

Discussion

Results for the 192 analytical units revealed no statistically significant differences among the four machine learning algorithms. Using feature vectors that contained 1-grams and 20 features, three of four algorithms achieved an accuracy, sensitivity, specificity, PPV, F-score and AUC of 96-100%. Calibrating the model further by choosing individual algorithms, at 1-grams and different feature sizes, performance measures increased to >98% for all diseases.

Regarding overall trends, we noted that model performance decreased when the total number of features decreased. Further, 1-grams were overall associated with improved performance when compared to models trained with 2- and 3-grams. We also noted that laboratory results for hepatitis B and C contained more complex results: they often included combinations of structured numeric and nominal values, as well as unstructured free-text interpretations. These more complex laboratory results required both 1) a larger feature set and 2) higher granularity 1-grams to achieve similar performance metrics when compared to less complex diseases like chlamydia and gonorrhea, which typically contained structured numeric or nominal values.

Conclusion

Machine learning models trained using feature vectors derived from real-world ("in-vivo") HL7 v2 transactions can feasibly and accurately identify notifiable diseases that should be reported to public health authorities in nearly every U.S. state. All four tested machine learning methods performed equally well. To attain highest performance, calibration of an algorithm to a specific disease is necessary. These results represent early evidence suggesting that machine learning algorithms can be feasibly implemented in operational systems to support automated detection of public health notifiable diseases. Our future work will evaluate machine learning performance for an expanded number of notifiable conditions and test their use in lieu of existing rules-based approaches like those in use currently within the IHIE (5).

References

Cardioprotective Drugs and Incident Dementias in Medicare’s Big Data

Fabricio S. P. Kury, MD1, Seo Hyon Baik, PhD1, Clement J. McDonald, MD1
1U.S. National Library of Medicine, Bethesda, MD

Introduction

Studies to assess a possible protective effect of cardioprotective drugs – antihypertensives and statins – on incidence of dementias demand large cohorts, detailed information about medication usage, and long follow-up because these diseases mostly occur late in life and only in a subset of patients. Among 17 studies of antihypertensives and dementias reviewed by Rouch et al., 2015,1 most had relatively limited cohorts and/or limited follow-up, leading to scarce events (new diagnoses of dementia during follow-up). In the context of statins, the 2015 review by Power et al.2 found that several secondary analyses of data from small randomized trials (≤ 6 months of duration) reported mixed conclusions, while the only two much larger ones failed to detect such protective effect.

As of mid-2017, the CMS Virtual Research Data Center (VRDC) provided exhaustive longitudinal data for up to 57 million beneficiaries ever enrolled in Medicare Part D, and up to 10 complete years of follow-up. It is a good example of big medical data. Most (96%) of dementia cases occur in Medicare-age patients (≥ 65 years), therefore the VRDC provides a prime dataset for assessing the potential effect of cardioprotective drugs on incident dementia. Here we report such an investigation using this dataset.

Methods

We analyzed all claims from a 20% random cohort of beneficiaries who originally joined Medicare based on old age criteria (≥ 65 years). To filter out prevalent cases, we excluded individuals with diagnosis of dementia within 2 years of their enrolment (washout period). We defined dementia according to the algorithm of the CMS Chronic Conditions Data Warehouse (CCW), which includes Alzheimer’s disease and other dementias. We used Part D prescription data from 2007, the first full year of Part D existence, until 2015, the last year available at the time. We only included patients who: a) were fully enrolled in Medicare Parts A, B and D; b) were never enrolled in managed care (Part C – for which claims data is not available); c) had at least 6 months of follow-up; and d) were exactly 65 years old at the end of 2007 (to filter out the important effect of age on risk of dementia). Given our large cohort size, we were able to analyze potential effects separated by drug classes. We also included proton pump inhibitors (PPIs) as a negative control. We quantified drug use by the sum of days of supply of the drugs. Our statistical analysis consisted of a competing risk regression (death is the competing risk) with time-varying covariates. The dependent variable was onset of dementia. Independent variables included all 57 chronic conditions recognized by the CCW’s algorithms, Medicaid eligibility status (as a surrogate for socioeconomic status), rural area residence, race and gender. We performed sensitivity analyses using only some of the aforementioned covariates. All source code we used, and the full set of results it produced, are available for free download from GitHub at https://github.com/fabkury/rxad.

Table 1. Hazard ratios for incident dementias across users of antihypertensives, statins, and proton pump inhibitors.

<table>
<thead>
<tr>
<th>Drug class</th>
<th>Beneficiaries</th>
<th>Wald test</th>
<th>HR (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Statins</td>
<td>331,607 (55.2%)</td>
<td>0.13</td>
<td>1.01 (0.95—1.07)†</td>
</tr>
<tr>
<td>Beta blockers</td>
<td>222,335 (37.0%)</td>
<td>4.59</td>
<td>0.94 (0.89—0.99)</td>
</tr>
<tr>
<td>Calcium channel blockers</td>
<td>155,634 (25.9%)</td>
<td>0.38</td>
<td>0.98 (0.92—1.04)†</td>
</tr>
<tr>
<td>Diuretics</td>
<td>245,374 (40.9%)</td>
<td>23.68</td>
<td>0.86 (0.81—0.91)</td>
</tr>
<tr>
<td>RAAS inhibitors</td>
<td>305,371 (50.9%)</td>
<td>10.91</td>
<td>0.91 (0.86—0.96)</td>
</tr>
<tr>
<td>Proton pump inhibitors</td>
<td>197,829 (33.0%)</td>
<td>0.27</td>
<td>0.98 (0.92—1.05)†</td>
</tr>
</tbody>
</table>

† Not statistically significant at 95% confidence level. All others are significant.

Results

From the 600,315 beneficiaries who met our eligibility criteria, 29,695 (5.0%) died, 13,444 (2.2%) developed dementia, and the remainder (92.8%) reached December 31st, 2015 and were censored. Follow-up time averaged 4 ± 2.5 years per beneficiary. In comparison to no statin use at all, continuous use of any statin was not significantly associated with risk of dementia (HR: 1.01, 95% CI: 0.95—1.07). Calcium channel blockers (HR: 0.98, 95% CI: 0.92—1.04) and proton pump inhibitors (HR: 0.98, 95% CI: 0.92—1.05) were also non-significant. Beta blockers, diuretics, and inhibitors of the renin-angiotensin-aldosterone system (RAAS) were significantly associated with reduced risk by
6%, 14% and 9% respectively. Female gender (HR: 0.91) and lung cancer (HR: 0.85) were associated with decreased risk. On the other hand, the following are some of the covariates that were significantly associated with increased risk: Black (HR: 1.31) and Hispanic (HR: 1.19) races, Medicaid eligibility (HR: 2.86), depression (HR: 2.06), stroke or TIA (HR: 2.11), hypertension (HR: 2.22), traumatic brain injury (HR: 2.28), schizophrenia (HR: 2.96). The hazard ratios and confidence intervals for the drug classes are presented in Table 1. In sensitivity analyses, when not including any covariate besides drug use, proton pump inhibitors (HR: 1.75), as well as beta blockers (HR: 1.34) and calcium channel blockers (HR: 1.25), were significantly associated with increased risk, while statins and diuretics were non-significant. When adjusting for race and socioeconomic status, statistical significance remained the same, but with less expressive hazard ratios: 1.47, 1.25 and 1.13 respectively. When adjusting for all covariates we had, PPIs became non-significant, diuretics became significantly protective (HR: 0.86, 95% CI: 0.81—0.91), and beta blockers became borderline-significantly protective (HR: 0.94, 95% CI: 0.89—0.99). Statins were non-significant in all models.

Discussion

The crude sensitivity analyses (without any covariates beside drug use) showed significant associations of drug use with increased risk. However, as we added more covariates the hazard ratios become smaller, then non-significant, and changed direction (decreased risk instead of increased) in the case of beta blockers. Diuretics, which were non-significant in sensitivity analyses, were significantly “protective” in our main analysis. Notably, statins were non-significant in all analyses we performed, in contrast to the two large observational studies that showed risk reductions by as much as 20%, but in agreement with the few randomized trials on this topic.2

In the context of antihypertensives, diuretics exhibited the largest risk reduction (-14%), which was similar in magnitude to the risk reduction found for angiotensin receptor blockers in the only other study (about antihypertensives & dementia) of size comparable to ours (Li et al., 2010). The other antihypertensives were also significantly protective, except for calcium channel blockers, albeit with less expressive hazard ratios.

We had originally included proton pump inhibitors as a negative control. Contradicting our expectations, after we started our work an observational study was published reporting increased risk of dementia due to use of proton pump inhibitors.3 In our analyses we found that such effect was absorbed by confounders to the point of becoming statistically non-significant. The large size of our cohorts increase the noteworthiness of statistical non-significance in our results.

In spite of using death as a competing risk, we still found death rates to be seemingly influential over the hazard ratios for dementia. For example, lung cancer had hazard ratio of 0.85 for dementia but 7.0 for death, suggesting that this disease looks protective against dementia because these beneficiaries die before having the “chance” of acquiring dementia. Conversely, other conditions such as other cancers, heart failure, atrial fibrillation, hypertension, among others, were associated with increased risk of both death and dementia. Therefore, understanding the influence of death over the hazard ratios for dementias remains challenging.

Conclusion

We were unable to find statistically significant associations between use of statins or proton pump inhibitors and incident dementias. Antihypertensives, except for calcium channel blockers, exhibited protective effect. Further study is needed to establish whether cardioprotective drugs prevent dementia and whether proton pump inhibitors cause it, with particular interest in comprehensive analyses and the potential influence of death rates.

References

Generating a Test of Electronic Health Record Narrative Comprehension with Item Response Theory

John Lalor, MS¹, Hao Wu, PhD², Li Chen, MS², Kathleen Mazor, EdD³, Hong Yu, PhD FACMI ³, ⁴
¹College of Information and Computer Sciences, University of Massachusetts, Amherst, MA; ²Psychology Department, Boston College, Chestnut Hill, MA; ³Department of Quantitative Health Sciences, University of Massachusetts Medical School, Worcester, MA; ⁴Bedford Veterans Affairs Medical Center, Center for Healthcare Organization and Implementation Research, Bedford, MA

Introduction

Allowing patients direct access to their electronic health record (EHR) notes has been shown to enhance medical understanding and improve healthcare outcomes. However, EHR notes contain complex medical terms, abbreviations, and domain specific jargon which patients have difficulty comprehending1–3. Health illiteracy is associated with a number of adverse health outcomes, including repeat health problems, insufficient follow up, and failure to follow post-visit instructions4, 5.

Instruments exist to evaluate health literacy in individuals6–8, but many focus on single term comprehension or word pronunciation. Existing instruments also have a number of problems that affect their ability to measure health literacy⁸, ⁹. EHR note comprehension requires knowledge that goes beyond terms. For example, one needs to comprehend relations between terms and clinical scenarios that incorporate complex relations and multiple events.

In this work, we report on the development of a new instrument to test patients’ ability to comprehend EHR notes. Our instrument consists of a test set of question-answer pairs that are based on the semantic content of EHR notes and selected using the psychometrics method Item Response Theory (IRT)¹⁰.

Methods

Our framework for question generation and validation consists of three steps. First, we developed a hybrid model integrating topic models and hierarchical clustering to automatically select representative EHR notes from a large hospital EHR system. Second, we developed a protocol to have researchers and medical doctors generate questionns from those notes. Finally, we used the crowdsourcing platform Amazon Mechanical Turk (AMT) to obtain responses for the generated questions and fit a model of ability using IRT to validate the questions.

We selected six conditions according to ICD-9 diagnostic codes to retrieve EHR notes from the University of Massachusetts Memorial Hospital EHR system. For each condition we randomly selected 1000 notes and annotated each note with MetaMap¹¹ to map the notes to Unified Medical Language System¹² (UMLS) concepts. We ran topic modeling on the notes according to the UMLS concepts and hierarchically clustered the notes according to topic similarities. One representative note was selected from each cluster, resulting in 30 notes (5 per condition). Our assumption is that the representative notes contain information that would be present in other similar notes.

Notes were presented to a group of researchers and medical doctors, who were asked to identify key sections of the note (a sentence or phrase from the EHR) and generate three statements about the section: (i) a paraphrase, (ii) a meaning change, and (iii) a distractor. 154 questions were created, 83 of which were selected for inclusion in the AMT task after manual inspection. Questions that were too specific for a single patient or dealt exclusively with lab results were removed.

AMT workers were given questions related to one of the 6 conditions and asked to select the correct definition of the concept from the three choices generated by the researchers. We collected responses from 1000 AMT workers to fit the IRT models. These responses were then used to fit an IRT model of latent ability and identify the subset of questions that can model ability. The subset of questions was subsequently used in a follow-up AMT task to confirm the validity of the model.
Results

Of the 83 questions used in the first AMT task, 55 were retained as good questions for evaluating latent ability. All 55 were retained after the subsequent confirmatory analysis. Retained questions were significantly more likely to be parsed as full sentences as opposed to noun phrases ($\chi^2 = 17.44, p < 0.001$). Question difficulty parameters range from -2.5 to 0, which is an appropriate range for evaluating low health literacy. The question set can be used to discriminate between low levels of ability in patients EHR note comprehension.

Discussion

In this work we identified a set of questions that can be used in a test of EHR note comprehension by applying the psychometric method of IRT. Using crowdsourced responses we model question difficulty and discriminatory ability to identify those questions that can discriminate between levels of ability across the normal human scale. We found that questions that consist of full sentences and not just noun phrases are more likely to be retained as good questions. This question set was generated directly from EHR notes, meaning that the questions have direct relevance to patients with the same condition. The questions are specific enough to a single topic that they can benefit a specific set of patients, but general enough that they are not unique to a single patient’s care and therefore not useful for other patients. Together they consist of a test to measure a patient’s latent ability to comprehend concepts from EHR notes.

These questions can be used as-is as a test of EHR note comprehension. The test can be administered to patients in preparation for receiving their own EHR notes. In addition, the framework we presented can be replicated to extract notes and questions for other diseases and conditions. Future work will evaluate existing tools for improving comprehension of EHRs to determine if they have a significant effect on the latent ability trait of test-takers. Finally, this question set can be used as a comparison set for systems that automatically generate questions from EHR notes.

References

Comorbidity Miner: An Open Source Interactive Tool for Mining Disparate Electronic Health Data Sources

Ashley S. Lee, MS1,2, Indra Neil Sarkar, PhD, MLIS2, Genevieve B. Melton, MD, PhD4, Yuanqing Liu, MA2,3, Vivekanand Sharma, PhD2, Elizabeth S. Chen, PhD2

1Data Science Practice, 2Center for Biomedical Informatics, 3Department of Biostatistics, Brown University, Providence, RI, USA; 4Department of Surgery and Institute for Health Informatics, University of Minnesota, Minneapolis, MN, USA

Introduction
The explosion of resources available to healthcare researchers — machine learning libraries, publicly accessible health data, and the pervasiveness of electronic health record (EHR) systems — has resulted in many studies and open source software for mining clinically actionable knowledge from patient data. However, a challenge persists in integrating knowledge from disparate data sources that use different coding systems and terminologies to represent diseases or conditions at varying levels of abstraction. This feasibility study aimed to develop an open source interactive tool for performing comorbidity analyses and demonstrate use for EHR and public health data in the context of anxiety disorders. This tool aims to address the knowledge integration challenge at multiple steps in the underlying pipeline by introducing: (1) versatility in the data ingestion step, (2) automated mapping of diagnosis codes to clinically significant groupings in the cohort selection and data mining steps, and (3) interactive visualization of comorbidities.

Methods
Comorbidity Miner is a web app built in R Shiny (v0.14.1) that performs data preprocessing, terminology mapping, cohort selection, data mining, and visualization. Preprocessing is customized to each data source, with subsequent steps standardized to work with generic data formats. The International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) was initially selected as the terminology standard for the generic format, which can be mapped to other terminologies such as the Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) and grouped by Clinical Classifications Software (CCS) as single- and multi-level categories or ICD-9-CM chapters. Apriori and SPADE algorithms implemented in arules (v1.5-0) and arulesSequences (v0.2-16) R packages are used to generate associations and sequences for co-occurring conditions at different levels of support and abstraction. Interactive visualization of results is done using the arulesViz (v1.2-0) and googleVis (v0.6.1) R packages.

Figure 1 shows the interface for data, cohort, and terminology selection, as well as some data mining parameters (e.g., support, confidence, lift, and maximum length) and filters (e.g., sequences containing “Anxiety”). To demonstrate the tool’s versatility, data representing clinical and public health perspectives were instantiated in the app: (1) Medical Information Mart for Intensive Care III (MIMIC-III; v1.3) database, which contains EHR data including ICD-9-CM diagnosis codes from Beth Israel Deaconess Medical Center from 2001 to 2012, and (2) Adverse Event Open Learning through Universal Standardization (AEOLUS; v1.0) database, which includes normalized data about SNOMED CT encoded drug reactions from case reports contained in the FDA Adverse Event Reporting System Database (FAERS).

Results
A cohort with any ICD-9-CM code in the CCS multi-level category 5.2 for Anxiety Disorders, was selected from each data set. Association rules and sequential patterns were generated at four levels of abstraction (ICD-9-CM codes, CCS for ICD-9-CM single- and multi-level categories, and ICD-9-CM chapters) and at varying levels of support, lift, and confidence for each cohort. Figure 2 shows a graph visualization of a subset of sequences with various filters and post-processing steps for MIMIC-III and AEOLUS. Source code at: https://github.com/bcbi/comorbidity-miner.

Discussion
Preliminary findings suggest that public health data may be complementary to clinical data, highlighting the potential value of Comorbidity Miner for facilitating knowledge integration for any disease or condition. The conditions found in AEOLUS are limited to drug adverse events, but are similar to those identified in MIMIC-III when abstracted to CCS categories (e.g., anxiety disorders associated with other psychiatric conditions, cardiovascular and urinary diseases, diabetes, and gastrointestinal disorders established in medical literature). Comorbidity Miner is being actively enhanced to add functionality such as incorporating other types of data (e.g., medications, procedures, and labs), accommodating data sources that use other terminologies (e.g., ICD-10-CM) and groupings (e.g., CCS for ICD-10-CM, PheWAS, and Unified Medical Language System hierarchies), and enhancing visualization capabilities.
Figure 1. Comorbidity Miner interface (left) and sequential pattern mining tab (right).

Figure 2. Graph visualization for association rules for Anxiety Disorders in: MIMIC-III (439 patients, support>0.02, sort statistic=lift, node size=lift, node color=lift) (left) and AEOLUS (1069 cases, support>0.01, sort statistic=confidence, node size=confidence, node color=lift) depicting adverse event co-occurrences (right).

References


Acknowledgments: This work was supported by NIH grants R01LM011364 and R01LM011963.
Usability and Feasibility of a Mobile App for Nursing Student Handoff Education

Mikyoung Lee, PhD, RN1, Catherine Bailey, PhD, RN, CNE2, Peggy Mancuso, PhD, RN, CNM, CNE1, & Patti Hamilton, PhD, RN1

1Texas Woman's University College of Nursing, Denton; 2Texas Woman's University College of Nursing, Dallas, TX, USA

Abstract

This study examines the usability and feasibility of a mobile nursing handoff reporting app (ON-SHARE) as an educational tool for baccalaureate senior nursing students to enhance handoff communication skills. The study uses a randomized pretest-posttest design. The usability of the ON-SHARE app is measured by 'think-aloud' comments, user-performance, duration of handoff reporting, and a usability questionnaire. Changes in students' confidence in handoff communication are compared within and between the experimental and control groups.

Introduction

Handoff is defined as the communication process that transfers essential patient information among healthcare providers during patient transitions in care. For decades, the importance of effective handoff communication has been highlighted through research findings of its relationships with patient safety and quality of care. In 2016, however, the Joint Commission1 reported communication failure as the second and third leading cause of serious unexpected, but preventable events (“sentinel events”) in U.S. healthcare settings with “handoff” documented explicitly as a contributing factor. With an increasing focus on strategies to improve handoff communication, studies on nursing handoff communications have accrued considerably2. Nevertheless, few studies have addressed educational designs and the effects of handoff training on clinical nursing education. The need for handoff education in pre-licensure nursing programs is certainly recognized by research findings that handoff failures have been implicated in seven out of eight novice nurses’ near miss and adverse events3.

Current literature reviews indicate that nursing students reported limited opportunities to develop and enhance handoff communication. The process of learning handoff communication in pre-licensure nursing programs is often unpredictable and dependent on clinical experiences4. Students remarked that they were unclear about communication components that constitute best practices5, 6, and they felt disorganized, anxious, and fragmented during change-of-shift interactions5, 7. Various omissions related to patient safety (e.g., missing patient names, incorrect infusion pump settings, and erroneous parameters) appeared in handoff during the clinical rotation of students8. The American Association of Colleges of Nursing's Baccalaureate Essentials9 includes appropriate and safe handoff communication skills as an essential competency of baccalaureate nursing students. Concurrently, there is increased attention towards innovations that integrate health information technology (HIT) to enhance patient safety. However, in many health care organizations, students and faculty have limited or no access to electronic health records because of logistics, security, and privacy issues10.

Students need opportunities to practice the handoff process to increase their confidence, promote teamwork, and enhance critical thinking skills. It is hypothesized that the use of the mObile Nursing Synergy HAndoff REporting (ON-SHARE)11, 12 provides a solution to these challenges in handoff education in a pre-license nursing program. The ON-SHARE app offers an innovative tool that students and faculty can use to document patient assessments, report patient conditions, plan care, and review practice performances during handoff simulations with it. This study aims to examine the usability and feasibility of a mobile nursing handoff reporting app (ON-SHARE) as an educational tool for baccalaureate senior nursing students to enhance handoff communication skills and to ultimately contribute to patient safety and quality of care in clinical learning experiences.

Method

The study uses mixed-methods, incorporating a randomized pretest-posttest design. The ON-SHARE app was developed by integrating the American Association of Critical-Care Nurses’ Synergy Model into the interface design11, 12. A convenience sample of 68 senior nursing students is randomly assigned into experimental and control groups (34 each). Prior to any handoff education and practice, all students complete the confidence in handoff communication skill questionnaire, which consists of 13 questions using a 5-point Likert scale (‘Most Confident’ (=5) to ‘Least
Confident’ (=1)) and an open-ended question for student feedback about the helpfulness of the simulated-handoff practice. With two simulated patient scenarios, the experimental group uses the ON-SHARE app for handoff practices and provides usability data measured by “think-aloud” comments, user-performance screen captures, duration of handoff reporting, and a user satisfaction questionnaire. The control group experiences standard handoff procedures using a paper-based template; the duration of handoff reporting is measured. Students in both groups complete the post measurements of handoff communication skills. The accuracy score is measured by checking if 25 items that correspond to patient condition changes are updated and reported in students’ handoff reports. Total accuracy scores are generated from each correctly captured/updated patient information item (‘Accurate’ (=1) and ‘Inaccurate’ (=0)).

**Results**

The following findings are reported: 1) differences in the accuracy scores associated with capturing patient condition information on handoff reports between the students in the experimental group and the students in the control group, 2) differences in confidence in handoff communication skills of the students within the experimental group between pre- and post-handoff practices, 3) differences in confidence in handoff communication skills after handoff practices between the students in the experimental group and the students in control group, and 4) usability evaluations from the analyses of ‘think-aloud’ comments, user-performances, duration of handoff reporting, and usability questionnaires collected with the use of the ON-SHARE app for handoff practices.

**Discussion/Conclusion**

The senior students’ clinical practices require well-thought out assessments and syntheses of the care rendered to patients. The handoff training via the ON-SHARE app will help students learn how to collect/synthesize patient data, promote an understanding of patient problems and situations, plan interventions, and communicate patient condition information to other healthcare providers while facilitating the application of theory to practice through the opportunity to use/review the HIT app. Strategies for an adoption/implementation of the ON-SHARE app for handoff education in a nursing program will be planned considering lessons from the process of this study and students’ comments.

**References**

Clustering Vital Sign Observations Using Unsupervised Random Forest

Dae Hyun Lee¹, Meliha Yetisgen, Ph.D.¹,²
¹Biomedical & Health Informatics, ²Linguistics, University of Washington, Seattle, WA, USA

Introduction

Vital signs are the primary observation in clinical practice. Compared to lab test values, they do not alter subtly due to the compensatory reflex. If altered, those vital signs tend not to be sustained for a long period because aggressive first-line interventions are applied in most cases to prevent further life-threatening deterioration. In addition, vital signs are measured at least once a shift, so they can be ideal features when illustrating a patient’s baseline physiology throughout their hospital stay. In this work, we conducted a clustering analysis using an unsupervised random forest (URF) dissimilarity matrix and k-means clustering to identify vital sign patterns over intensive care unit patients.

Method

MIMIC-II¹ demo dataset was used for the analysis. The dataset includes 4,000 patients with 34,828 intensive care unit (ICU) patient-days. In total, 41.33% of patients died in hospital. Regarding patient demographics, the average age was 72.20 years old (standard deviation: 16.28). For hospital admission, 4,438 emergency, 408 elective, 201 urgency and 27 newborn admission were presented among 5,074 hospital admissions. There were 5,844 ICU admissions among those patients with an average length of stay of 5.46 days (standard deviation: 9.00).

The proposed clustering analysis was based on the hypothesis that abnormal vital sign patterns would be observed frequently among expired patients. We considered systolic blood pressure, diastolic blood pressure, heart rate, respiratory rate, and body temperature to be vital signs, although consensus varies in this regard. As the dataset presented the vital signs as time-series observations, proper abstraction was needed to conduct a clustering analysis. The observations were grouped based on the same hospital day for each patient, and calculated summary statistics (minimum, maximum, average) for each variable. In addition, we introduced sustainment and instability quantifiers from our previous study². Sustainment quantifiers aimed to quantify the atypia of variables using the p-value from t-test and instability quantifiers aimed to quantify instability using the p-value from f-test. Both statistical tests compared population and the sampled observations.

URF³ is a method used to calculate dissimilarities between instances. The technique introduces synthetic data generated from original data, with the objective of the random forest model being to distinguish the former from the latter. Synthetic data is generated by shuffling features by columns, which will introduce a new dataset that ignores inherent feature correlations with the new class label while preserving the distribution of each feature with the same number of instances. In our study, a similarity matrix was calculated using terminal node assignment from random forest sub-trees. The similarity between two instances Iᵢ and Iⱼ, similarityᵢⱼ, was calculated by dividing the number of subtrees that assign the same terminal node on two instances with the total number of subtrees. The dissimilarity matrix was derived by the following: dissimilarityᵢⱼ = √1 − similarityᵢⱼ. K-means clustering was used to cluster instances using the resulting dissimilarity matrix.

Result

To find a URF model for the dissimilarity calculation, various model complexity was considered by changing hyper-parameters, and we chose the one with the lowest skewness (negatively skewed) on the probability of original instances. In an ideal case, we would expect the ideal URF model to assign a low probability of original instance to synthetic instances (positively skewed) and a high probability of original data to original instances (negatively skewed). In addition, we defined cluster mortality as the proportion of the instances from expired patients within the cluster to evaluate the quality of k-means clustering results. For choosing the number of clusters, k, we choose the smallest k that achieves two-fold difference between the highest cluster mortality and the lowest cluster mortality.

We used 34,420 day-level abstracted instances for the cluster analysis after excluding instances with missing values and instances of newborns. The selected URF model resulted in -0.5333 Pearson’s skewness for the assigned probabilities on original instances. We selected k as 20 where the highest cluster mortality (76.73%) was 2.58 times higher than the lowest cluster mortality (29.72%) to generalize our findings although higher k will achieve higher inter-cluster mortality difference.
**Discussion**

Figure 1a presents the cluster mortality for each cluster after k-means clustering, \( k = 20 \). As can be seen from the figure, cluster 16 included more vital signs from expired patients, and observed heart rate and respiratory rate in the cluster tends to be comparably lower while other variables followed population distribution (Figure 1b). For cluster 0, on the other hand, instances with lower heart rate and higher systolic blood pressure were gathered compared to population distribution. In addition, Cluster 3 included the majority of instances (76.73%), while other clusters included the remaining instances (29.72%), which were distributed with minor variations. The k-means clustering approach, using dissimilarity, first aggregated instances that deviated from population distribution, while instances wherein all vital signs followed population distribution were assigned a lower priority in clustering tasks.

The information gain from the selected URF model showed that it assigned different weights to features when classifying original and synthetic instances, thereby indicating that feature contribution would also be different when calculating dissimilarity between instances (Figure 1c). Unweighted Euclidean distance is a common way to calculate the distance between instances when conducting k-means clustering. However, it assumes the contribution of features equally when calculating distances. On the other hand, the URF dissimilarity considers the feature relationships represented in the original dataset. Therefore, URF dissimilarity is more feasible compared to Euclidian distance for the clustering task on clinical variables because the features are known to interact each other.

**Conclusion**

Due to the magnitude of EHRs, it is hard to review all patient vital signs manually to verify potential abnormalities. Therefore, we provided a clustering analysis using patient-day abstraction to identify the groups of potentially abnormal patient day-level vital signs, which could present during patient deterioration.

We are currently in the process of creating a comprehensive clinical dataset from University of Washington Medical Center and Harborview Medical Center. This new dataset will include structured clinical observations, unstructured clinical notes, and time-stamped clinical events created during hospitals stays of a large population of patients. We are planning to conduct the clustering analysis using a broader feature space.

**References**


Patient-Reported Outcomes in Clinical Use: Clinician Perspectives
Cynthia LeRouge, PhD1, Andrea Hartzler, PhD 2, Liz Kellogg, MPH1, Mary Beth Hasselquist, MD1, Danielle Lavallee, PhD, PharmD1
1: University of Washington, Seattle WA; 2: Kaiser Permanente Washington Health Research Institute, Seattle WA

Introduction
Patient-Reported Outcomes (PROs) such as functional health, quality of life (QoL), and well-being represent a patient’s perspective about their health without interpretation or filtering from clinicians or caregivers. Growing evidence demonstrates that collecting PROs increases clinician awareness of patients’ concerns1, reduces symptom distress2, and enhances patient-clinician communication without increasing visit length2,3. When patients share PROs with their healthcare team, patient satisfaction, perceptions of QoL, and clinical outcomes improve3. Thus, shifts towards patient-centered healthcare and payment reform initiatives call for the inclusion of PROs in clinical practice. Given this environment, evolving healthcare systems are increasingly exploring the use and integration of PROs into routine clinical care.

Although there is promise in clinical use of PROs, most clinicians and care teams do not routinely or systematically capture such metrics. Rather, clinicians generally gather informal assessments of patient outcomes, often based on their own impressions. Many clinicians remain skeptical about using PROs in clinical care and uncertain of how to incorporate this data into routine practice. Although clinicians express concern that adding PRO displays to clinical workflow could add complexity, some studies highlight the increased efficiency and usability of PRO measures that are integrated with existing EHR systems. The development of meaningful and sustainable PRO capture and reporting systems for clinical practice requires further study of the perspectives, needs, and characteristics of clinicians.

Pilot work that informed the current study through interviews with a diverse sample of clinical stakeholders (n=12) across specialties revealed the following clinician perspectives: a) the need for long-term collection of PRO to achieve value; b) the need for new tools to both capture and report PROs as part of practice workflow; c) concerns about data visualization tools that may slow or impede workflow; d) concerns over what PRO data might show and how their practice might change as a result of sharing and having to explain such data; and e) concerns about how to interpret data of questionable validity or accuracy, and f) lack of clarity regarding how to manage missing data or data that conflicts with clinical findings. Despite those obstacles, pilot participants voiced enthusiasm for leveraging PRO data for enhancing healthcare practice and outcomes.

Funded by the Agency for Healthcare Research and Quality (AHRQ), we are expanding on pilot findings by further exploring the design, implementation, and workflow for use of PROs in clinical care at University of Washington Medicine. UW Medicine is performing a system-wide deployment of the EPIC PRO module. The context provides an opportunity to extend early pilot interviews to more deeply discern clinician perceptions of PRO use and workflow, the challenges and affordances resulting from integration of PROs with electronic health record (EHR), and visual preferences for PRO displays. The clinician network from which we recruited participants includes both novices to PRO systems as well as seasoned PRO users with experience using homegrown and other commercial systems.

Methods
We are conducting in-depth interviews with clinicians engaged in implementing the EPIC PRO module to explore their perceptions about clinical use of PROs. The 19 participants interviewed to date represent the following specialties: orthopedics, cardiology, urology, and general surgery. The semi-structured interview protocol contains two sections. The first section explores perceptions of PRO use and workflow and both challenges and affordances associated with integration of PROs with an EHR. The second section of the protocol asks participants to react to various graphical PRO visualizations (i.e., line graphs, bar charts, and scatterplots) and detailed aspects of each visualization. The protocol includes coverage for three types of PRO visualizations: 1) longitudinal PROs for a specific patient; 2) comparative PROs for one patient relative to other patients; and 3) performance information that compares PROs for a clinician’s or practice’s patients to other patients.
Interviews are recorded, transcribed, and checked for accuracy. Two independent coders perform open coding with periodic reconciliation of general themes aided by a third reviewer who also performs a detailed review of all coded quotes post-reconciliation. This procedure aligns with other published open coding reconciliation procedures.

**Results**

**PRO Use:** Interest and enthusiasm for the use of PROs in patient care was greater when PROs provided information that participants felt they could not obtain by conventional means (e.g., clinical tests and general conversation). Participants indicated that they would most frequently use PRO data to discern longitudinal trends of PROs as a check of patient status and would use PRO visualizations that compared their patient to other patients to assist with shared decision making. Even when probed, none of the clinicians interviewed indicated that they would ordinarily review PROs data outside of encounters with patients, barring situations where they may be conducting a research study and using PRO data to support that study.

Regarding the use of PROs for quality improvement and practice performance, participants’ reactions varied widely between strong support and strong opposition. Mirroring results from our pilot related to using PRO data to make comparisons among clinicians, several participants expressed concerns over data ownership and unintended consequences of making PROs available outside the clinic, whether accessed by system administrators or the public.

**PRO Workflow:** Although electronic PRO capture from patients outside of clinic visits was perceived to be vastly more efficient for staff workflow, participants recognized this would not be feasible for all patients and that electronic means of capturing data from patients in waiting rooms (e.g., tablet, kiosk) and paper back-ups would be required. Data capture outside of clinic visits was preferred to ensure that the most recent points of data were included. Regarding review of PRO data captured, most participants reflected time constraints and indicated that they would “glance” at PRO data to check patient status prior to entering the exam room and possibly again during an encounter to share with a patient, particularly if PRO trends were favorable after a procedure or to assist with shared decision making. Although all participants expressed the preference for integration of PROs in the EHR to facilitate workflow by avoiding the need to connect with an external system, there were mixed perspectives regarding the actual need to integrate clinical data with PRO data.

**PRO Visualization:** Participants universally preferred simple graphs (e.g., line and bar graphs) to other graphical options (e.g., box plots and scatter plots) and referenced pictographs as being useful to share with patients. The preference for illustrating benchmarks or minimally clinically important difference scores on PRO visualizations varied widely among specialties. For PROs involving an intervention (e.g., surgery, new medication), participants preferred a clear indication of the point in time of the intervention indicated on longitudinal graphs.

**Discussion**

Overall, our findings suggest that many clinicians perceive a PRO system as a potentially valuable addition to augment their patient care process if preferences for use and visualizations are met and if PRO workflow is streamlined. Our findings reveal many commonalities among clinician perspectives of PROs as well as a few nuances among specialties. This suggests that a moderate degree of customization may be needed for a system-wide deployment of PROs to meet clinician expectations of PRO usefulness and ease of use. In addition, our findings reveal both the opportunity to share PRO data across specialties as well as the challenge of defining a systemwide scale for common measures of interest (e.g., quality of life or depression) to avoid redundancy. Our work to date provides referential data regarding patient process and preferences; future work will integrate these findings with patient perspectives or PRO systems.

**References**

When to re-order laboratory tests? Learning lab shelf-life

Gal Levy-Fix, MA1, Sharon Lipsky Gorman, MS1, Noémie Elhadad, Phd1
1Department of Biomedical Informatics, Columbia University, 622 W. 168th Street, New York, NY, USA

Introduction

A lab test result reflects the health state of a patient for only a certain time period, after which the test is outdated and needs to be re-administered. We refer to this duration as the lab shelf-life. Currently this information is not centrally available anywhere and largely depends on implicit knowledge of doctors. In this work we propose an automated method to learn lab-specific shelf-life or shelf-lives (if there are several) by mining lab orders patterns reflected in patients’ electronic health record (EHR). Lab shelf-life information has multiple use cases including encoding in decision support tools such as those aimed to prevent over-testing or those reminding to test as well as determining the informational utility of available data for modeling tasks.

Previous work have utilized occurrence of lab orders as inputs for probabilistic phenotyping and clinical order recommendations.1–3 Manually identified patterns in the time between lab orders have been used to identify guideline adherence, emergence of new practices, and order context.4, 5 By contrast to previous work, our proposed method automates the identification of prevalent order patterns using time between orders across many patients.

To identify lab shelf-life we utilize an anomaly detection method originally developed for the analysis of long-term time series. We implement this method on patient EHR data from New York-Presbyterian Hospital (NYPH) for 69 lab types and evaluate our findings based on the method’s ability to identify common gap lengths in the data (internal validity), clinical interpretability, and its utility in modeling phenotypes (external validity). The method was found to perform well on all three tasks.

Methods

To identify prevalent patterns in the time between lab orders we calculated the time gaps for each lab in the records of about 14,000 patients (Figure 1(a)). The frequency of each gap length was then aggregated for each lab over all patients in the studied cohort. We then applied an anomaly detection method called Seasonal Hybrid ESD (S-H-ESD)6 on each lab series (Figure 1(b)). S-H-ESD combines seasonal time-series decomposition7 and the generalized Extreme Studentized Deviate (ESD) test.8 The method removes any seasonal and trend detected in the data before implementing the ESD estimation, since anomalies are more difficult to detect in data with such forces at play. To make identified peaks robust and clinically relevant we excluded all identified gap lengths longer than 365 days and which were exhibited in the records of less than 40 patients.

Figure 1: Method approach for a specific lab (A1c). Given all patients’ A1c measurement gaps the method finds gaps at about 3 months and 6 months; consistent with clinical guidelines for A1c testing.
Experimental Setup

To evaluate the performance of the proposed method we designed three complimentary evaluation tasks. In Task 1 we set out to assess the **internal validity** of the anomaly detection method to find peaks in lab-gap frequencies. We do this by comparing gap lengths found by the method to those found by visual inspection of histogram graphs of the data as a sort of sanity check. Task 2 evaluates the **clinical interpretability** of the identified gap lengths by a review of a domain expert. The domain expert was to assess whether the identified gap-lengths correspond to clinically meaningful shelf-lives. Task 3 focuses on task-based evaluation, where the task is to learn high-throughput phenotypes at scale (K=250 phenotypes). The Phenome model was used and its standard metrics for quality were used to assess the learned phenotypes qualitatively and quantitatively.\(^1\) We compare the output of Phenome to one that incorporates gap lengths.

Results

In Task 1, the average precision at K across the labs was found to be 90% and the average recall at K was 89%. Hence, a large majority of the identified gap lengths were confirmed visually in the lab histograms, while some manually identified gap lengths were missed by the method. Before the post-processing steps the average precision at K was 79% and average recall at K was 86%. Hence the post processing steps removed several false positive gap lengths according to the manual review. In Task 2, the results for 96% of the labs were found to be largely consistent with the medical opinion of the expert (such as that of Hemoglobin A1c seen in Figure 1(b)) while a few lab- gaps identified in the data remained unexplained by medical guidelines. Finally, in Task 3, the Phenome model using the alternative lab feature was found to preform as well as the original Phenome model in terms of log-likelihood of unseen patients and in terms of the qualitative assessment of the clinical expert.

Discussion

We show that contextual anomaly detection method can be used to identify laboratory shelf-life lengths from patients’ lab order histories. Identified shelf-lives using this method were found to be highly consistent with a visual evaluation of the lab order patterns and clinical knowledge, with few exceptions. Most of the analyzed labs were found to have more than one shelf-life, pointing to multiple context of use. The shelf-lives were shown to be useful in categorizing patient lab data into gap-length types for phenotype modeling.

References


Acknowledgments This work was supported by NLM T15 LM007079 Data Science Supplement to GL and NSF award #1344668 to NE.
User-Centered Design of a Collaborative Genetic Variant Interpretation Tool

Wayne H Liang, MD MS¹, Bas de Veer, MSc², Diane M Korngiebel, DPhil², Annie T Chen, MSIS PhD², Gail P Jarvik, MD PhD², Peter Tarczy-Hornoch, MD², David R Crosslin, PhD²
¹University of Alabama at Birmingham, Birmingham, Alabama,
²University of Washington, Seattle, Washington

Introduction

Precision genomic medicine is reliant upon accurate genetic variant interpretation knowledge¹. However, multiple studies have reported high levels of inter-rater discordance in variant interpretation²⁻⁵. Contributing factors include heterogeneity in sequencing pipelines, inadequate knowledge sharing, and institution-specific variant interpretation criteria⁶. In 2015, the American College of Medical Genetics and Genomics (ACMG) and the Association for Molecular Pathology (AMP) published joint guidelines standardizing variant interpretation⁷. The Clinical Sequencing Exploratory Research (CSER) consortium evaluated the guidelines with 97 test variants across nine laboratories and found inadequate inter-laboratory concordance (K alpha = 0.76, 66% discordance rate⁸). Contributing factors include calculation errors, differing guidelines interpretation, and information siloing. Concordance improved significantly after a collaborative consensus process with inter-rater discussion and knowledge sharing.

Computerized tools supporting collaborative variant interpretation may improve consistency in genetic variant interpretation⁹⁻¹¹. We report on a user-centered design of an openly available computerized tool that supports role-based collaboration, knowledge sharing, and consensus-making in variant interpretation.

Methods

Part A. We used literature review and informal expert input to a) characterize the existing variant interpretation workflow; and b) create initial designs for a collaborative workflow, web-based graphical user interface (GUI), and technology architecture for a collaborative variant interpretation tool. Part B. We used user-centered design methodology, including user interviews and design feedback sessions with molecular geneticists, to a) further characterize the existing workflow; b) define user requirements and desired functionality; and c) refine designs for the collaborative workflow, GUI, and technology architecture. User sessions were audio-recorded, transcribed, and coded for themes using an analysis methodology based on Grounded Theory.

Results

Part A. We characterized the existing workflow and developed a proposed collaborative workflow based on the Delphi Method, an established structured communication process to facilitate consensus among multiple experts¹². We developed an initial GUI and REDCap-based technology architecture to support this workflow (Figure 1C).

Part B. Four faculty molecular geneticists representing different laboratories at the University of Washington enrolled in the user-centered design process (six invited, response rate = 67%). We conducted four 45-60 minute interviews, which were transcribed and coded for thematic content. User feedback data were used to further characterize the existing workflow, define user requirements, define desired functionality, and refine designs.

In our revised design, MEGA (MEdicine Gene Annotation) is a tool integrated within the existing workflow to support role-base collaboration, knowledge management, and knowledge sharing for variant classification (Figure 1A). MEGA consolidates all clinical, phenotypic, and variant annotation data required for variant interpretation. InterVar, an open-source guidelines-based pathogenicity classification calculator, is used to perform semi-automated classification predictions⁸. MEGA consolidates interpretation opinions from collaborating geneticists and external consultants. The Delphi Method-based workflow was removed based on user feedback. MEGA supports the report author in finalizing the variant interpretation. Variant data is saved in an internal knowledge base and can be exported to external systems (e.g. ClinVar).

The GUI was revised to support the revised workflow (Figure 1B). The Interpretation Panel is designed to resemble the ACMG/AMP Guidelines evidence table. This interface captures and displays the evidence and interpretation assignments entered by the predictive rules engine, all collaborators, and the user, to enable collaboration. The
Evidence Panel displays all necessary phenotypic and annotation data for classification. The technology architecture is unchanged (Figure 1C).

Figure 1. Final Design. (A) Workflow. (B) Graphical user interface. (C) Technology architecture.

Discussion

We report on a user-centered design of an openly available tool supporting role-based collaboration, consensus-making, and knowledge sharing in variant interpretation. We used user-centered design methodology, which is not commonly used in bioinformatics, to design this tool. Our experience demonstrates that user-centered methodology uncovers rich user data that can significant change the direction of the design process. The Delphi Method-based workflow was eliminated based on user feedback. Future direction include conducting additional user interviews and design sessions across multiple sites, building a working prototype, conducting usability testing with the working prototype, and testing within an actual laboratory workflow.

References

Promoting Adoption and Effective Use of Continuous Patient Monitoring Technology in the Acute Care Setting

Graham Lowenthal1, Stuart R. Lipsitz, ScD1, Catherine S. Yoon, MS1, Perry G. An, MD2, Suzanne Salvucci, RN2, Christine Shaughnessy, RN2, Sharon Keogh, RN2, David W. Bates, MD, MSc1,3, Patricia C. Dykes, PhD, RN1,3

1Brigham and Women’s Hospital, Boston, MA; 2Newton-Wellesley Hospital, Newton, MA; 3Harvard Medical School, Boston, MA

Introduction
In the acute care setting, patients often exhibit physiological signs of deterioration, such as adverse changes in heart and respiratory rates, hours before experiencing critical events.1,2 Continuous patient monitoring systems have demonstrated potential to aid clinical staff in detecting these early physiological warning signs and taking appropriate actions sooner to prevent patient deterioration before conditions become too severe.1,3 In January of 2013, a continuous monitoring system was implemented on all adult medical-surgical beds in an Eastern Massachusetts acute care hospital. This particular system uses contact-free sensors under the patient’s mattress to continuously measure heart rate, respiratory rate, and motion. Through real-time analysis of its collected measurements, the system can accurately recognize trends, identify clinical deterioration, and deploy audible and visual alerts to influence earlier care team intervention.3 Controlled trials had linked use of this system to improved quality of care, length of stay, and cost outcomes.1,3 However, observation of real-world staff workflow at the implementation site revealed much resistance to system adoption and to adherence to practice guidelines, limiting projected benefits. This podcast presentation will describe the process our team used for identifying and overcoming barriers to adoption and effective use of the continuous monitoring system.

Methods
We established a committee of key stakeholders: physicians, clinical champions (nurses, managers, educators, and patient care assistants) from all study units, researchers, and vendor partners. Our committee hosted regular meetings to identify barriers to adoption and effective use of the continuous monitoring system. We then designed practical solutions based on staff and unit needs using a cyclical Plan Do Study Act (PDSA) framework. We developed strategies and tactics for addressing each discovered barrier (PLAN), executed the intervention (DO), evaluated effectiveness of our methods (STUDY), and revised then redeployed our approach (ACT). Our desired outcome was improved system adoption and adherence to practice guidelines. We measured the success of our PDSA interventions toward achieving this outcome through observed changes in staff alert response times: the interval between the triggering of an alert from the continuous monitoring system, patient assessment at bedside, administration of treatment, and electronic dismissal of the alert on the patient’s bedside monitor. The smallest response times were considered the most ideal.

Results
Working with hospital staff, we identified and addressed several barriers to adoption and effective use of the continuous patient monitoring system. Barriers, as well as the strategies and tactics we developed to overcome those barriers, are detailed in Table 1. Using our PDSA solutions, staff produced statistically-significant improvements (trend P-value < 0.001) to response times between January 2014 and January 2017. During this period, monthly mean cardiac alert response times improved from 105 (range: 17 to 312) minutes to 21 (range: 5 to 43) minutes. Mean respiratory alert responses improved from 91 (range: 26 to 360) minutes to 26 (range: 8 to 61) minutes. Mean combined (cardiac and respiratory) alert responses also improved, from 92 (range: 17 to 360) minutes to 25 (range: 5 to 61) minutes.

Discussion
Using our PDSA framework allowed us to effectively address barriers to adoption and support proper use of the continuous monitoring technology in the acute care setting. We chose to employ a PDSA because its cyclical nature enables users to continuously adapt improvement approaches to changing needs. This allowed us to facilitate consistent progress in an often hectic and constantly changing environment. Involving hospital staff, the system’s end-users, was imperative for the implementation to succeed. Empowering staff to voice their own concerns and observations allowed for much more effective identification of barriers. Likewise, permitting staff to contribute directly to development of practical solutions and workarounds facilitated greater unit engagement in improvement
efforts. We used response times as the unit for measuring progress toward outcomes because in order to maximize the benefits of continuous monitoring systems staff must respond to alerts as quickly as possible. Therefore, the response time improvements that we facilitated may indicate greater opportunities for staff to recognize, act appropriately on, and prevent patient deterioration earlier. To test this, future research will focus on measuring the impact of our documented improvements with the continuous monitoring system on clinical outcomes.

<table>
<thead>
<tr>
<th>Identified Barriers</th>
<th>Strategies and Tactics Developed to Address Barriers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inadequate Training/Education</td>
<td>Host a Kickoff Event to emphasize importance of the system and proper use. Provide in-the-moment hands-on and didactic teaching sessions to promote awareness of and familiarity with using system and features.</td>
</tr>
<tr>
<td>Clinical Workflow Challenges</td>
<td>Empower staff as clinical champions to aid in identifying barriers and solutions, and to enforce practice guidelines on units. Increase leadership involvement to support change and promote consistency. Strengthen partnership with vendor to address concerns and drive improvements based on user feedback. Regularly review anecdotes which present examples of peers’ best practice workflows.</td>
</tr>
<tr>
<td>Lack of Communication</td>
<td>Integrate system reporting and scripted discussion into rounds. Incorporate system alerts and trends into documentation and handoff reporting. Champions share content and findings from committee meetings with colleagues.</td>
</tr>
<tr>
<td>Mismanagement of System Alerts</td>
<td>Coach staff on how to reduce unnecessary alerts and mitigate alarm fatigue. Enable nurses in special cases to change alert settings within specified limits. Provide managers with daily-emailed alert reports to review performance with individuals and correct issues in real time. Establish unit protocol to standardize methods of receiving and acknowledging alerts notifications. Appoint central person (such as Unit Coordinator) to monitor status boards and notify staff about unaddressed alerts.</td>
</tr>
<tr>
<td>Lack of Multidisciplinary Involvement</td>
<td>Provide peers with counseling to promote understanding of system use and encourage them to support nursing practice and communication. Host medical grand rounds to educate interdisciplinary team about the system and its importance. Involve PCAs to assist nurses in lesser system priorities, such as patient turns and system upkeep.</td>
</tr>
<tr>
<td>Inability to See Device’s Value</td>
<td>Regularly review Care Team &quot;Saves&quot;: peer-witnessed cases in which proper system use allowed for early intervention and prevention of patient deterioration.</td>
</tr>
<tr>
<td>External Barriers</td>
<td>Reaffirm and align leadership at every level to support staff through all changes and keep them engaged in improvement efforts.</td>
</tr>
</tbody>
</table>

Table 1. Identified Barriers to Adoption and Effective Use of the Continuous Monitoring System, and Strategies and Tactics Developed to Overcome Them

Acknowledgments: The authors wish to acknowledge the clinical champions at Newton-Wellesley Hospital, and EarlySense, Ltd for research grant support. David W. Bates, MD, MSc has also consulted for and received consulting fees and Honoraria from EarlySense, Ltd.

References

132
Recurrent Neural Networks for Classifying Relations in Clinical Notes

Yuan Luo

1Department of Preventive Medicine, Northwestern University, Chicago, IL, US

Introduction

Relation extraction from text documents is an important task in knowledge representation and inference in order to create or augment structured knowledge bases and in turn support question answering and decision making. In the clinical domain, extracting relations from clinical narratives has been gaining attraction over the past decade. Part of the advances in the state-of-the-art specialized clinical NLP systems for identifying medical relations have been documented in the 2010 i2b2/VA challenge workshop, which attracted international teams to address successive shared classification tasks. The challenge focused in part on identifying the relations that may hold between medical problems and treatments, between medical problems and tests, as well as between pairs of medical problems. Support Vector Machines (SVMs) were a common theme among the systems that tackled the relation extraction task in the challenge. All challenge participating systems involved heavy feature engineering; they explored lexical, semantic, syntactic, general domain and medical domain ontology features etc.

Systems that use many human engineered features often do not generalize well to new datasets. Recent progress on applying recurrent neural networks (e.g., Long Short Term Memory, LSTM) to clinical datasets aims to minimize the amount of engineered features and has achieved some success on specific tasks such as early detection of heart failure onset using structured clinical data and de-identification of patient notes. In this work, we use recurrent neural networks to classify the semantic relations from the 2010 i2b2/VA challenge, which include relations from the following three categories. The medical problem-treatment relations include: treatment administered for medical problem (TrAP), treatment is not administered because of the medical problem (TrNAP), treatment improves medical problem (TrIP), treatment causes medical problem (TrCP), a patient’s medical problem has deteriorated or worsened because of or in spite of a treatment being administered (TrWP), and no relation. The medical problem-test relations include: test has revealed some medical problem (TeRP), test was performed to investigate a medical problem (TeCP), and no relation. The medical problem-medical problem relations include: two problems are related to each other (PIP) and no relation.

Methods

Word embedding

For NLP applications, recurrent neural network models are most used together with word embeddings. The word embedding is designed to capture semantic similarity of words. Neural language modeling tools such as word2vec can learn embedding vectors from an unlabeled large text corpus, based on the word’s contexts in different sentences. We trained word vector on clinical notes from MIMIC-III database using word2vec tool. Of note, the MIMIC-III dataset contains clinical notes for over 46,000 patients with a total of 100 million words.

Segment level LSTM for relation classification

Some of the top performers in the i2b2/VA relation classification challenge reported that distinguishing the concepts vs. context text, and further subdividing the context text into text preceding the first concept, between the concepts, and succeeding the second concept. To explicitly model the concept and context text, we construct segment level LSTM architecture for relation classification, as shown in Figure 1. We divide the concept and context text into five segments: before the first concept (preceding), of the first concept (concept 1), between the two concepts (middle), of the second concept (concept 2), and after the second concept (succeeding). For each segment, we feed the sequence into a LSTM layer then a pooling layer to learn the \( n_{hu} \)-dimensional hidden feature representation. We then concatenate the hidden features from the five segments into one \( 5n_{hu} \)-dimensional feature vector, input the feature vector to a softmax layer to produce the relation class labels. In our input data from the i2b2/VA challenge, some of the concepts are annotated on the head word, while others are annotated including preceding and succeeding modifiers. To more consistently capture concept characteristics, we allow the text of the concepts to be padded before and after by neighboring words.

Experiments and Results

In order to fairly compare our models with those from the i2b2/VA challenge participants, we used the same split of the training and test datasets by the challenge organizers. To optimize the parameters for our models, we further randomly selected 10% of training dataset as the validation set and used it to tune the parameters. For segment level LSTM models, we experimented with a series of padding sizes (from 3 to 10) for padding the concept text with their...
context. We experimented with multiple numbers of hidden units (100, 150, and 200 in this work). For regularization on the LSTM models, we used dropout technique in the output of the pooling layer. We compared our systems’ performance with the top four teams from the i2b2/VA challenge participants, as shown in Table 1. From the comparison of the micro-averaged f-measure, we see that segment LSTM model ranks the second in classifying the TrP relations, the third in TeP relation classification, and the third in PP relation classification. Overall segment LSTM achieved good performance that are comparable to state-of-the-art systems from i2b2/VA challenge participants with heavily engineered features, even though segment LSTM uses only the basic word embedding as features.

Table 1 Performance of the LSTM models. Performance of top four i2b2/VA challenge participating systems are also included for comparison. Best micro-averaged f-measures are in bold. R = Recall, P=Precision, F=F-measure.

<table>
<thead>
<tr>
<th>System</th>
<th>Problem-Treatment (TrP) Relations</th>
<th>Problem-Test (TeP) Relations</th>
<th>Problem-Problem (PP) Relations</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R</td>
<td>P</td>
<td>F</td>
</tr>
<tr>
<td>Segment LSTM mean</td>
<td>0.641</td>
<td>0.683</td>
<td>0.661</td>
</tr>
<tr>
<td>Roberts et al.</td>
<td>0.686</td>
<td>0.672</td>
<td><strong>0.679</strong></td>
</tr>
<tr>
<td>deBruijn et al.</td>
<td>0.583</td>
<td>0.750</td>
<td>0.656</td>
</tr>
<tr>
<td>Grouin et al.</td>
<td>0.646</td>
<td>0.647</td>
<td>0.647</td>
</tr>
<tr>
<td>Patrick et al.</td>
<td>0.599</td>
<td>0.671</td>
<td>0.633</td>
</tr>
</tbody>
</table>

Conclusion

In this work, we proposed the first system based on LSTM for classifying relations from clinical notes. We showed that our segment LSTM models achieved a micro-average f-measure of 0.661 for classifying medical problem-treatment relations, 0.800 for medical problem-test relations, and 0.683 for medical problem-medical problem relations. These results support the use of LSTMs for classifying relations between medical concepts, as they show comparable performance to state-of-the-art systems from the i2b2/VA relation classification challenge while requiring no manual feature engineering.

References

A Pilot Randomized Trial to Train Vulnerable Primary Care Patients to Use an Online Patient Portal Website

Courtney Lyles PhD, Lina Tieu MPH, Stephen Kiyoi MLIS, MS-HAIL, Shobha Sadasivaiah MD, Mekhala Hoskote, Neda Ratanawongsa MD MPH, Urmimala Sarkar MD MPH, and Dean Schillinger MD

University of California San Francisco, Department of Medicine at Zuckerberg San Francisco General Hospital and Trauma Center, San Francisco, CA

Introduction
Patient portals, websites that provide patients and caregivers access to personal health information, are becoming ubiquitous in the US, especially with the support of Meaningful Use financial incentives. Previous research has documented substantial usability barriers, especially among patients with limited health literacy. We conducted a randomized pilot trial to determine the effectiveness of a scalable online video-based training program to increase use of a portal website among patients in a safety net healthcare setting. We further explored whether training outcomes varied by mode of delivery (administered in person vs. distributed via a link) or by patient characteristics.

Methods
Based on our previous work documenting usability barriers to portal use, we created a curriculum with simple instructions and 11 how-to videos for accessing features of the patient portal available in the public healthcare system in San Francisco (videos available at http://www.sfhealthnetwork.org/about-myshealth), distributed through a link to an online learning platform, Learnerweb. These videos were developed in close consultation with both a patient advisory board (members included primary care patients from the system with varying experience with technology) and project advisory board (members included experts on health information technology research, adult education, and clinical operations) to ensure that the content was appropriate in terms of digital and health literacy levels of the target population.

We then randomized 93 patients to receive either: 1) an in-person tutorial with a trained research assistant versus 2) a link to watch the videos on their own. To identify eligible individuals, we identified English-speaking patients with at least 1 chronic condition (heart disease, diabetes, hypertension, heart failure, asthma, kidney disease, anxiety, depression) through a chart review of patients with recent clinic visits, as the portal website was only available in English. Through phone screening, we excluded individuals with no previous email use. We also excluded those who reported use of the portal website in our setting in the previous year. We assessed health literacy status using a screening item, categorizing participants noting any lack of “confidence filling out forms” as having limited health literacy.” Follow-up measures were collected via phone after participants had a clinic visit ≥3 months post-baseline, or at 6 months if they had no visit.

The primary outcome was portal use (proportion of patients who signed on to the portal website and number of log-ins) between 3 and 6 months post-training, assessed through electronic chart review. Data on use of specific portal features was not available through the electronic health record. We collected baseline and follow-up survey measures regarding participants’ perceptions of the importance of the portal website, experiences with their chronic illness healthcare (using the Self-Efficacy for Managing Chronic Disease Scale,2 and Morisky Patient Medication Adherence Scale3), and digital health literacy ratings (using 4 questions from the eHealth Literacy Scale4). Using the tracking capabilities of the online training platform, we also measured whether participants viewed the training videos over the course of the study.

Results
The trial enrollment was completed in September 2016. From a list of 833 provider-referred patients, we contacted 377 via phone, of whom 52 declined and 186 were ineligible primarily due to lack of email use (108, 58%). Of the remaining 139 individuals, 93 participants were enrolled. The mean age of the sample was 54 years, and 51% had self-reported limited health literacy; 60% were non-white, 52% were female, 45% were in fair or poor health, 25% reported speaking English less than very well. While the majority of participants used the Internet and email daily (76% and 65%, respectively) and reported that it was very important to access their medical information
electronically (75%), at baseline over a quarter (27%) reported that they believed they lacked the skills to use a portal website to manage their healthcare.

By design, everyone in the in-person training arm watched the videos at least once, compared to 44% of those in the take-home arm (p<0.01). Overall, among the 93 enrolled, 35 participants (41%) logged into the portal at least once at follow-up (including the 17 those who had logged into the portal before the study); 18 participants (21%) logged into the portal website during the 3 to 6 months post-training. These proportions did not differ by the in-person vs. take-home training assignment (37% vs. 45% (p=0.41) for having ever logged in and 21% vs 20% (p=0.91) for logging in 3 to 6 months post-training). The mean number of log-ins to the portal website in the 3 to 6 months post-training was 1.4 (s.d.=3.2). Number of log-ins did not differ by study arm (average 1.2 for take-home, 1.6 for in-person, p=0.57). Overall, we found significant improvements in ratings of digital literacy (p<0.01) and medication adherence (p<0.01) from baseline to follow-up, and a trend toward an increase in self-efficacy for managing chronic disease from baseline to follow-up (p=0.09).

When we analyzed rates of portal login during the follow-up period, we found significant differences in portal use by age, health literacy status, and number of chronic conditions. Participants who were 60 or older were more likely to log in to the portal than their younger counterparts (33% vs. 13%, p=0.03). In addition, participants with adequate health literacy were more likely to log in than those with limited health literacy (35% vs. 7%, p<0.01). Participants with only 1 chronic condition were significantly more likely to log in than those with 2 or more conditions (35% vs. 15%, p=0.04).

To measure per-protocol outcomes, we examined the subgroup of participants in either arm exposed to the training videos at least once (n=60): 23% logged into the portal in the follow-up period, which was not statistically higher than the 15% among those who did not watch the videos (p=0.41). However, there was a significant improvement in patients’ self-reported skills in using the portal website: 65% agreed or strongly agreed that they had the skills to use the website at baseline, and this improved to 80% at follow-up (p=0.04). There was also a trend toward higher patient ratings of the Internet being useful for their health needs (58% at baseline to 74% at follow-up, p=0.09). Notably, we found that interest in using the portal among those who accessed the training decreased significantly: 91% reported moderate to high interest in using the portal at baseline versus 75% at follow-up (p=0.02).

Of note, 58% (n=26) of those in the take-home arm and 49% (n=20) of those in the in-person arm had been registered for portal access prior to enrolling in our study but did not have self-reported experience in using the website. Among this subsample of already registered patients, 12% (n=3) in the take-home arm and 25% (n=5) in the in-person arm logged into the portal following training, but this trend did not reach significance among this limited sub-sample (p=0.23).

Conclusion

Patients receiving care in safety net health care settings are interested in using a portal to help manage their healthcare. An online training program appeared successful in improving patients’ confidence in their skills to use the website, motivating subsequent 6-month portal use among about 1/5 of the patients regardless of the training vehicle (i.e., in-person vs. take-home link). While this moderate rate of portal use is likely higher than the secular trends in portal use we have observed in this safety net healthcare system, there is room for improvement in patient engagement in this and similar settings. In particular, a one-time training session may be insufficient for patient activation, particularly for patients with limited health literacy. More research is needed to determine what additional patient-level and provider-level interventions are needed, as well as to determine how to make portal interfaces more engaging and user-friendly.

References

Leveraging SNOMED CT Relationships for Mapping Disease Codes with Different Levels of Abstraction between EHR Systems

Sina Madani, MD, PhD, Shari Just, RN, BSN, Scott D. Nelson, PharmD, MS
S. Trent Rosenbloom, MD, MPH, Asli Weitkamp, PhD
Vanderbilt University Medical Center, Nashville, TN

Abstract
There is an increasing demand for rapid and consistent mapping of data elements to terminology standards to support data migration between clinical information systems. However, the level of detail used for representing clinical concepts in the source and target systems may not be the same. Furthermore, such translation relies heavily on domain experts’ knowledge and consistency in mapping. We have developed a framework for efficient mapping between two clinical data sets within two EMR systems using SNOMED CT hierarchical relationship.

Introduction
Mapping clinical concepts to standard terminologies like SNOMED CT leads to numerous benefits, including unambiguous data capture and support for semantic interoperability in clinical and research domains. Migrating between Electronic Health Record (EHR) systems requires rapid and coherent translation of clinical data into the new system. To support the integrity of such data during EHR migration, mapping between source and target concepts should be kept consistent and semantically meaningful. Such mappings can maintain the quality of patient data for health care providers during the implementation and data migration to a new EHR system.

Vanderbilt University Medical Center (VUMC) is in the process of data conversion from various modules of the its home grown EHR, such as Family Medical History (FMH), to a commercial EHR system. Although clinical concepts used for family medical history in our legacy and target EHR systems were both bound to SNOMED codes, FMH data migration process underwent an additional mapping procedure due to the differences in the granularities of the concepts represented in the source and target data sets. To improve the pace and consistency of the mapping between the two systems we used existing SNOMED CT hierarchical relationships, such as the is-a relationship. The main objective of this study is to evaluate a framework for identifying equivalent concepts between two data sets that represent family medical history findings with different levels of abstraction in two EHR systems. We have used structural resemblance of the concepts within SNOMED CT hierarchy in order to find the best match based on the closeness of the parent concepts from the source to the target concepts.

Methods
We identified all Family Medical History (FMH) concepts represented by SNOMED CT codes within the current VUMC EHR system that have no direct match to SNOMED CT concepts in the target EHR system. All retired SNOMED concepts were excluded or replaced with an active concept based on the 2016 version of SNOMED CT. We used a graph version of SNOMED CT (v2016AA) to identify all hierarchical parents, grandparents, and ancestors of FMH concepts in our legacy system by using SNOMED CT is-a relationships. During the query process, we calculated the distance of each extracted parent concept to its child FMH concept. We then compared SNOMED CT codes associated with parent concepts to the existing SNOMED CT codes in the target EHR system to find possible matches based on the closest distance. Subject matter experts (SMEs) evaluated the mapping to validate whether the suggested closest parent concept was the best match among the target data set repository (Figure 1).

Figure 1. Mapping schema between VUMC and target EHR concepts
Results

We identified 2,014 unique SNOMED CT codes associated with Family Medical History (FMH) context in patient records within our legacy system without any direct equivalency to SNOMED CT codes bound to the target EHR FMH concepts. However, one or more parent concept(s) of the source FMH concepts in our system had a direct match to the limited set of 637 SNOMED CT codes within the target EHR system (Figure 1).

The SMEs confirmed 88.6% of the mappings of the nearest parent concept as appropriate (Table 1). For 34 VUMC FMH concepts, with more than 1 parent with matching target EHR concepts, another parent concept other than the nearest concept was considered as the best map to a target EHR concept. For example, for Myoclonic Seizure two parent concepts identified as Movement Disorder and Seizure Disorder with a distance of 2 and 6 to Myoclonic Seizure respectively. Despite the fact that Movement Disorder is the closest concept to Myoclonic Seizure (based on SNOMED CT hierarchy) the domain expert decided to select Seizure as the best match within the target EHR. Moreover, in 124 cases where there was only 1 suggested parent match to a target EHR concept, the SMEs decided to reject the map and pick another concept from the target EHR concept that seemed more relevant. For example, there were 44 instances of different types of Arrhythmias in FMH (like Tachyarrhythmia) where the nearest parent with a map to a target EHR concept, Heart Disease, was replaced with Conductive Disorder of Heart by SMEs. The reason for this and similar cases (like cancer related concepts) were due to the fact that less granular concepts can be located at the same hierarchical level as their semantically related concept with more granularities under the same axiom or other axioms in SNOMED CT. There were 72 VUMC FMH concepts that SMEs couldn’t find any direct match between their parent concepts and the target EHR concepts or any other relevant concepts in the target list.

Table 1. Family medical history concepts mapping results based on the number of parent concepts match

<table>
<thead>
<tr>
<th></th>
<th>FMH concepts with 1 parent match</th>
<th>FMH concepts with &gt;1 parents match</th>
</tr>
</thead>
<tbody>
<tr>
<td>Closest parent concept match accepted</td>
<td>891</td>
<td>893</td>
</tr>
<tr>
<td>Closest parent concept match not accepted</td>
<td>124</td>
<td>34</td>
</tr>
<tr>
<td>None of the parent concepts accepted</td>
<td>17</td>
<td>55</td>
</tr>
<tr>
<td>Total</td>
<td>1032</td>
<td>982</td>
</tr>
</tbody>
</table>

Discussion

Standard clinical terminologies, such as SNOMED CT, that tend to cover medical domain are often used in EHR systems. These terminologies, with different levels of granularities, may overlap with other terminological sources that are used in different applications. Identification of equivalent clinical concepts with lexical and semantic similarities between terminologies is an important contributory factor in interoperability and re-usage of data of the underlying clinical data repositories. Therefore, during the migration process between clinical information systems, ad hoc mapping between existing and new terminologies becomes an important and often arduous task.

Conclusion

The successful automatic mapping of family medical history concepts based on SNOMED CT hierarchies has facilitated the migration and reuse of patient data from our legacy system into the new EHR system. The results of this study can be used as a framework in similar environments where concepts with higher granularities can be mapped to the relevant but less-granular concepts based on SNOMED CT is-a relationship. We have also used lexical matching method for mapping the concepts that were not bound to SNOMED CT codes in the source and target data repositories. However, additional evaluation of this method was beyond the scope of this study.

References

Development of a Wrist-Worn Sensor to Improve Medication Adherence:
Designing for Diverse User Behaviors and Technology Preferences

Jenna Marquard PhD1, Barry Saver MD MPH2, Swaminathan Kandaswamy MS1, Vanessa Martinez1, Jane Simoni PhD3, Joanne Stekler MD MPH3, Deepak Ganesan PhD1, Sean Noran1, James Scanlan PhD2,
1University of Massachusetts, Amherst, MA, 2Swedish Medical Center, Seattle, WA, 3University of Washington, Seattle, WA

Introduction

Studies have demonstrated that, for optimal outcomes of antiretroviral therapy (ART), both high (at least 95%) and consistent levels of adherence are necessary for continuously suppressing HIV replication1,2. For preventing new HIV infection with pre-exposure prophylaxis (PrEP), daily dosing is currently the only recommended regimen3. These required adherence levels are substantially higher than the 80% typically used for other health conditions, and maintaining such high levels of adherence is a challenge. For providers, lack of reliable, real-time information about ART and PrEP adherence makes it difficult to identify patients struggling with adherence until adverse outcomes, such treatment resistance, have already occurred. Current medication adherence-promotion devices, for this and other patient populations, do not capitalize on advances in wearable sensors. We are developing “USE-MI” (Unobtrusive Sensing of Medication Intake), a low-cost and innovative system designed to improve measurement and adherence. USE-MI uses a wrist-worn device and a tagged medication container to unobtrusively sense individuals’ gestures related to opening a pill bottle followed by hand-to-mouth arm movement, using this data to determine pill-taking behaviors, trigger pill-taking reminders to the wrist-worn device and a smartphone app, and to generate summative data on individuals’ adherence levels. To guide the design of this device, we conducted patient interviews, questionnaires, and observations to understand their current medication-taking practices and technology preferences. Our approach and findings provide generalizable guidance for similar devices and interventions.

Methods

Participants (n=17) were current patients at clinics providing medical care and social services for persons living with HIV/AIDS. In an interview, we asked participants questions about how they store, take, and remember to take their medications and 2) their feedback about how two candidate devices (MS Band, Android Wear Watch) could support this routine. We observed participants wearing a wrist device while taking placebo pills. Participants also completed a questionnaire asking about their perceived adherences, additional medication-taking practices, and preferences for gaining feedback about their medication-taking patterns, and general demographic measures. Interview responses were coded and reviewed by two team members. Questionnaire responses were analyzed using descriptive statistics.

Results

Almost all participants were male and HIV+ taking ART, but were relatively diverse in age and race (Table 1). Participants reported taking a mean of 88% (36%-100%) of their medications over the previous 4 weeks. Participants reported taking medications once (82%) or twice (18%) per day, split between the morning (41%) and night (53%); one participant took medications at lunch. Participants reported taking their medications from pill bottles (35%), blister packs (12%), and rolls with pills heat-sealed into bags labeled with date/time of intended administration (53%); 3 participants reported deviating from this method five or more days a month (e.g., transferring containers because of travel). Participants used a range of medication reminder approaches, including placing the medications in visible locations (e.g., by the coffeepot), aspects of the storage method itself (e.g., time and date on pill roll), annotating calendars, self-motivation, and reminders from others.

Participants exhibited varied handedness, with 12 right-handed (RH), 4 left-handed (LH) and 1 ambidextrous participant. When asked to wear the wrist device and take a placebo pill from a bottle (even if they typically take pills a different way), there was significant variation in which hands participants used to perform medication-taking tasks. Figure 1 shows what percent of participants used their right hand for each task; 1 RH individual did not have a video. Over 75% of all participants used their right hand to turn the pill bottle cap, regardless of their handedness. Participants were evenly split on which hand they used to put the pill in their mouth. Most participants (82% RH, 75% LH) used their dominant hand to drink the liquid.
After testing the devices, when asked for their preference for the devices, two participants (12%) stated they would only wear the MS Band and two (12%) stated they would only wear the Android Wear Watch. Most participants (76%) had no preference between the MS Band or Android Wear Watch, though 77% of this group preferred the MS Band over the Android Wear Watch. These preferences related to a variety of criteria, including size, comfort, fashionability, other functionality, battery life, and the charging mechanism. Two preferences related to drug use arose; one participant did not want a valuable device that might be sold or stolen, another wanted a waterproof device that could withstand sweating and showering at bathhouses.

Most participants (65%) stated that they would like to use the system. Those who said no (18%) felt the feedback would be helpful, but in the end did not want to wear the device. Two participants (12%) were unsure as to whether they would use the system. Participants were most interested in receiving medication reminders via a mobile app (67% extremely/very interested), the wrist-worn device (59%), or text message (53%); phone calls and emails were not of interest. Most participants (86%) wanted to receive feedback on their pill taking over time; most (86%) also wanted to share that information with their doctor. Mobile apps (58%) and email (33%) were the preferred mechanisms for receiving this summative feedback.

Discussion

While most participants were interested in using the proposed device, our results highlighted diverse user behaviors and preferences. First, our initial design required users to store their medications in pill bottles; we are working to determine how the system could accommodate pill rolls. Second, we are working to improve medication-taking gesture recognition, given the variation in which hand users perform tasks with. Third, we will allow some flexibility in the method for sending users reminder alters and summative reports. Finally, we are developing the system to work with multiple Android Wear watches, allowing user choice in wrist device style. While we focused on designing an intervention for individuals with HIV/AIDS, the aforementioned findings may align with the preferences and behaviors of other patient populations. Our findings are limited by a fairly small sample size.

References


Table 1: Participant Demographics

<table>
<thead>
<tr>
<th>Gender</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>16</td>
</tr>
<tr>
<td>Female</td>
<td>1</td>
</tr>
</tbody>
</table>

| Age Range | 26-62 |

<table>
<thead>
<tr>
<th>Status</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>HIV+ taking ART</td>
<td>16</td>
</tr>
<tr>
<td>HIV- taking PrEP</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Race</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>White</td>
<td>9</td>
</tr>
<tr>
<td>Black</td>
<td>7</td>
</tr>
<tr>
<td>American Indian</td>
<td>3</td>
</tr>
<tr>
<td>Hispanic</td>
<td>2</td>
</tr>
<tr>
<td>Prefer not to answer</td>
<td>1</td>
</tr>
</tbody>
</table>

Acknowledgements

The National Institutes of Health (1R01MH109319) supported this work. The content is solely the responsibility of the authors and does not represent the official views of NIH.
Turning Off Medication Alerts to Reduce Clinical Decision Support Overrides

Allison B. McCoy, PhD¹, Eric J. Thomas, MD, MPH², Marie Krousel-Wood, MD, MSPH¹,³, Ryan P. Walsh, MD, MMM²,⁴, Adam Wright, PhD⁵, Dean F. Sittig, PhD⁴

¹Tulane University School of Public Health and Tropical Medicine, New Orleans, LA, ²McGovern Medical School, Houston, TX, ³Tulane University School of Medicine, New Orleans, LA, ⁴The University of Texas School of Biomedical Informatics at Houston, Houston, TX, ⁵Brigham and Women’s Hospital, Partners HealthCare, Harvard Medical School, Boston, MA

Introduction
Computerized alerts that warn clinicians about drug interactions or provide dosing guidance are commonly implemented forms of clinical decision support (CDS) to improve patient safety. However, high rates of inappropriate alerts and alert overrides prevent healthcare providers and the settings in which they work from realizing the desired improvement in clinical processes and patient care and can create additional work for already busy clinicians.¹,² Evaluation and improvement of alert implementations, and turning off inappropriate alerts, may help reduce overrides and subsequently increase patient safety.³,⁴

Methods
We previously developed an interactive dashboard that allows informatics personnel to review CDS alerts and provider responses (InSPECt: Interactive Surveillance Portal for Evaluation Clinical decision supporT).⁵ We used InSPECt to review medication interaction alerts in a large multi-specialty, ambulatory practice that provides care for all ages throughout the Houston community. During this process, we identified high override rates for a large number of alerts displayed that were deemed to rarely require a change in clinical care. As a result, we changed the alerting criteria from displaying alerts for all severities to only the highest severities (i.e., the lowest setting).

To evaluate the effect of the intervention, we extracted log data from the electronic health record (Allscripts Enterprise EHR) for 20 weeks before and 20 weeks after the intervention. We reviewed the number of alerts displayed to providers and the number of alerts overridden versus accepted, and we compared these outcomes across alert type before and after the intervention, and by week. We performed a chi-square test to identify whether there was a significant difference in the override rate between the pre-intervention and post-intervention periods.

Results
Providers received 532,672 alerts during the pre-intervention period and 191,836 alerts during the post-intervention period. Despite receiving 64.0% fewer alerts, override rates persisted at 69.2% pre-intervention and 70.1% post-intervention (p<0.0001). Figure 1 depicts overridden and accepted alerts for pre-intervention and post-intervention by alert type. Drug-condition alerts were the most frequent, with 333,204 pre-intervention alerts and 9,840 post-intervention alerts; 66.5% and 66.2% were overridden for pre-intervention and post-intervention, respectively.

Figure 1. Intervention results by alert type.
We observed the largest change in the override rate for drug-dose error alerts, with 78.1% of 85,417 pre-intervention alerts and 67.6% of post-intervention alerts overridden (p<0.0001). Changes were also significant for drug-drug (p=0.047), duplicate therapy (p<0.0001), and drug-dose alerts (p<0.0001); however, the change in override rates was not clinically significant.

Figures 2, 3, and 4 depict overridden and accepted alerts by week for all alerts, drug-condition alerts, and drug-dose error alerts. The override rates remain similar across weeks in the pre-intervention and post-intervention periods for all alerts combined and drug-condition alerts, and lower in the post-intervention period for drug-dose error alerts. We observed no other temporal trends.

**Conclusion**
Through use of a previously developed dashboard for evaluating CDS alerts, we identified high rates of overrides and inappropriate alerts. After implementing an intervention to turn off low severity alerts, we reduced the alerting rate by 64.0% compared to the pre-intervention period. Despite a statistically significant change, there was no clinically significant change in override rates (69.2% pre-intervention to 70.1% post-intervention). Additional research is necessary to further improve CDS alerts to reduce overrides and improve patient safety.

**Acknowledgments:** This work was supported by NLM Grant K22 LM 011430-01A1, a UTHealth Young Clinical and Translational Sciences Investigator Award (KL2 TR 000370-06A1), and NCRR Grant UL1 RR 024148.

**References**

Introduction
Smartphone apps for self-tracking of chronic disease hold promise both for patients trying to manage their condition and researchers aiming to characterize diseases and behavior patterns at scale. In 2015, Apple introduced ResearchKit, a platform to facilitate app-enabled research through streamlined survey functionality and wide recruitment of study participants through electronic consent. Studies that use ResearchKit have been successful over traditional studies in enrolling a large number of participants, but opportunities remain to overcome the limitations of passive recruitment, in which a ResearchKit app is released but not actively marketed, and for app engagement via active contribution of data consistently over the study period. For example, the mPower study on Parkinson’s disease had 48,104 app downloads but only 898 study participants, of which 150 were patients (0.3%), used the app for at least 5 days within the 6-month study. Fostering engagement (broadly defined as the quality of experience in using a technology that makes it appealing to continuous use) is critical for realizing the full potential of research and general tracking apps. In this study, we explore the impact of active recruitment strategies as well as the use of specific app functionalities on user engagement. The context for this work is Phendo, an app-based study to capture and characterize endometriosis, a prevalent disease (estimated to affect 10% of women) and systemic chronic condition with a high burden on quality of life, from the standpoint of patients.

Methods
Phendo aims to capture directly from patients a wide range of variables about endometriosis through time. It was developed as a ResearchKit-based tracking app. The design of the app, in particular, identifying the range of variables to track, along with determining their temporal granularity for patients to track (e.g., momentary assessment vs. day-level) was carried out through a user-centered, participatory process with endometriosis patients via 5 focus groups with 27 adult women with the disease. We incorporated a number of features beyond the vanilla ResearchKit capabilities: customized user interface of surveys to allow users to enter multiple responses to a survey question; smart notifications triggered only when a user has not tracked any data as well as “positive” notifications to encourage returning users, calendar-based review and insights screen that provide a personalized timeline of tracked data; and a Citizen screen showing how much their individual data contributed to the overall study. For this study, we explore three novel functionalities and their traction with study participants: (i) customizable surveys to track medications, hormones, supplements, foods, and exercises; (ii) ability to track specific variables on a moment-by-moment basis (repeatedly if needed) throughout the day and other variables at the day-level; and (iii) ability to track all variables retroactively. The app is available on the iTunes Store for use on iPhone, and as such passive recruitment is available. For active recruitment, we created a website (citizenendo.org) and a social media presence, maintained a blog, and engaged with patient advocacy groups and celebrities who advocate for the disease to promote Phendo. We focus on the first 100 days of the app since its launch (11/13/2016-2/20/2017).

Results
1,184 women with self-reported endometriosis diagnosis from 39 countries consented to the Phendo study, during which time no participants opted out. Participant demographics follow the adult US female population who use apps: 56.7% have at least a college degree, live in suburban (46.4%) or urban (38.1%) areas, and are White non-Hispanic (82.5%). Their median age is 29 years old, and their age at endometriosis diagnosis was consistent with epidemiological characterization of the disease. Fig. 1 shows the impact of active recruitment efforts on enrollment. In the first 100 days since the launch of the app, participants each tracked 1 to 80 days, and 50% of them tracked 6 days at least. Participants made use of the customizable surveys: 84.5% used at least 1 customized variable, and on average created 6 variables across the food, exercise, medications, supplements, and hormones categories. Overall, participants tracked both day-level (e.g., “Do you have your period today?” and “What did you do to self-manage today?”) and moment-level variables (e.g., “Are you experiencing any GI/urinary issues now?”) (Fig. 2). We note the variability in amount of tracking both at the moment- and day-level, indicating that participants made sure to track only what is relevant to them or what they deem useful to let researchers know (e.g., menstrual periods were tracked more than sexual activity). Fig. 3 shows the times at which participants engaged with the app.
to track (i.e., local time at which participant used their phone). They used the app throughout the day for momentary tracking, whereas they answered day-level questions at the end of the day (7pm is default notification time). 53% of actively tracking participants made use of retroactive tracking. Within the same day, they tracked retroactively primarily for moments (graph not shown) (e.g., tracked a moment that occurred in the morning at 4pm). But they also tracked retroactively up to several days in the past (Fig. 4), mostly for day-level questions.

Discussion
Beyond ResearchKit’s many benefits, ways to retain and engage patients are needed. We found that active recruiting was critical to not only gather interest but also to target the right population of study participants (in our case, women with endometriosis). Aligning tracking functionalities with the way in which participants experience the disease showed promise in engaging participants: they made heavy use of customized tracking and both moment and day-level surveys, with specific patterns of tracking throughout the day. Furthermore, ability to track retroactively was shown convenient and useful to participants. When used retroactively, moment-level tracking was done within the same day, while day-level tracking went back 1-2 days in time, suggesting users are reasonable in their recall. These patterns are encouraging: these functionalities enable more data points to be tracked, and there is likely validity to them given the short retroactive tracking times. Although we focus on a specific condition, engagement functionalities are applicable to other chronic disease populations that experience a wide range of symptoms dynamically over time.

References
Usability Testing to Guide Development of a Clinical Decision Support System for Substance Use Screening and Interventions in Primary Care

Jennifer McNeely, MD, MS1, Andre Kushniruk, PhD2, Sarah Farkas1, Aida Vega, MD3, Eva A. Waite, MD3, Lauren Peccoralo, MD3, Melanie Harris1, Christine Chollak1, Richard N. Rosenthal3, John Rotrosen, MD1, Joseph L. Kannry, MD3
1New York University School of Medicine, New York, NY; 2University of Victoria, Victoria, BC, Canada; 3Icahn School of Medicine at Mt. Sinai, New York, NY

Introduction
Alcohol and drug use are significant drivers of preventable morbidity and mortality that are not routinely identified or treated in medical settings.1 Screening, brief intervention, and referral to treatment (SBIRT) to address substance use in primary care is widely promoted, but has proven challenging to implement.2 Validated screening tools that can quickly and accurately identify substance use have been designated common data elements (CDEs),3 but are not widely integrated into electronic health records (EHRs). As part of a study of the NIDA National Drug Abuse Treatment Clinical Trials Network, we developed a clinical decision support system (CDSS) to support the delivery of substance use screening and interventions in primary care, and thereby facilitate data capture of the CDEs.

The four-phase study began by conducting qualitative interviews with key clinical stakeholders to identify barriers and facilitators to substance use screening, and to inform a clinical workflow for implementing SBIRT. In Phase Two, results of which are reported here, we conducted multiple rounds of usability testing to iteratively develop the content and delivery of a substance use CDSS. The later study phases will examine the performance of the CDSS in capturing the CDEs and facilitating substance use interventions, in two large health systems. This work is significant because it provides a CDSS to support the integration of behavioral health in primary care, and addresses a significant knowledge gap among medical providers, who receive little education on substance use treatment. To enhance generalizability, we developed the CDSS in a widely used EHR (Epic™) using standard functionality.3

Methods
We employed the ‘Framework for Usable and Effective Clinical Decision Support’ to guide our CDSS development, usability testing, and implementation approach.4 We had previously applied this framework in a study to develop a CDSS for antibiotic prescribing that achieved high levels of adoption and acceptance in primary care.4 Key features of the framework are 1) active and actionable decision support; 2) multiple rounds of usability testing with iterative development to increase user acceptance; 3) numerous context sensitive triggers; 4) dedicated training and support for users of the CDSS to foster adoption; and 5) support from clinical and administrative leadership.

Prior to usability testing, we drew upon findings from the stakeholder interviews to build an initial prototype CDSS in the Epic test environment. We then conducted 3 rounds (total of 27 sessions) of usability testing in June-August 2016, with 19 end-users in one primary care clinic. Participants were medical assistants and primary care providers (PCPs), including both faculty and resident physicians. Some PCPs from the initial round were invited to repeat testing in subsequent rounds, to provide feedback on whether changes made were responsive to their suggestions.

During usability testing, participants were given a clinical scenario of a patient with moderate- or high-risk alcohol or drug use, and were led through the scenario by the facilitator and asked to ‘think aloud’ as they used the EHR tools. The medical assistants used the EHR tool to enter screening results, while the PCP users reacted to the data and alerts and were asked to complete a patient progress note. Alerts notified providers that screening is due, and alerted them to a high-risk screening result. Use of the system during a medical visit was simulated by having the participant interact with a simulated patient (i.e. one of the investigators playing the role of a patient using a script). Data was collected using Camtasia™ screen capture software, audio and video recordings, and written notes.

For analysis, first all of the audio data was transcribed to create a log file. The video data was then reviewed, and usability problems were identified and annotated in the log file. Problems were classified as being either ‘human-computer interaction’ findings (e.g. number of mouse clicks) or ‘workflow’ findings (e.g. context and timing of the information delivered). Following each round of testing, the investigators met with the medical informatics team to review the results and make decisions on how to implement changes. Once the changes were completed, the subsequent round of testing was performed.
Results
Participants were medical assistants (n=3), faculty physicians (n=10), and resident physicians (n=6). The main problems identified and changes made to the CDSS in each round of usability testing are summarized in Table 1. This presentation focuses on workflow findings, which proved more challenging to solve than the human-computer interactions findings. For example, we found that participants were sensitive to the timing of alerts, which in some cases were triggered before having had a conversation with a patient. The majority of problems were identified in the first round of usability testing, after which significant changes to the CDSS were made. By the third round of testing the number of identified problems had decreased considerably and solutions focused on training issues.

Table 1. Findings and changes made in response, by usability testing round.

<table>
<thead>
<tr>
<th>Testing round</th>
<th>Problems and suggestions identified</th>
<th>Changes made to the CDSS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Round 1 (13 sessions)</td>
<td>1. Difficult for MD to see the screening result that had been entered by the medical assistant. 2. Alert for ordering a referral to social work if a patient has been in the hospital for more than 2 weeks. 3. Decision support text guidance is too lengthy for a medical visit. 4. Documenting current alcohol or drug consumption is redundant with the standard EHR social history section.</td>
<td>1. Clicking on a ‘last filed’ button shows the most recent result. 2. Triggered alert to fire when the MD selects a visit diagnosis, which is typically done at the end of the medical visit. 3. Separated guidance into two sections (counseling and plan); MD can select one or both. 4. Not possible to modify the existing social history section. We placed the CDSS screening tools adjacent to the standard EHR sections for drug and alcohol history.</td>
</tr>
<tr>
<td>Round 2 (4 sessions)</td>
<td>5. Some users prefer order sets (SmartSet) over text guidance. 6. MDs want patient educational handouts to include in patient instructions.</td>
<td>5. Created an order set for social work referrals, triggered via practice alert for high-risk patients. 6. Developed substance-specific patient education materials and worksheets that can be inserted into the EHR-generated instructions.</td>
</tr>
<tr>
<td>Round 3 (10 sessions)</td>
<td>7. Difficult for MDs to remember the steps of brief intervention.</td>
<td>7. Could not accomplish this within the EHR; created a reference card.</td>
</tr>
</tbody>
</table>

Discussion
Usability testing identified important problems with the substance use CDSS, many of which we were able to address during the development process. Similar to observations of the CDSS development framework paper, triggers and workflow proved challenging in a commercial EHR. Standard EHR functionality was not optimal for collecting the substance use CDEs because it did not support the reconciliation of the CDE data with the social work and drug fields in the social history section. We found it difficult to use the EHR to facilitate team-based behavioral health interventions in primary care; a problem that has also been reported in other studies. Notably, usability testing of our EHR-integrated substance use CDSS revealed issues related to the tool’s content, in addition to identifying more basic human-computer interaction issues. Limitations of our study include (1) the decision to include some repeat participants, which reduced our sample size and potentially complicates our analyses; and (2) having a member of the study team play the role of simulated patient, which could bias our findings. Despite these limitations, our iterative process of usability testing and content adaptation allowed us to successfully modify the EHR to support a complex behavioral intervention for substance use, though some limitations to its functionality remain. Future phases of this study will examine adoption of our CDSS in multiple primary care clinics.

References
Identification of Clinically Meaningful Clusters of Multi-morbidity in a National Cohort of Adults Using Unsupervised Learning

Che Ngufor, Ph.D., Rozalina G. McCoy, M.D., Lixia Yao, Ph.D., Lindsey R. Sangaralingham, M.P.H, Shannon M. Dunlay M.D., Nilay D. Shah, Ph.D.
Mayo Clinic, Rochester

Introduction

The number of individuals in the U.S. and around the world affected by multiple chronic conditions (MCC) has increased substantially and continue to rise.¹ Patients with MCCs require significantly more health care resources, incur high cost of care, and face greater mortality, functional decline, and worse quality of life.²³ However, identifying, treating, and improving the care and health outcomes of patients with MCCs have been hindered by the marked complexity and heterogeneity of this population. Current Comparative Effectiveness Research and Evidence Based Practices (EBP) focus on one disease at a time, or on highly prevalent co-occurring two-way or three-way combinations of conditions, thus leaving a large gap in our knowledge about how to optimally manage individuals with complex MCC. Alternatively, comorbidities are treated as being equal; with risk stratification based on weighted or unweighted comorbidity counts. Yet not all comorbidities, affect patients the same way. Nor do they incur the same burden or pose the same risk to the patient and the healthcare system. Information about the interactions and subgroups of MCC can facilitate diagnosis, enhance preventive strategies, improve quality of life, and help create smart EBP guidelines.

To address this critical knowledge gap, we use data driven unsupervised learning methods to identify clinically meaningful homogeneous groups of MCC from a large observational data and to characterize frequent disease combinations, relationships, and phenotypes of patients within subgroups.

Methods

Study population: We used data for 604,712 adults, age ≥ 18 years, included in OptumLabs Data Warehouse (OLDW), a national administrative claims database of commercially insured and Medicare Advantage beneficiaries across the U.S. A two-year sample of Medicaid beneficiaries with at least 12 months of continuous medical coverage and affected by at least two chronic conditions was used for the analysis.

Clustering and Association Rule Mining: We propose a two-stage method that allows the discovery of stable, robust, and clinically meaningful homogeneous multimorbid subgroups. First, we applied the agglomerative hierarchical clustering algorithm with the Ward’s minimum variance criterion to cluster patients based on similar multimorbid patterns. Similarity between patients was measured by the Jaccard index.⁴ Agglomerative hierarchical clustering starts by assuming that each patient was a cluster, and successively merge similar clusters to form larger clusters. Because of the hierarchical tree structure, different number of clusters can be obtained by restricting the tree at specific heights. The Ward’s method minimizes the sums of squares between clusters and merges similar clusters. To evaluate the goodness of the clustering results and select the optimal number of clusters, we apply a validation technique that makes use of external information not used to generate the clusters. Specifically, based on an assumption that individuals with similar MCC combinations are expected to consume about the same amount of healthcare resources, we validated the clustering by measuring the area under the ROC curve (AUC) for a logistic regression model predicting health utilization (Frequent ED Visits and Readmission) based on cluster memberships. By the assumption, the more homogeneous the clusters, the higher the AUCs, which implies that the AUC increases with the number of clusters. We selected the optimal number of clusters by plotting the AUC against the number of clusters and applied the elbow criterion. The elbow is the point on the graph where addition of a cluster does not lead to a significant gain in AUC and corresponds to the optimal number of clusters. As the elbow method can be ambiguous, we also applied the principle of parsimony when selecting the number of clusters.

In the second step, we applied association rule mining in each cluster to characterize frequent disease combinations (3 or more) and disease-disease association rules. A disease-disease rule is a statistically meaningful pattern identified in the data, such as "patients with {Cardiac Dysrhythmia, Cardiomyopathy, Chronic renal Failure and Asthma} tend to have {Congestive Heart Failure}". Mining frequent disease combinations can generate large number of frequent comorbidity sets, most of which might contain a small number of high prevalence conditions (e.g., hypertension, hyperlipidemia, and dementia), while missing out the low prevalence but more important rare MCCs. To address this problem, we weighted the conditions by dividing by their frequency of occurrence in the data to reflect their importance. This strategy is a reflection that not all chronic conditions affect patients equally.
Results

The median age of the study population was 66 years, with 55.51% female, and mean number of conditions per patient was 3. Based on the Wards criterion and the external validation technique, we identified 14 optimal clusters of patients with 69 chronic conditions defined by the Agency for Healthcare Research and Quality (AHRQ [https://www.hipxchange.org/Comorbidities] and 5 demographic variables (age, gender, race, and region). We also tried other merging algorithms such as the single, complete, and average linkage methods, but they all produced unstable and sparse clusters. The Wards method produced stabled, dense, and approximately equal sized clusters. Figure 1 (a) shows that the AUC for predicting health utilization increases gradually as the number of clusters increases (solid lines) compared to the poor and all flat-out performance of randomly generated clusters (dotted lines). There is a big jump in performance from the 12 and 13 clusters to 14, but after the 14th, no observed gain in AUC can be seen. Thus, 14 clusters correspond to the turning point or elbow of the curve and was selected. Clusters beyond the 14th did not add much clinical value and some were sparse (with respect to the outcome) and unstable.

Figure 1 (b) shows the distribution of the top frequent disease combinations in each cluster by the support (fraction of patients with the combination). The mean age (A) and percentage of females (F) are shown above each group. Several clusters described patient groups with the common and high prevalent conditions, however, our weighting scheme revealed some of the less prevalent conditions compared to when no weighting was used. Cluster 8 can be described as the Mental Subgroup, characterizing young female adults (18-34 years) with primarily mental illness (top disease-disease rule: \{anxiety, depression, substance-use disorder\} \rightarrow \text{bipolar}), while cluster 14 can be considered as the Hypertensive Heart Disease Subgroup, describing older male (70+ years) with top frequent disease combinations: \{Hypertension, Coronary artery atherosclerosis, and CHF\}.

Conclusion

In this study, we combined two data driven unsupervised learning techniques: hierarchical clustering and association rule mining to discover clinically meaningful clusters of patients with multiple chronic conditions. Based on the assumption that similar MCC patients tend to consume the same amount of health care resources, we were able to select 14 optimal clusters by measuring the predictive ability of cluster memberships in discriminating patients with high health care utilization. These clusters can help in comparative effectiveness research in identifying patients who are more likely to benefit from certain interventions.

References

Designing and Evaluating an Automated System for Real-time Medication Administration Error Detection in a Neonatal Intensive Care Unit

Yizhao Ni, Todd Lingren, Eric Hall, Matthew Leonard, Kristin Melton*, Eric Kirkendall*
Cincinnati Children’s Hospital Medical Center, Cincinnati, OH *Equal contribution

Introduction

Medication administration errors (MAEs) are the most common medical errors experienced by patients. Neonates are particularly vulnerable to MAEs due to medication dosing that is influenced by weight, gestational age and postnatal age. Previous efforts to reduce MAEs, including voluntary incident reporting and trigger tools, rely heavily on manual vigilance and detect only a fraction of errors. Sustaining effective and accurate detection of MAEs remains a challenging issue among pediatric healthcare institutions. Our objective is to develop an automated system that utilizes comprehensive Electronic Health Record (EHR) information to detect dosing-related MAEs in real-time and to facilitate mitigation of medication safety events in neonatal patients.

Methods

We focus on automating detection of MAEs among 11 high-alert medications prescribed to inpatients in the neonatal intensive care unit at Cincinnati Children’s Hospital Medical Center. The medications include total parenteral nutrition (TPN), lipid, intravenous fluids (IVF), insulin, morphine, fentanyl, milrinone, vasopressin, dopamine, dobutamine, and epinephrine. Our study consists of three phases: 1) development of the MAE detection system, 2) observation and evaluation of system performance without intervention (i.e., without step 4 in Figure 1), and 3) intervention and system deployment into clinical practice. This work reports results of the pilot system at the observation phase (November 2016 through February 2017).

Figure 1 diagrams the piloted MAE detection system. The system extracts medication use information from the EHR (step 1), including medication orders and modifications, medication administration records (MARs) that document doses (or infusion rates) administered to patients, and free-text physician to nurse communications that describe dose/rate changes. By utilizing logic-rules and natural language processing (NLP) techniques, the system’s detector identifies discrepant doses/rates between MARs and other extracted information (step 2). The detector builds on MAE detection algorithms from our earlier work1,2. If a discrepancy is identified, the notifier will send a MAE message to clinicians in real-time (step 3). The clinicians will investigate the notification and take appropriate action (e.g., communicating with bedside nurses; step 4). Clinician feedback will be sent to the reporter (step 5), which generates case reports for system evaluation and improvement (step 6). Using the case reports, study investigators will periodically evaluate system performance and calculate the time window after a MAE event during which a patient is potentially at risk for harm (step 7). Their feedback will be fed into the learner module to adjust system parameters (e.g., modifying regular expressions in NLP to improve dose detection from clinical narratives; step 8).

During the observation phase, study team physicians (Drs. Kirkendall and Melton) double-reviewed MARs for the targeted medications and identified all MAEs. We reviewed two months of orders for high frequency medications (TPN, IVF, lipid, morphine and epinephrine) and four months of orders for the others. Two medications (dobutamine and dopamine) had no orders during the observation period and were excluded from this evaluation. Adjudicated MAEs were used as a gold standard to assess system performance, where positive predictive value (PPV), sensitivity (SEN), negative predictive value (NPV), and specificity (SP) were calculated in aggregate and for
each medication. We also evaluated expected mitigation of exposure to potential harm following a MAE event with the implemented system. Without the system the time of exposure was calculated between the MAR time of erroneous dose/rate administration and time of documented clinician correction. With the system it was calculated between the MAR time and time of MAE notification, assuming a clinician could respond immediately upon receiving a MAE message. If a MAE was missed by the system (false negative), the time of exposure was identical to that without the system.

Results and Discussion

Table 1 presents descriptive statistics of the medication use data. The physicians reviewed 5,333 MARs for 3,124 medication orders and identified 80 MAEs (Cohen’s kappa of inter-annotator agreement: 87.8%). Both the physicians and the system identified a notable number of insulin dose discrepancies. However, the retrospective analysis suggested that many discrepancies were documentation errors rather than true MAEs. Eighteen errors resulted from non-standard documentation for specifying insulin dose change were excluded from the subsequent time window evaluation. Table 2 shows the system performance in aggregate and for each medication. The MAE detection system achieved an overall PPV of 85.5% and SEN of 81.3%. The PPV was over 95% for the majority of the medications and over 75% for medications with frequent dose adjustments via physician to nurse communication (TPN and IVF). The system achieved similar SENs across medications except lipid and morphine, where two lipid and one morphine MAEs were missed. Finally, the automated system had potential to reduce the median time window for exposure to harm, from 298 minutes to 35 minutes (p<0.001 with paired t-test). In particular, the time windows were reduced substantially for long-time intravenous medications such as TPN and lipid.

The high PPV and sensitivity suggest that the implemented system has capacity for identifying MAEs while guarding against alert fatigue. Furthermore, the system could significantly reduce a patient’s exposure to potential harm following a MAE event. Consequently, we hypothesize that the automated MAE detection system, once fully deployed, has great potential to significantly mitigate medication safety events among neonatal patients.

Table 1. Descriptive statistics of the medication use data.

<table>
<thead>
<tr>
<th>Medication</th>
<th>#Patients</th>
<th>#Encounters</th>
<th>#Orders</th>
<th>#MARs</th>
<th>MAR Error Rate</th>
<th>Observation Timeframe</th>
</tr>
</thead>
<tbody>
<tr>
<td>TPN</td>
<td>64</td>
<td>65</td>
<td>1317</td>
<td>2072</td>
<td>1.59%</td>
<td>2 months</td>
</tr>
<tr>
<td>Lipid</td>
<td>64</td>
<td>65</td>
<td>1313</td>
<td>1469</td>
<td>0.20%</td>
<td>2 months</td>
</tr>
<tr>
<td>IVF</td>
<td>121</td>
<td>125</td>
<td>350</td>
<td>1013</td>
<td>1.18%</td>
<td>2 months</td>
</tr>
<tr>
<td>Insulin</td>
<td>2</td>
<td>2</td>
<td>14</td>
<td>50</td>
<td>46.00%</td>
<td>4 months</td>
</tr>
<tr>
<td>Morphine</td>
<td>27</td>
<td>27</td>
<td>84</td>
<td>421</td>
<td>0.24%</td>
<td>2 months</td>
</tr>
<tr>
<td>Fentanyl</td>
<td>4</td>
<td>4</td>
<td>7</td>
<td>41</td>
<td>0.00%</td>
<td>4 months</td>
</tr>
<tr>
<td>Milrinone</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>24</td>
<td>0.00%</td>
<td>4 months</td>
</tr>
<tr>
<td>Vasopressin</td>
<td>3</td>
<td>3</td>
<td>9</td>
<td>70</td>
<td>0.00%</td>
<td>4 months</td>
</tr>
<tr>
<td>Epinephrine</td>
<td>10</td>
<td>10</td>
<td>26</td>
<td>173</td>
<td>4.62%</td>
<td>2 months</td>
</tr>
<tr>
<td>Total</td>
<td>127</td>
<td>131</td>
<td>3124</td>
<td>5333</td>
<td>1.50%</td>
<td></td>
</tr>
</tbody>
</table>

Table 2. Performance of the automated medication administration error detection system.

<table>
<thead>
<tr>
<th>Medication</th>
<th>PPV (%)</th>
<th>SEN (%)</th>
<th>NPV (%)</th>
<th>SP (%)</th>
<th>Median Exposure to Harm with System (minutes)</th>
<th>Median Exposure to Harm without System (minutes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TPN</td>
<td>80.0</td>
<td>72.8</td>
<td>99.6</td>
<td>99.7</td>
<td>37.0</td>
<td>584.0</td>
</tr>
<tr>
<td>Lipid</td>
<td>100.0</td>
<td>33.3</td>
<td>99.9</td>
<td>100.0</td>
<td>298.0</td>
<td>1087.0</td>
</tr>
<tr>
<td>IVF</td>
<td>75.0</td>
<td>100.0</td>
<td>100.0</td>
<td>99.6</td>
<td>30.0</td>
<td>101.5</td>
</tr>
<tr>
<td>Insulin</td>
<td>95.7</td>
<td>95.7</td>
<td>96.3</td>
<td>96.3</td>
<td>34.5</td>
<td>46.0</td>
</tr>
<tr>
<td>Morphine</td>
<td>100.0</td>
<td>0.0</td>
<td>99.8</td>
<td>100.0</td>
<td>3.0</td>
<td>3.0</td>
</tr>
<tr>
<td>Fentanyl</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Milrinone</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Vasopressin</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Epinephrine</td>
<td>100.0</td>
<td>75.0</td>
<td>98.8</td>
<td>100.0</td>
<td>31.0</td>
<td>323.0</td>
</tr>
<tr>
<td>Total</td>
<td>85.5</td>
<td>81.3</td>
<td>99.7</td>
<td>99.8</td>
<td>35.0</td>
<td>298.0</td>
</tr>
</tbody>
</table>

References

Reducing Variation in Core Measures data through Automation

Matthew A. Oberhardt, PhD¹, Hojjat Salmasian, MD PhD,¹,² Dena Goffman, MD³, David Vawdrey, PhD¹,²
¹: Value Institute, NewYork-Presbyterian Hospital, New York, ²: Division of Quality & Patient Safety, NewYork-Presbyterian Hospital; ³: Department of Obstetrics & Gynecology, Columbia University Medical Center

Introduction

The Centers for Medicare & Medicaid Services (CMS) defines a set of “Core Measures,” key metrics that hospitals are required to track in order to benchmark and improve quality outcomes. The OB-GYN core measure named “PC-02” assesses the rate of Cesarean sections (C-section) among NSTV women -- Nulliparous, Full-Term women who delivered a Singleton baby in Vertex position (i.e., numerator = C-sections, denominator = women who had NSTV deliveries) ¹. This rate is traditionally determined manually by nurse abstractors who, in our study setting, randomly assess ~40-120 delivering women per quarter (approximately 10% of deliveries) and determine who among those are NSTV (typically ~10-50 women). They then pull the C-section rate among this NSTV subsample, and extrapolate that rate to the hospital as a whole. This manual abstraction process may be hindered by lack of consistency and also by high intrinsic variance due to the small number of patients sampled each quarter. Automating data abstraction has potential to reduce costs and improve consistency in reporting, but must be done carefully to ensure quality ².

For these reasons, an effort has been underway to develop electronic Clinical Quality Metrics (eCQMs) to automate abstraction of quality metrics. A preliminary eCQM was developed for PC-02, but it has yet to be officially reported ³ or assessed to our knowledge. Here, we develop an automated algorithm for determining PC-02 C-section rates that is functional within our healthcare setting. We evaluated whether this method significantly tightened the confidence intervals on PC-02 rates and increased data consistency.

Methods

We implemented a rules-based automated method for determining PC-02 rates across four hospitals in the New York City area (see Table 1).

<table>
<thead>
<tr>
<th>Nulliparous</th>
<th>From notes: MAX(Living children)=0 (more reliable in our institution than parity)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FullTermDelivery</td>
<td>From notes: MAX(Gestation weeks) &lt;= 50 and &gt;= 37</td>
</tr>
<tr>
<td>Singletonbirth</td>
<td>ICD10 code: ‘Z370’ (Single live birth)</td>
</tr>
<tr>
<td>Vertex</td>
<td>From notes: ‘OB delivery note’ or ‘OB delivery record’ or ‘OB Labor Progress note’ or ‘OB Labor and Delivery Progress Note’. The value ‘Vertex’ in a fetal presentation field in each note.</td>
</tr>
<tr>
<td>OtherExcludedPC02</td>
<td>(Age &lt; 8 OR Age &gt;= 65 ) OR Length of Stay &gt; 120</td>
</tr>
<tr>
<td>Csection</td>
<td>MSDRG code: ‘765’ or ‘766’</td>
</tr>
<tr>
<td>NTSV_Delivery</td>
<td>Nulliparous AND FullTermDelivery AND Singletonbirth AND Vertex AND PrimaryDelivery AND NOT OtherExcludedPC02</td>
</tr>
<tr>
<td>NTSV_Csection</td>
<td>NTSV_Delivery AND Csection</td>
</tr>
</tbody>
</table>

Table 1: Pseudocode for automation of PC-02 measure. Note, when measures appear in multiple notes for a patient (e.g., parity), we take the maximum value.

Identifying appropriate rules required detailed conversations and shadowing of clinicians about how the data is entered into the EHR as well as conversations with data abstractors about how they obtain the measures. Conversations with leading clinicians and physicians assistants led, for example, to the use of the ‘living children’ field instead of the ‘parity’ field for determining that a woman is Nulliparous, as this was deemed more reliable in our records. The rules-definition process was iterative with intensive

involvement of both the data abstractors and clinicians. After automating the measure, NTSV determinations were compared directly against determinations from manual abstraction. We found that the automated methods were able to identify approximately 10x as many NTSV women as the sample examined manually by abstractors, giving NTSV rates with significantly tighter confidence intervals (Figure 1).

We performed a comparative analysis of the patients determined to be NTSV through the automated vs manual methods, and found strong overlap (average precision = 0.9, recall=0.79 across 4 hospitals for the automatic method to identify the same NTSV women from the subset manually abstracted).

**Results**

The rates for both automated and manual extraction were found to be consistent, both drawn from the same population for three of the hospitals (1, 2, and 4). However, the last hospital (number 3) was inconsistent (Table 2). The inconsistency is found in patients assessed only by the automated algorithm, in that they showed a different NTSV rate than patients assessed by both the abstractors and the manual abstractors. In contrast, the patients identified by both methods have a high overlap. i.e., for the patients that the automated method was able to judge, it often accurately assessed NSTV status (precision=0.95, recall=0.87 for NTSV identification of manually abstracted women in hospital 3). To further understand this, we looked into factors contributing to PC-02, and found that they are not always recorded consistently by the abstractors. An analysis of the ‘nulliparous’ status revealed that 9% of patients have discrepancies in their parity scores based on the data in medical notes, and 7% of abstractor parity determinations conflicted with at least one note. To address this, we worked with the abstractors to define a priority ranked list of notes that should determine parity labeling, thus improving consistency and hopefully accuracy of the manual method.

<table>
<thead>
<tr>
<th>Hospital</th>
<th>quarter</th>
<th>numerator</th>
<th>denominator</th>
<th>rate</th>
<th>p-value (manual vs auto)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Q1</td>
<td>46</td>
<td>175</td>
<td>26.3%</td>
<td>0.3</td>
</tr>
<tr>
<td>1</td>
<td>Q2</td>
<td>49</td>
<td>152</td>
<td>32.2%</td>
<td>0.5</td>
</tr>
<tr>
<td>2</td>
<td>Q1</td>
<td>86</td>
<td>327</td>
<td>26.3%</td>
<td>0.3</td>
</tr>
<tr>
<td>2</td>
<td>Q2</td>
<td>81</td>
<td>346</td>
<td>23.4%</td>
<td>0.5</td>
</tr>
<tr>
<td>3</td>
<td>Q1</td>
<td>176</td>
<td>512</td>
<td>34.4%</td>
<td>0.004</td>
</tr>
<tr>
<td>3</td>
<td>Q2</td>
<td>172</td>
<td>507</td>
<td>33.9%</td>
<td>0.007</td>
</tr>
<tr>
<td>4</td>
<td>Q1</td>
<td>103</td>
<td>331</td>
<td>31.1%</td>
<td>0.1</td>
</tr>
<tr>
<td>4</td>
<td>Q2</td>
<td>89</td>
<td>346</td>
<td>25.7%</td>
<td>0.6</td>
</tr>
</tbody>
</table>

Table 2: NTSV C-section rates (=PC-02) automatically extracted. P-values denote Z-tests that the proportions of women determined to be NTSV are consistent between manual and automatic abstraction.

**Conclusions**

Automation of core measures and other manually abstracted metrics is a major priority for hospitals across the country, as it can liberate hospital human resources for other tasks. Here we find that a rules-based automated process can indeed improve consistency, but care must be taken to ensure the accuracy of the methods and generalizability. The effort was helped by strong communication between clinicians (data enterers), abstractors (data pullers), and technical quality analysts (data analyzers) to get appropriate feedback and develop a working model. Knowledge of how to achieve this kind of effort is highly generalizable to many other data metrics, which require multiple types of input to be successfully automated.
The Impact of Simulation on Electronic Health Record Use Patterns among Pediatric Residents

Evan Orenstein, MD1, Irit Rasooly, MD2, Wanczyk Phillips BS3, Mark V. Mai, MD MHS1, Adam Dziorny, MD PhD1, Levon Utidjian, MD MBI1,2, Anthony Luberti, MD1,2, Jill Posner, MD, MSEd4, Rebecca Tenney-Soeiro, MD MSEd2, Chris Bonafide, MD MSCE1,2

1Department of Biomedical and Health Informatics; 2Department of Pediatrics; 3Information Services & 4Division of Emergency Medicine; Children’s Hospital of Philadelphia, Philadelphia, PA

Introduction:
As electronic health record (EHR) adoption has grown, training clinicians to efficiently and effectively use the EHR has become increasingly important to provide safe, high-quality patient care. While most formal EHR training occurs in classroom and lab settings,1 simulation with realistic, clinically rich test patients has improved recognition of safety concerns on subsequent simulation2,3 as well as outpatient provider self-efficacy ratings.4 However there is a paucity of data on how simulation changes EHR use in real clinical encounters. In this study, we aimed to determine if exposure to an EHR simulation that exhibited specific tools was associated with enduring changes in EHR use patterns.

Methods:
We created an informatics simulation curriculum for first-year pediatric residents at the Children’s Hospital of Philadelphia aimed at increasing their capacity to detect patient safety concerns using the EHR. Once a month on the General Pediatrics/Hematology team, 4-6 residents participated in a 1 hour teaching session focused on a simulated case in the EHR aimed at increasing awareness during emergency department (ED) to inpatient handoffs. In this case of an infant with hyperbilirubinemia, exploration in the EHR beyond the ED note revealed vital sign instability suggestive of neonatal sepsis and evidence of hemolytic anemia. During the debrief session, we emphasized the utility of specific data visualization and information retrieval tools in the EHR to detect these safety issues. While the curriculum is intended for all first-year residents, in the first 16 months of the program some first-year residents never participated because their clinical rotations did not overlap with scheduled simulation administrations.

Using EHR access logs, we assessed the frequency with which these tools were accessed by residents who did and did not participate in the simulation. We combined these data with resident schedules and simulation schedules to identify additional potential predictors including class year and weeks of inpatient experience. We restricted the analysis to inpatient weeks scheduled at the Children’s Hospital of Philadelphia. Among those who had participated in the simulation, we also examined the number of weeks post-simulation as a predictor to assess for a possible waning effect. We excluded partial weeks from holidays or class retreats. We used a mixed-effects logistic regression model with the resident as random effect in all models.

Results:
From July 1, 2015 to February 20, 2017, 59 residents participated in the neonatal sepsis simulation while 80 residents did not. We collected EHR use patterns for 3262 resident-weeks where the resident had not been exposed to the simulation and 1049 resident-weeks after simulation exposure. Residents who completed the simulation were followed for a median of 19 weeks post-simulation (IQR: 8 - 32). For a given week, if the resident had not been exposed to simulation, there was a 47% chance of having used the data visualization tool at least once and 36% chance of having used the information retrieval tool at least once. If the resident had been previously exposed, those rates increased to 73% and 85% respectively.

In univariate analysis, previous participation in the simulation increased the chance of using the data visualization tool at least once in a given week (OR 5.8, CI: 4.4 – 7.8) and of using the information retrieval tool (OR 10.5, CI: 7.5 – 14.8). After adjustment for the number of inpatient weeks completed, exposure to simulation remained a significant predictor of use of the data visualization tool (OR 2.8, CI: 2.1 - 3.9) and information retrieval tool (OR 3.0, CI: 2.0 – 4.5). The number of inpatient weeks the resident had previously completed was also associated with increased
likelihood of use of the data visualization tool (OR 1.05, CI: 1.04 - 1.06) and use of the information retrieval tool (OR 1.08, CI: 1.06 – 1.09).

After simulation exposure, the number of weeks post-simulation adjusted for total inpatient weeks completed was not a significant predictor of use of the data visualization tool (OR 0.98, CI: 0.94 - 1.02) or use of the information retrieval tool (OR 1.01, CI: 0.96 - 1.07).

**Discussion:**
Our study found that an informatics simulation was associated with changes in pediatric residents’ EHR use patterns. These changes did not wane after simulation exposure during the study period, suggesting that demonstration of new tools through well-designed simulation changes the long-term behavior of simulation participants who find the tools clinically useful. These effects may alternatively be explained by increasing awareness across the institution, although changes to institutional culture may also have been caused in part by the informatics simulations themselves. Residents who were not exposed to the simulation may also have learned about these tools through diffusion of practices from their peers, their supervisors, or annual lectures on EHR efficiency. However, diffusion of EHR practices from simulation participants to non-participants would be expected to bias the impact of simulation towards the null. Nonetheless, the strong association between simulation exposure and weekly use of these tools was robust to adjustment for class year and weeks of inpatient experience.

Simulation requires substantial effort to create and administer cases compared to other modalities such as lectures or online modules. We did not compare the impact of these simpler educational approaches on EHR use patterns. However, previous studies have demonstrated additional benefits of simulation on provider performance. Assessing the impact of simulation on patient outcomes is complex; while procedural simulations have been shown to improve resident performance in randomized trials, EHR simulations have focused on indirect measures showing improvement in recognition of simulated safety issues, novel characterization of EHR-induced errors, better measures of the quality of resident notes, and improved provider satisfaction. Our study also examines an indirect measure, but suggests that the effects of a single simulation may impact users’ long-term use patterns.

EHR simulation presents a new opportunity to improve patient safety through provider education that integrates directly into their workflow. This study suggests that simulation impacts EHR use patterns, and that those providers maintain those patterns long-term. Future efforts to compare the impact of different educational modalities on EHR use patterns and downstream effects on provider efficiency and quality of care will help define how best to incorporate simulation into EHR education.

**References:**

Improving Electronic Inpatient Progress Notes Using Voice: Results from the VGEENS Project

Thomas H. Payne, MD, J. Andrew Markiel, PhD, W. David Alonso, MD, Ross Lordon, BS, Kevin Lybarger, MS, Meliha Yetisgen, PhD, Jennifer M. Zech, MS, Andrew A. White, MD

University of Washington, Seattle, WA

Abstract

We implemented a system to create inpatient progress notes on hospital rounds (VGEENS), integrating voice recognition, automated note formatting, and EHR links. In a randomized trial we compared VGEENS with usual note writing on note timeliness, quality and physician satisfaction. Results show VGEENS notes were available within 5 minutes after dictation. Notes were on average available sooner, and physicians’ satisfaction greater in control, perhaps due to copy/paste practices. Workflow changes may improve note timeliness.

Introduction

Physician progress notes are an important record for clinical care and communication with care team members and patients. However, electronic notes are criticized for poor readability, overuse of copy and paste, and excessive note length. Physicians have voiced concerns that writing notes in EHRs takes more time than using paper or dictation; a consequence is that inpatient progress notes may not be completed and available to other team members until long after rounds.

This project is an attempt to address these problems. Here we describe the development, implementation, and evaluation of a voice-generated enhanced electronic note system, integrating voice recognition and transcription with natural language processing (NLP) and integration with the EHR, designed to match physician rounding workflow. We also present results of a randomized controlled trial to determine the effect of using this new method of writing inpatient progress notes on note timeliness, quality, and physician satisfaction, in comparison with writing notes in the usual way, through typing into partially populated templates.

Methods

VGEENS (voice-generated enhanced electronic note system), was used by study intervention physicians while on hospital rounds. At the bedside or later, the physician records a voice file on a cell phone application we (DA) developed. The completed dictation voice file is securely sent to a server where it is converted to text using automated speech recognition software (Dragon Medical Practice Edition, Nuance). Voice commands are used to break the note into sections corresponding to the preferred UW progress note format and to insert formatted patient vital signs and select laboratory results. The transcribed note is sent to the EHR Inbox.

We randomly assigned physicians on medical services of two UW teaching hospitals to the intervention group, using VGEENS, and control group, entering notes using a keyboard. We compared: 1. The time between when the patient is seen on hospital rounds and the availability of the note in the EHR; 2. Physician satisfaction with note writing and 3. Note quality as assessed by manual quality review using instrument PDQI-9.

Results
Thirty-one subjects wrote 1,852 inpatient progress notes during the study period, 1,143 by controls and 709 notes by intervention subjects. The median number of minutes between the patient encounter and the availability of a progress note in the EHR for others to view was 190 minutes for the control group and 227 minutes for the intervention group. For the subset of physicians who used VGEENS on rounds (intended workflow—circled at right), notes were available within 5 minutes.

Physician satisfaction survey response rate among the 31 of the 49 subjects who completed at least one note was 100%. Among intervention subjects, an equal number (40%) rated satisfaction with the VGEENS tool as either highly or moderately satisfied (6) and moderately dissatisfied or not at all satisfied (6). Among controls, 50% of subjects rated their satisfaction with note writing as either highly or moderately satisfied (8) and one subject (6%) was moderately dissatisfied. Note quality assessment is underway. (18 subjects were not on a medical service rotation in which their responsibilities included writing daily progress notes during the study period or for other reasons.)

Discussion

We successfully developed and implemented a new note writing method using voice to create inpatient progress notes. Where the physician used VGEENS on rounds, notes were available within 5 minutes, were properly formatted and included patient data in response to voice command. The system was integrated with a commercial EHR. Physicians preferred traditional note writing methods, in part because younger physicians are inexperienced with dictation yet facile with copying the previous day’s note, editing and saving as the current day’s note. VGEENS began each day’s note with the voice dictation and did not carry forward information such as problem list and ‘checklist’ information, though these features could be added.

Preliminary comparison of progress note content shows more preservation of text between successive days’ notes in control than with VGEENS, likely a reflection of copying/pasting workflow common in control notes. Note accuracy was not assessed, but highly similar physical examinations in successive days’ notes raise questions of accuracy. Using voice can potentially permit history and exam findings to be quickly documented reducing need to copy them from previous notes; this was a motivator for our work. We have not yet leveraged advanced NLP techniques to correct semantic errors within the note, nor to extract encoded concepts from the narrative text. Perhaps the greatest promise for this work is that we have developed a system to create notes that capture physician thinking as close to rounds as possible; we have the potential to suggest diagnostic and therapeutic interventions based on that thinking in near-real-time. The VGEENS approach has potential to directly address physician concerns with excessive documentation time requirements and declining note quality, and may also improve progress note accuracy.

Conclusion

VGEENS permits voice dictation on rounds to create progress notes and can reduce note availability and may reduce dependence on copy/paste within notes. Timing of dictation determines when notes are available; in this early trial most notes were dictated after rounds, delaying note availability. Capturing notes in near-real-time has potential to apply NLP and decision support sooner than when notes are typed later in the day, and to improve note accuracy.

Acknowledgements

This project was supported by grant number R21HS023631 from the Agency for Healthcare Research and Quality. The content is solely the responsibility of the authors and does not necessarily represent the official views of the Agency for Healthcare Research and Quality or the Department of Health and Human Services.

References

Text Mining Radiology Reports for Deep Learning Radiology Images

Yifan Peng, Ph.D. 1, Xiaosong Wang, Ph.D. 2, Le Lu, Ph.D. 2, Mohammadhadi Bagheri, M.D. 2, Ronald Summers, M.D., Ph.D. 2, Zhiyong Lu, Ph.D. 1
1National Center for Biotechnology Information, National Library of Medicine, NIH, Bethesda, MD, USA; 2Clinical Center, NIH, Bethesda, MD, USA

Introduction

Chest X-rays are a common radiological examination for screening and diagnosis of lung diseases. Although hospitals have accumulated a large number of raw radiology images and reports in Picture Archiving and Communication Systems (PACS) and their related reports in Radiology Information System (RIS), it is not yet known how to effectively use them to build high precision computer-aided diagnosis systems (CAD). In the last few years, machine-learning techniques are attracting considerable interest due to tremendous progress on medical image recognition and understanding. In particular, recent developments regarding deep-learning models (e.g., deep convolutional neural networks) have led to promising performance in the recognition of objects within medical images 1.

Generally speaking, deep learning models depend upon labeled data. That is, annotators must transfer their knowledge to the dataset in order for a model to learn the correlation between labels and images. However, it is both costly and time consuming to annotate a large-scale corpus to facilitate the data-hungry deep learning models. Thus, crowdsourcing is often used in the general domain. On the other hand, crowdsourcing is frequently infeasible for biomedical images annotation, because it requires comprehension of domain-specific medical knowledge, even if the security and privacy issues with disseminating patient information in a public platform could be adequately addressed.

In this work, we propose a text-mining approach to automatically generate a “weakly” labeled dataset for detecting thoracic diseases in radiology images. These labels can be used as a silver standard by deep learning algorithms. Using this framework, we construct a dataset, CXR-XIV, which contains over 100,000 frontal-view chest X-ray images weakly-labeled with 14 common thoracic diseases. Given such a large-scale dataset, we further propose a unified weakly-supervised multi-label image classification framework, which uses deep learning to effectively detect common thoracic diseases in radiology images.

Methods

CXR-XIV dataset. To build a corpus for deep learning radiology images, we first identified 14 common thoracic diseases/patterns that are frequently observed and diagnosed in radiology reports (e.g., pneumonia and cardiomegaly). Using a list of hand-crafted keywords for each disease, we retrieved all the radiological images and associated reports from our PACS system. We then developed a text-mining approach (described below) to filter negative and equivocal disease asserted in the reports. The resulting database, which we call CXR-XIV, is composed of 108,948 frontal-view chest X-ray images from 32,717 patients and each image/report is associated with one or more weakly labeled (i.e. computer-generated) pathology category or “normal” otherwise. To evaluate our text-mining algorithm, we also manually hand-annotated a random set of 900 radiology reports due to the high cost of human annotation.

Text-mining approach. To detect the positive findings and rule out negative/equivocal findings, our approach processes the radiology reports in two steps (Figure 1). First, we text mine all diseases mentioned in the reports using a state-of-the-art NLP tool, MetaMap 2. Next, we developed a new negation detection tool called NegBio for filtering negative and equivocal diseases in our radiology reports. Specifically, NegBio uses hand-crafted heuristic rules to search the dependency parse of each sentence in the report to determine if a disease mention is covered in its dependency graph by a negated or uncertainty cue (e.g., “no evidence of” or “suspicious”). If so, this disease will be marked as negative or equivocal and subsequently discarded.

Radiology images classification for thoracic diseases. With CXR-XIV, we then built a framework for detecting the presence of multiple thoracic diseases in radiology images (Figure 1). A weakly-supervised deep learning paradigm was adopted by utilizing the text-mined image level labels for the multi-label image classification task. Specifically, we tailored Deep Convolutional Neural Network (DCNN) architectures for multi-label disease classification.
Results

Text Mining Results. We first evaluated our method on the 900 human-annotated radiology reports: we achieved 94.4% in precision, 94.4% in recall and 94.4% in F-measure overall. These highly accurate results meet our need to generate a corpus with weak labels, which serves as a solid foundation for the later process of image classification. To demonstrate the robustness of our method, we applied NegBio to another independent dataset OpenI3. Furthermore, we compared NegBio with NegEx⁴, a rule-based tool for negation detection in clinical NLP. Our results show that NegBio achieves an improvement in precision (over 16%) and overall F1-score (over 8%) on both datasets.

Radiology image classification results. We evaluated and validated the unified DCNN framework with the CXR-XIV dataset¹. Similar to other large-scale dataset released in the computer vision community (e.g., ImageNet and MS COCO), we randomly shuffled the entire dataset into training, validation, and testing subsets. CNN models were optimized via stochastic gradient descent. In general, the quantitative performance varies for different disease patterns, in which the model based on ResNet-50 achieved the best results⁵. For example, the “cardiomegaly” (AUC=0.8141) and “pneumothorax” (AUC=0.7891) classes are consistently well-recognized, compared to other groups where the detection ratios are relatively lower for diseases of small objects, e.g., “nodule” classes (AUC=0.7164).

Discussion

To our best knowledge, this work represents the first attempt to combine text mining with radiology imaging analysis in the era of deep learning. We demonstrate the unique role of text mining for generating weak labels in the use of radiology imaging analysis by deep convolutional neural network. From the text mining perspective, a novel algorithm (NegBio) was developed with state-of-the-art performance for identifying not only negative findings, but also equivocal findings which have largely been ignored in the past.

The subsequent imaging analysis (i.e., recognizing and locating disease patterns) supports the idea that a weakly-labelled dataset can be as effective as one generated by expert annotators in training a deep learning model to detect common thoracic diseases. However, the current results also suggest that building fully-automated high precision CAD systems remains challenging. We hope that these findings will add to a growing body of literature to our understanding of reading radiology images and reports.

Acknowledgements. This work was supported by the Intramural Research Program of the National Institutes of Health, at National Library of Medicine and Clinical Center. We are also grateful to Robert Leaman for his editorial comments. We thank NVIDIA Corporation for the GPU donation.

References

Introduction

End of life care and cost continues to trouble patients, providers, and policy makers. 25% of Medicare expenditures are associated with end of life care. It has been shown that shared decision making has been linked to high quality end of life care, which could lower costs. However, Advanced Directives, which capture patient preferences, have low adoption rates, and are often completed 1-2 months prior to death. Accurately predicting mortality risk could allow providers to develop advanced care plans with their patients earlier in the course of disease. Research indicates that the presence of multiple chronic conditions, is a critical factor in health care spending; however, the trajectory of non-cancer illnesses involve exacerbations of disease followed by stabilization may make it difficult for providers to recognize when the final turn for the worse is occurring.

While literature exists on acute mortality, there are few models that address the period six months prior to death. While we identified disease specific models (e.g., CHF and Diabetes), having many disease specific models could result in obfuscation. Multiple, possibly competing mortality risk scores, would force providers to “juggle” multiple predictions. Therefore, we assessed the feasibility of a generalized mortality predictive model to assisting providers and population health organizations to identify patients six months prior to death.

Methods

Data was obtained from Medicare Shared Savings Program (MSSP) Accountable Care Organization (ACO). From eligibility data, we identified a sample of 2,307 members with: (1) a deceased date between 2014 and 2016, and (2) at least 12 months of eligibility history prior to death. We identified 52,298 members with at least 12 months of eligibility between 2014 and 2016. Deceased status was coded as a binary outcome. For the decedents, the surveillance window was the first 6 months of the 12-month period preceding the deceased date, and our prediction window was the 6-month period immediately preceding the deceased date. For the non-decedents maximum eligibility date was a proxy for deceased date. We manually curated 36 concepts mapped to four classes of features: medical claims, pharmacy claims, lab results, and socio-economic data. The selected criteria were selected based on published data on prognosis. We began with common clinical concepts (i.e., Diabetes, CHF and Charlson) and common utilization metrics such as the number of distinct diagnosis codes (Number of DXs). Our clinical expert guided features that reflect frailty, such as Skin Ulcers and difficulty with Activities of Daily Living (ADLs), and recent Body Mass Index (BMI). From that information we built concepts such as Discharge to a Skilled Nursing Facility (SNF) and procedure codes that could indicate impairments to ADLs (Functional Issues). We leveraged Median Income and Poverty rate data (Below Poverty) from the US Census Bureau. For labs we sought signal clinical indicators that can qualify an individual for hospice care: GFR, Albumin, and Creatinine. Given strong performance and ability to handle missing data without imputation, we employed Gradient Boosted Decision Trees. We leveraged Dismo in R with the parameters: folds = 10, complexity = 6, bag rate = 0.5, and learning rate = 0.01.

Results

The mean training c-statistic was 0.88. The mean validation c-statistic was 0.86. Table 1 highlights positive predictive value (PPV), sensitivity, and specificity at various risk thresholds. Figure 1 shows key influential variables.

<table>
<thead>
<tr>
<th>Percentile</th>
<th>PPV</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>99th</td>
<td>73.60%</td>
<td>17.40%</td>
<td>99.72%</td>
</tr>
<tr>
<td>95th</td>
<td>38.62%</td>
<td>45.42%</td>
<td>97.29%</td>
</tr>
<tr>
<td>90th</td>
<td>25.60%</td>
<td>60.58%</td>
<td>92.23%</td>
</tr>
</tbody>
</table>

Table 1: Model Performance

Figure 1: Key Influential Variables
Discussion

It appears feasible to predict morality risk within six months without restricting the population to a particular condition. This is significant because it could allow for a singular risk score to identify patients for end-of-life support services. As a result this model could be of interest to Population Health and providers whom a) may desire a single mortality risk prediction, b) manage an array of comorbidities, or c) manage a population with varied health needs. Our findings are consistent with other research, which found that various factors impact mortality risk. For example, laboratory results has been used previously to predict frailty and subsequent mortality. In addition, the influence of poverty on mortality aligns with literature associating income with life expectancy. While studies on mortality risk factors, specifically for Kidney Disease, have demonstrated that laboratory data are most predictive of short term mortality whereas demographic data is predictive of long term mortality, our research suggests that both socio-demographics and laboratory measurements affect six month mortality. Our data came from a single ACO, so it is questionable whether our results are generalizable to other populations. More validation is needed to determine if the model is generalizes to other populations, and if Gradient Boosted Decision Trees yields the best classifier.

Accurately predicting 6-9 month mortality may allow physicians to engage patients in advance care planning, which may facilitate better coordinated end of life care. However, accurate mortality predictions do not automatically translate into shared decisions between patients and providers, higher quality care, and lower costs. Research has indicated that policies promoting end of life discussions, expanded access to palliative care, and more education on end of life discussions are needed. Therefore, further work is needed to evaluate how end of life discussions impact both quality and cost of care in addition to assessing whether or not accurate mortality predictions facilitate end of life care discussions.

References

Linking Resident Behavior to Health Conditions in an Eldercare Monitoring System

Mihail Popescu, Andrew Craver, Lorraine Phillips, Richelle Koopman, Gregory Alexander Laurel Despins, Marilyn Rantz

University of Missouri, Columbia, MO

Introduction

Early detection of health changes among older adults is the key to maintaining health, independence, and function. Non-wearable sensors such as depth cameras, motion sensors (passive infrared, PIR) and bed sensors (based on ballistocardiography) are able to detect changes in gait activity and sleep, and have emerged as a possible solution for early detection of health changes. Since 2005, our interdisciplinary research team has investigated, developed and tested a state of the art sensor monitoring system for older adults at TigerPlace, a unique eldercare facility in Columbia, MO1.

Analyzing and acting upon sensor data remains a challenge for clinicians due to data variety (many sensor types) and velocity (continuous monitoring). To view the context of the health alerts sent by the monitoring system, clinicians currently use a secure interface to review multiple data displays that may take 7 minutes per alert. To save clinicians’ time and make alerts easier to interpret, we are investigating a new knowledge-generation methodology based on linguistic summaries as a tool to provide more meaningful and easier to interpret alerts to clinicians.

While certain nonspecific behaviors were shown to be linked to diseases in the elderly2 and non-wearable sensors can capture those behaviors3, more specific information is needed for our monitoring system. As an initial step in creating linguistic summary alerts, our team conducted a survey of clinicians to determine which signs and behaviors captured by the monitoring system they find most relevant in evaluating and treating health conditions among older adults. The results of this survey will guide the development of linguistic summary methods.

Methods

Our survey instrument consisted of a 7-item questionnaire in which clinicians were asked to rate how useful they would find a set of sensor measurements when evaluating or treating health conditions among the elderly. Respondents used a 1-4 scale to assess the usefulness of each behavior and sign: 1 = Not Useful; 2 = Somewhat Useful; 3 = Useful; 4 = Very Useful. Resident behavior measurements were computed by our monitoring system based on the non-wearable sensors shown in Table 1. For example, the number of bathroom visits was computed based on the data provided by the bathroom motion sensor, and the walking speed and stride length were computed based on depth sensor (Kinect) data (see http://eldertech.missouri.edu/papers for more details).

Table 1. Resident behavior captured by given sensors in our monitoring system

<table>
<thead>
<tr>
<th>Resident Behavior</th>
<th>Sensor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apartment activity (overall motion)</td>
<td>Motion sensors</td>
</tr>
<tr>
<td>Number of bathroom visits</td>
<td>Motion sensors</td>
</tr>
<tr>
<td>Amount of time spent in bed</td>
<td>Bed sensor</td>
</tr>
<tr>
<td>Rate of respiration while in bed</td>
<td>Bed sensor</td>
</tr>
<tr>
<td>Restlessness while in bed</td>
<td>Bed sensor</td>
</tr>
<tr>
<td>Pulse rate while in bed</td>
<td>Bed sensor</td>
</tr>
<tr>
<td>Walking speed and stride length</td>
<td>Depth sensor</td>
</tr>
</tbody>
</table>

We assessed 12 health conditions in this survey: depression, dementia, mental status change, chronic obstructive pulmonary disease (COPD) exacerbation, chronic health failure (CHF) exacerbation, atrial fibrillation, transient ischemic attack (TIA)/stroke, fall risk, hypo/hyperglycemia, urinary tract infection (UTI), pneumonia and pain.

The sample for this preliminary survey consisted of 11 physicians, 5 registered nurses and 6 licensed practical nurses. These clinicians, all users or familiar with our monitoring system, were recruited from TigerPlace, from our assisted...
living research sites, and via existing contacts in the MU Health System through referral. Clinicians took part in the survey in-person and or via an online Qualtrics (http://www.qualtrics.com) instrument. All materials and protocols were approved by the University of Missouri IRB.

Results

Three resident behaviors (signs) in order of their importance for assessing 12 health conditions are shown in table 2.

Table 2. Three resident behaviors in order of their importance for assessing 12 health conditions among the elderly

<table>
<thead>
<tr>
<th>Condition</th>
<th>Behavior (sign) 1</th>
<th>Behavior (sign) 2</th>
<th>Behavior (sign) 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Atrial Fibrillation</td>
<td>Pulse: 91%</td>
<td>Respiration: 50%</td>
<td>Restlessness: 36%</td>
</tr>
<tr>
<td>CHF Exacerbation</td>
<td>Pulse: 82%</td>
<td>Respiration: 82%</td>
<td>Restlessness: 55%</td>
</tr>
<tr>
<td>COPD Exacerbation</td>
<td>Respiration: 96%</td>
<td>Pulse: 68%</td>
<td>Restlessness: 50%</td>
</tr>
<tr>
<td>Dementia</td>
<td>Restlessness: 68%</td>
<td>Overall Motion: 59%</td>
<td>Time in Bed: 59%</td>
</tr>
<tr>
<td>Depression</td>
<td>Time in Bed: 91%</td>
<td>Overall Motion: 82%</td>
<td>Restlessness: 59%</td>
</tr>
<tr>
<td>Fall Risk</td>
<td>Stride length: 100%</td>
<td>Bathroom Motion: 82%</td>
<td>Overall Motion: 68%</td>
</tr>
<tr>
<td>Hypo/Hyperglycemia</td>
<td>Bathroom Motion: 32%</td>
<td>Restlessness: 28%</td>
<td>Overall Motion: 24%</td>
</tr>
<tr>
<td>Mental Status Change</td>
<td>Time in Bed: 82%</td>
<td>Restlessness: 77%</td>
<td>Overall Motion: 68%</td>
</tr>
<tr>
<td>Pain</td>
<td>Stride length: 82%</td>
<td>Respiration: 73%</td>
<td>Pulse: 73%</td>
</tr>
<tr>
<td>Pneumonia</td>
<td>Respiration: 86%</td>
<td>Pulse: 59%</td>
<td>Overall Motion: 41%</td>
</tr>
<tr>
<td>TIA/Stroke</td>
<td>Stride length: 55%</td>
<td>Pulse: 46%</td>
<td>Respiration: 27%</td>
</tr>
<tr>
<td>Urinary Tract Infection</td>
<td>Bathroom visits: 96%</td>
<td>Restlessness: 46%</td>
<td>Overall Motion: 46%</td>
</tr>
</tbody>
</table>

From the above table we see that for 9 out of the 12 conditions, more than 80% of the clinicians seem confident (scores 3 and 4) that our system could provide useful information. For three conditions listed above (TIA, hypo/hyperglycemia and dementia) our monitoring system doesn’t seem able to provide useful information.

Conclusions

Based on data in table 2, for residents with a history of one of the above conditions (extracted from our nursing EMR), we will provide linguistic summaries that provide context for the health alerts of the form: “Observed an increase in bathroom visits and an increase in bed restlessness. Possible UTI.” Note that the data in table 2 can be used as a fuzzy rule decision support system for producing alerts, with rules of the form: “IF the number of bathroom visits is high AND night restlessness is high and apartment motion is low THEN the possibility of UTI is high.” Our team is currently exploring the timeframe and directionality of our set of health behaviors through further clinician surveys in order to refine a system of thresholds for deploying alerts.

Acknowledgments

The National Library of Medicine grant number R01LM012221 funded this work.

References

Identifying Potentially Missing Hierarchical Relations in SNOMED CT based on Lexical Features – Impact of Synonyms and Lexico-syntactic Constraints

Satyajeet Raje, PhD, Olivier Bodenreider, MD, PhD
U.S. National Library of Medicine, National Institutes of Health, Bethesda, MD, USA

Introduction
The quality assurance of large bio-ontologies is extremely critical for their effective and continued use and is an active area of research. For example, recent investigations highlighted issues in the hierarchical structure of SNOMED CT and its detrimental effects on biomedical applications. Previous work by one of the authors established a method to identify potentially missing hierarchical relations leveraging lexical features in SNOMED CT. It used the preferred term for each concept in SNOMED CT to create logical definitions for concepts. These definitions were used to identify missing hierarchical relations. The method was tested on a subset of SNOMED CT concepts and showed limited precision (20% of false positives). In this paper, we propose two improvements on our original method: 1) by adding lexico-syntactic constraints based on shallow parsing, we expect to increase precision; 2) by processing all synonyms, we expect to increase recall. This work is a contribution to quality assurance in SNOMED CT.

Methods
We used the Sept. 2016 release of SNOMED CT (US Edition) in this study. Our methodology can be summarized as follows. We create logical definitions for concepts leveraging all concept labels (preferred terms and synonyms) provided by SNOMED CT and their lexico-syntactic analysis. We infer hierarchical relations from the logical definitions and filter out those already present in the original hierarchy of SNOMED CT. Finally, we review the potentially missing relations. We compare the results of the enhanced method to those of the original (baseline) method.

A. Creating logical definitions. In the original method, each concept in SNOMED CT is represented as a set of words using description logics (see example below in Table 1). In the enhanced method:

1) We distinguish between head nouns and modifiers through lexico-syntactic analysis using the minimal commitment parser provided by SemRep. E.g., “Photogenic epilepsy” has head “epilepsy” and has modifier “photogenic”. We then create logical definitions representing the head and modifier roles of words in the term. Complex terms are ignored. (In the definition below, “some” is the existential qualifier in the OWL syntax.)

2) We create logical definitions for each synonym of the concept in addition to the preferred term. For example, the concept “Photogenic epilepsy” has a synonym, “Television epilepsy” in addition to its preferred term, “Photogenic epilepsy”. Definitions created for terms are later mapped to the corresponding concept.

Table 1. Example logical definitions from the original and enhanced methods.

<table>
<thead>
<tr>
<th>Original</th>
<th>“Photogenic epilepsy” = disorder AND (has_word some photogenic) AND (has_word some epilepsy)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Enhanced</td>
<td>“Photogenic epilepsy” = disorder AND (has_head some epilepsy) AND (has_mod some photogenic)</td>
</tr>
<tr>
<td></td>
<td>“Television epilepsy” = disorder AND (has_head some epilepsy) AND (has_mod some television)</td>
</tr>
</tbody>
</table>

We apply this enhanced method to the same hierarchies as the original study, namely the Disorder of head (disorder) (118934005) and Operative procedure on head (procedure) (89901005).

B. Inferring and filtering hierarchical relations. We use the ELK reasoner to infer a hierarchy of terms from the logical definitions (expressed in OWL 2 EL profile). For instance, ELK infers that disorder AND (has_head some epilepsy) AND (has_mod some photogenic) is a subclass of disorder AND (has_head some epilepsy). We further derive a hierarchy of concepts from the hierarchy of terms. For example, Photogenic epilepsy (disorder) is a subclass of Epilepsy (disorder). We only keep those subclass relations (between concepts) that are not already present in SNOMED CT (transitively closed).

C. Comparing results from original and enhanced methods. We compare the performance of the original and enhanced methods. To validate the results, we manually review all potentially missing subclass relations generated by either method. The results were pooled during review in order to mask the source of the relations.

Results
A. Creating logical definitions. We created logical definitions for the 12,088 concepts of the subhierarchy rooted with the concept Disorder of head (disorder) and for the 3798 concepts from Operative procedure on head (procedure).
by the original method. With the enhanced method, we could create definitions for 15,757 terms (9687 concepts) and 3129 terms (2171 concepts), respectively.

B. Inferring and filtering hierarchical relations. A total of 612 potentially missing relations were found by the original method, and 525 by the enhanced method. Common to both methods were 225 relations, giving a combined total of 912 relations.

C. Comparing results from original and enhanced methods. Of the 912 combined relations, 657 (72%) were judged valid. Table 2 gives a breakdown of the results per method. Contrary to our expectation, the addition of lexico-syntactic constraints and synonyms does not provide the expected gain in performance. However, there is significant gain in precision (~10 points) for the 225 relations retrieved by both methods (intersection).

Table 2. Performance of the original and enhanced methods.

<table>
<thead>
<tr>
<th></th>
<th>Original (A)</th>
<th>Enhanced (B)</th>
<th>Both (A ∩ B)</th>
<th>Total (A ∪ B)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Valid</td>
<td>488</td>
<td>370</td>
<td>201</td>
<td>657</td>
</tr>
<tr>
<td>Invalid</td>
<td>124</td>
<td>155</td>
<td>24</td>
<td>255</td>
</tr>
<tr>
<td>Total relations</td>
<td>612</td>
<td>525</td>
<td>225</td>
<td>912</td>
</tr>
<tr>
<td>Precision</td>
<td>0.797</td>
<td>0.705</td>
<td>0.893</td>
<td>0.72</td>
</tr>
<tr>
<td>Recall*</td>
<td>0.742</td>
<td>0.563</td>
<td>0.304</td>
<td>-</td>
</tr>
<tr>
<td>F1 measure</td>
<td>0.769</td>
<td>0.626</td>
<td>0.456</td>
<td>-</td>
</tr>
</tbody>
</table>

*The combined 912 relations are considered as reference. Recall is based on total 657 “valid” relations in this set.

Discussion

A. Adding lexico-syntactic constraints increases precision. The gain in precision for the 225 relations retrieved by both methods (intersection) indicates that the lexico-syntactic constraints reduce false positives (adding precision). For example, Removal of fixation of mandible (procedure) is no longer recognized as subclass of Fixation of mandible (procedure), because the two terms have different head nouns. This level of precision has potential application to quality assurance. However, it also causes a significant drop in recall, because our method excludes complex syntactic patterns.

B. Enhancements degrade performance. We observed that adding synonyms created additional valid relations that cannot be inferred by the original method. For example, Enlarged parietal foramina (disorder) is recognized as subclass of Craniolacunia (disorder). However, it also adds many false positives. For example, Acquired keratoglobus (disorder) is inferred as subclass of Congenital keratoglobus (disorder), because it has a synonym “Keratoglobus”, reflecting the predominant congenitality of this condition.

Conclusion

We have evaluated the effects of adding synonyms and lexico-syntactic constraints on the identification of potentially missing hierarchical relations in a subset of SNOMED CT. In the future, we would like to apply our method to all of SNOMED CT and study the relative contribution of each enhancement (adding synonyms vs. lexico-syntactic constraints). Importantly, the missing hierarchical relations identified by our methods may only indicate underlying issues in SNOMED CT’s concept definitions, which will require domain expertise to address.

Acknowledgements

This work is supported by the Intramural Research Program of the U.S. National Library of Medicine (NLM) and an appointment to the NLM Research Participation Program administered by ORISE through an interagency agreement between the U.S Dept. of Energy and the NLM.

References

Abstract

Transitioning from hospital to home is a difficult undertaking for many patients, often putting them at risk for adverse events. We are developing a multicomponent care transition intervention (CTI) founded on the concept of a virtual nurse that corresponds with Veterans pre-discharge as an onscreen relational agent, and post-discharge via two-way automated tailored text messaging. We report the process of selecting and refining the characteristics and onscreen personality of our relational agent.

Introduction

Successfully managing complex health conditions requires patients to be engaged in their health intellectually and emotionally. Often, patient-facing eHealth technologies reach patients who are highly engaged, leaving behind those less inclined to participate meaningfully in their own care, or to effectively use technology. Strategies are needed to engage all patients more uniformly. We are examining how to empower, motivate and engage patients within the transition from hospital to home. When transitioning from hospital to home, many patients feel overwhelmed and face limited support for post-discharge self-management tasks. Clinical teams often cannot engage in proactive, tailored, post-discharge communication with patients beyond the scope of usual care. Thus, patients are often readmitted to the hospital due to factors such as limited self-monitoring, suboptimal adherence, or lack of follow-up.

Our CTI is founded on the concept of a virtual nurse to engage and support Veterans with chronic heart failure (CHF) and/or chronic obstructive pulmonary disease (COPD) in their self-management efforts. The virtual nurse first communicates with Veterans in-hospital as an onscreen relational agent, a computer-animated character that simulates conversation. Research shows the effectiveness of relational agents to provide education and counseling to patients, including discharge information. Post discharge, the virtual nurse continues to converse with Veterans via two-way, automated computer-tailored text messaging. The CTI is accompanied by a detailed safety plan including system monitoring to identify instances requiring clinical assistance. We report our process for selecting

Reference
and refining the characteristics and onscreen personality of our virtual nurse, and developing the script that drives the dialogue and interaction with Veterans.

**Methods**

We followed an iterative, multimethod process, with input from Veterans and subject matter experts. We initially developed 12 prototype virtual nurse characters based on previous work conducted in the Veteran population. To choose the physical attributes, we conducted a card-sorting exercise of photos of the characters and semi-structured interviews with six Veterans selected in accordance with the principles of maximum variation sampling (e.g., intentionally gathering data from individuals with varying characteristics). We revised the materials, removing the most disliked characters, and conducted another card-sort and interviews with an additional three Veterans. Analysis of data was descriptive and included frequencies and means. Study team members thematically analyzed the interview data using a brief code list to support deductive coding for key points of feedback.

To design the script, we consulted with experts in CTIs, health behavior, and clinical management of CHF and COPD. CTI experts organized the content of the script to reflect four important conceptual domains or “pillars” (understanding one’s condition, minding “red flags” suggesting a worsening condition, medication self-management, and keeping follow-up appointments) of care transitions. Our clinical experts tailored the script to reflect key symptoms and management for CHF and COPD. Our health behavior experts incorporated verbal and non-verbal cues, along with targeted social talk to build rapport. We role-played the script with two Veterans and solicited feedback regarding the script’s content, flow, and length.

**Results**

The virtual nurse selected had physical attributes that Veterans found realistic, inviting, and compassionate, appropriate for a clinical setting, and reflective of a nurse that they “would feel comfortable speaking with about medical matters.” Veterans also suggested that the virtual nurse should be wearing hospital scrubs and be situated in a “neutral” hospital setting like an exam room. They also emphasized that the virtual nurse’s voice should be soft, professional and without excessive intonation. The Veterans who participated in the script role-playing session reported that the script was understandable, time appropriate, educational, and engaging.

**Conclusion**

Leveraging diverse stakeholder input, we developed a key component of a multi-component CTI intended to meaningfully engage patients. Veterans expressed strong preferences for the virtual nurse attributes, and subject matter experts were instrumental to developing a script that supports clinically-appropriate and engaging interaction. Our virtual nurse CTI is unique as it embodies a personality that strives to relate to patients, making connections that otherwise may not be possible due to the heavy workloads faced by clinical team members. Our next step is to conduct a usability pilot. We will then test the effectiveness of our CTI in a three-site randomized control trial. Our paper will include preliminary results from our pilot as well as a multimedia presentation of the virtual nurse.

**Acknowledgements**

This work is supported by the U.S. Department of Veterans Affairs, HSR&D Grant #15-101. The views expressed in this article are those of the authors and do not necessarily reflect the position or policy of the Department of Veterans Affairs or the United States government.

**References**

Information Retrieval for Biomedical Datasets: The 2016 bioCADDIE Challenge

Kirk Roberts, PhD1, Anupama Gururaj, PhD1, Xiaoling Chen, PhD1, Saeid Pournejati, MS1, Trevor Cohen, MBChB, PhD1, William R. Hersh, MD2, Dina Demner-Fushman, MD, PhD3, Lucila Ohno-Machado, MD, PhD1, Hua Xu, PhD1
1 The University of Texas Health Science Center at Houston, Houston, TX, 2 Oregon Health and Science University, Portland, OR, 3 U.S. National Library of Medicine, Bethesda, MD, 4 University of California San Diego, San Diego, CA

Introduction

Biomedical research’s increasing dependence on digital data, as well as recent concerns on reusability and replicability, has led to a significant increase in the number and types of datasets available to biomedical researchers. Finding datasets relevant to one’s own project amid the massive quantity available, however, can be quite challenging due to the diversity of information types associated with a dataset. For this reason, new methods of information retrieval (IR) are necessary. To encourage the rapid development of novel dataset retrieval methods, a publicly available test set was created to provide a benchmark dataset on which IR researchers can compare their methods. The release of this dataset was accompanied by a shared task—the 2016 bioCADDIE Dataset Retrieval Challenge—to spur rapid development and dissemination of ideas for biomedical dataset IR.

The IR challenges in searching biomedical datasets are numerous and complex. Dataset searches are not typically answerable entirely from the structured meta-data (e.g., organism, array type) associated with the datasets. For example, a researcher might be interested in “genome data for IDH1 and IDH2 in humans with glioma”. Answering such a query requires innovative search strategies not only for the structured and unstructured dataset description, but often the linked scientific articles to better place the dataset’s qualities in perspective. Due to this complexity and the many potential IR solutions, evaluating dataset retrieval methods on a common benchmark is vital.

Background

The biomedical and healthCAre Data Discovery Index Ecosystem (bioCADDIE) project1,2 seeks to provide a prototype platform for researchers to find, reanalyze, and revise biomedical data. bioCADDIE is designed to be a common data index, connecting with existing biomedical data repositories (e.g., dbGaP, ClinicalTrials.gov). The meta-data from these sources are collected, normalized, and indexed. The bioCADDIE search engine–DataMed3–utilizes the complex and varied meta-data describing each dataset. Currently, DataMed is a relatively baseline IR system, indexing structured and unstructured data, then performing standard IR search mechanisms. An immediate benefit of the Dataset Retrieval Challenge, then, is to identify innovative search strategies for integration into the DataMed search engine.

Beyond the scope of bioCADDIE and dataset retrieval, IR shared tasks have been enormously successful in fostering innovative methods and encouraging collaboration between IR researchers and biomedical experts. These shared tasks have largely been organized as part of the annual Text Retrieval Conference (TREC), organized by the U.S. National Institute of Standards and Technology (NIST). From 2003-2007, the TREC Genomics track4 focused largely on retrieving scientific articles of interest to genomics researchers. From 2011-2012, the TREC Medical Records track5 switched the focus to retrieving clinical notes. Finally, from 2014-2016, the TREC Clinical Decision Support track6 focused on retrieving scientific articles of interest to clinicians. All these tasks garnered significant interest and participation, well above that of most TREC tracks. The bioCADDIE Dataset Retrieval Challenge most resembles the original TREC Genomics track in its focus on researchers as the primary user. However, bioCADDIE’s focus on datasets and not literature articles is a substantial shift. Not only do datasets have their own meta-data and descriptions, they often are linked to multiple scientific articles. The focus of those articles is generally the scientific findings, as opposed to the potential reusability of the data. This presents a dataset search engine with three very different sources of information: (i) structured data providing basic constraints of the dataset, (ii) textual descriptions providing a high-level overview of the data, and (iii) scientific articles that generally are not focused on the dataset itself, but rather illustrate potential uses of the dataset. This combination of challenges has not been addressed by any previous IR shared task.
Materials and Methods

The bioCADDIE Dataset Retrieval Challenge participants were provided a snapshot of the bioCADDIE index comprising 20 repositories in XML and JSON formats. They were also provided with 6 sample queries with manual judgements and 30 sample queries without judgement. Two weeks before the submission deadline, participants were provided with 15 test queries, for which they were allowed to return 1,000 results each. Up to five submissions were allowed per participant.

Prior to the submission deadline, four baseline IR systems were used to identify 18,416 results for manual judgement. After the deadline, the participant submissions were pooled to identify a further 1,767 results for manual judgements, similar to the TREC approach for creation of judgement pools. Each judgement determines whether a given dataset is relevant, partially relevant, or not relevant to a particular query.

Results

A total of 45 submissions were received from 10 participants. The results are shown in Table 1.

<table>
<thead>
<tr>
<th>Participant</th>
<th>infAP</th>
<th>infNDCG</th>
<th>P@10 (+P)</th>
<th>P@10 (-P)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BioMed</td>
<td>0.2568</td>
<td>0.4017</td>
<td>0.7733</td>
<td>0.3333</td>
</tr>
<tr>
<td>Elsevier</td>
<td>0.3283</td>
<td>0.4368</td>
<td>0.8267</td>
<td>0.4267</td>
</tr>
<tr>
<td>Emory</td>
<td>0.2818</td>
<td>0.4241</td>
<td>0.7200</td>
<td>0.2667</td>
</tr>
<tr>
<td>HITSC-IRCC</td>
<td>0.2576</td>
<td>0.3850</td>
<td>0.7000</td>
<td>0.2800</td>
</tr>
<tr>
<td>IAI PUT</td>
<td>0.0876</td>
<td>0.3580</td>
<td>0.5333</td>
<td>0.1600</td>
</tr>
<tr>
<td>Mayo Clinic</td>
<td>0.1628</td>
<td>0.3933</td>
<td>0.7467</td>
<td>0.2600</td>
</tr>
<tr>
<td>OHSU</td>
<td>0.3193</td>
<td>0.4454</td>
<td>0.7600</td>
<td>0.3333</td>
</tr>
<tr>
<td>SIBTex</td>
<td>0.3664</td>
<td>0.4258</td>
<td>0.7533</td>
<td>0.3467</td>
</tr>
<tr>
<td>UCSD</td>
<td>0.2901</td>
<td>0.5132</td>
<td>0.7600</td>
<td>0.3333</td>
</tr>
<tr>
<td>UIUC GSIS</td>
<td>0.3228</td>
<td>0.4502</td>
<td>0.7133</td>
<td>0.2867</td>
</tr>
</tbody>
</table>

Table 1: Best participant run on each metric. infAP: inferred average precision; infNDCG: inferred normalized discounted cumulative gain, P@10: precision of top 10 results (+/-P: whether partial results are considered relevant). See Yilmaz et al. for more information on inferred measures.

Discussion and Conclusion

To the degree that the bioCADDIE Challenge elicited new and innovative approaches for dataset retrieval beyond those already employed by DataMed, the task was quite successful. Participants utilized a highly diverse set of IR methods for retrieving datasets. These methods ranged from rule-based query processing for identifying specific query elements for special handling (e.g., identifying the organism of interest for use against the corresponding structured metadata field) to machine learning-based ranking frameworks incorporating dozens of features. As a publicly available benchmark dataset, the bioCADDIE Challenge data opens future work by IR researchers in many directions. Further, the bioCADDIE team plans to incorporate many of the approaches used by the participants into the DataMed search engine and further analyze the specific sub-problems related to dataset retrieval.

Acknowledgements This project was supported by the National Institutes of Health grant U24AI117966.

References

Association of BMI and Obesity Genetic Risk Score with Surgical Procedures Through a Procedure-wide Association Study

Jamie R. Robinson, M.D., Zongyang Mou, B.S., Lisa Bastarache, B.S., Wei-Qi Wei, M.D., Ph.D., Robert J. Carroll, Ph.D., Joshua C. Denny, M.D., M.S.
Vanderbilt University Medical Center, Nashville, TN

Introduction
Body mass index (BMI), both below and above the optimum, is known to be a strong predictor of comorbidities and overall mortality. We sought to analyze the association of BMI and known obesity genetic risk loci via Mendelian randomization with comorbidities and invasive procedures via phenome-wide association studies (PheWAS) and procedure-wide association studies (ProcedureWAS). We aimed to demonstrate the utility of ProcedureWAS in validating, as well as identifying novel, phenotypic and genetic associations.

Methods
A retrospective, cross-sectional study of all adult individuals with documented BMI at Vanderbilt University Medical Center (VUMC) was performed. Data collection was accomplished using the VUMC Synthetic Derivative (SD), a de-identified version of over 2.4 million patient electronic health records. After removal of outliers and BMIs recorded during pregnancy, median BMIs for each individual were categorized based upon world health organization (WHO) classifications. PheWAS, logistic regression of aggregated International Classification of Diseases (ICD) 9-CM codes against BMI categories, was performed to determine phenotypic associations. Similarly, current procedural terminology (CPT) codes were aggregated into clinically meaningful categories as defined by the Healthcare Cost and Utilization Project of the Agency for Healthcare Research and Quality. ProcedureWAS, logistic regression models of procedure aggregates against BMI categories, was performed. Replication analysis was performed with a separate ProcedureWAS with logistic regression models of procedure categories against a genetic risk score (GRS) for obesity constructed as a sum of risk alleles from 37 single nucleotide polymorphisms (SNPs) with known relation to obesity. The GRS was calculated for adults of European ancestry genotyped on the Illumina HumanExome array platform. All analyses were adjusted for age and gender.

Results
A total of 737,066 individuals had at least one recording of BMI for inclusion in the PheWAS and ProcedureWAS against BMI. Of those, 24,363 individuals had genetic data for inclusion in the ProcedureWAS against the obesity GRS. In PheWAS, WHO categories of underweight (<18.5 kg/m²) and obesity class 3 (≥40 kg/m²) individuals showed significant increases in many comorbidities compared to those with normal BMI (Figures 1-2).

Figure 1. PheWAS of WHO underweight BMI with comorbidities. Red line is Bonferroni significance.
In ProcedureWAS (Table 1), underweight individuals showed significant increases in gastrointestinal (GI), respiratory, and non-cardiac vascular procedures compared to those with normal BMI. In contrast, class 3 obesity was significantly associated with coronary intervention, liver biopsy, bariatric and orthopedic surgery, cesarean section, cholecystectomy, and hernia repair. Coronary arteriography, liver biopsy, gastric bypass, knee arthroplasty, and hernia repair associations were replicated in ProcedureWAS using the obesity GRS (noted by an asterisk in Table 1).

Table 1. ProcedureWAS top associations of WHO underweight and class 3 obesity with invasive procedures. Significance determined by Bonferroni corrected p-value of $5.7 \times 10^{-6}$.

<table>
<thead>
<tr>
<th>Procedure</th>
<th>OR</th>
<th>P-value</th>
<th>Procedure</th>
<th>OR</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gastrostomy</td>
<td>5.00</td>
<td>2.2x10^{-10}</td>
<td>Gastric bypass*</td>
<td>156.96</td>
<td>7.3x10^{-195}</td>
</tr>
<tr>
<td>Vascular catheterization</td>
<td>1.66</td>
<td>1.2x10^{-43}</td>
<td>Arthroplasty knee*</td>
<td>7.84</td>
<td>5.6x10^{-136}</td>
</tr>
<tr>
<td>Upper GI endoscopy, biopsy</td>
<td>1.67</td>
<td>8.5x10^{-44}</td>
<td>Coronary arteriography*</td>
<td>2.26</td>
<td>8.2x10^{-136}</td>
</tr>
<tr>
<td>Upper GI therapeutic procedure</td>
<td>2.99</td>
<td>2.4x10^{-22}</td>
<td>Biopsy of liver*</td>
<td>3.89</td>
<td>5.3x10^{-130}</td>
</tr>
<tr>
<td>Ileostomy and other enterostomy</td>
<td>2.66</td>
<td>1.1x10^{-16}</td>
<td>Arthrocentesis</td>
<td>1.98</td>
<td>3.3x10^{-120}</td>
</tr>
<tr>
<td>Tracheostomy</td>
<td>2.29</td>
<td>1.7x10^{-14}</td>
<td>Other hernia repair*</td>
<td>3.82</td>
<td>3.9x10^{-90}</td>
</tr>
<tr>
<td>Procedure on respiratory system</td>
<td>1.97</td>
<td>3.3x10^{-11}</td>
<td>Cesarean section</td>
<td>2.83</td>
<td>1.8x10^{-77}</td>
</tr>
<tr>
<td>Thoracentesis</td>
<td>1.88</td>
<td>1.7x10^{-10}</td>
<td>Cholecystectomy</td>
<td>2.90</td>
<td>8.1x10^{-64}</td>
</tr>
<tr>
<td>Treatment of hip/femur fracture</td>
<td>1.82</td>
<td>8.4x10^{-8}</td>
<td>Coronary angioplasty*</td>
<td>2.08</td>
<td>2.4x10^{-33}</td>
</tr>
</tbody>
</table>

*Replicated (p<0.05) in separate analysis of BMI categories against obesity GRS with Mendelian randomization

Conclusion
This study demonstrates that BMI, both underweight and overweight, are strongly correlated with increased risk for comorbidities and undergoing invasive procedures. This is also the first known study to show that procedure-wide association studies using aggregated procedure codes and pairwise analyses can be used to analyze associations with clinical phenotypes and genetic data, quantifying the magnitude of increased probability due to a specific risk factor.

References
Deep recurrent neural networks identify transgender patients

Joseph D. Romano¹, Kenrick D. Cato¹,², Rami Vanguri¹, Nicholas P. Tatonetti¹
¹Columbia University, New York, New York; ²NewYork-Presbyterian Hospital, New York, New York

Abstract

We demonstrate a state-of-the-art deep learning approach to classifying patients by transgender status. Our model consists of performing word-embedding on clinical free-text notes, followed by a long short-term memory recurrent neural network classifier. This study represents a machine learning approach for identifying transgender patients, and may prove useful in improving clinical care for transgender patients.

Introduction

Transgender individuals have unique health concerns that should be considered during clinical encounters¹. However, due to stigmas and other social factors, patients may not inform their health care providers of their transgender status. All previous methods for note-based identification of transgender patients are based on simple heuristic searches (e.g., via substring matching or billing/diagnosis codes) that lack sufficient rigor. In response to these and other observations, LGBT researchers and advocacy groups are seeking methods for informing care providers of potential transgender status in a scalable manner.

Recurrent neural networks (RNNs) provide an architecture for performing deep learning on ordered sequences², such as natural language. Currently, RNNs are a state-of-the-art approach to performing text classification, and have seen success in many areas, including biomedicine³. The performance of RNN text classifiers can be further improved using word embedding to convert words to high-dimensional vectors, where the relationships between word vectors represent the natural semantic relationships between their corresponding words.

In this study we demonstrate a novel method for classifying transgender patients using word embedding and an RNN model applied to clinical notes.

Methods

We manually identified an initial cohort of 39 transgender patients—using a string search for terms and pronouns indicative of transgender status —in all clinical free-text notes within the Columbia University Medical Center EHR.
system. We subsequently defined a control cohort as 500 randomly selected patients in the EHR that are not in the case group. For each of these two groups, we extracted all available clinical notes (inpatient and outpatient) and constructed training, testing, and hold-out datasets (via random sampling without replacement of 70%, 20%, and 10% of patients, respectively), where notes were labeled according to the case/control status of the patients to whom they corresponded. An overview of the analysis pipeline is shown in Figure 1.

The classification model we used consists of two components: (1) Learning a high-dimensional embedding for the words using the skip-gram model⁴, and (2) training a dynamic RNN to predict whether a note is from a control or a case patient. The RNN was trained using stochastic gradient descent and a cross-entropy loss function, and monitored for overfitting using the aforementioned holdout validation data set. We implemented the model in the Tensorflow numerical computation library using a long short-term memory architecture⁵ for the RNN, and performed training and evaluation on a CUDA-accelerated Linux server with an Nvidia Tesla k40 supercomputing GPU.

Results

We extracted 4,859 clinical notes for the 39 transgender patients and 13,314 notes for the 500 control patients, each of which we then divided as described above into training, testing, and hold-out sets. The trained model performed surprisingly well, yielding 83.3% classification accuracy, with 87% precision and 93% recall ($F_1$ score of 90.2%). These results are summarized in Table 1.

Learning the word embeddings for the corpus of 18,173 notes completed within 5 minutes, and training the RNN completed within 10 minutes, which is approximately one order of magnitude faster than when tested on a desktop computer without GPU acceleration.

Table 1: Performance of embedding + RNN model trained on transgender notes.

<table>
<thead>
<tr>
<th>Performance measure</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classification accuracy</td>
<td>83.3</td>
</tr>
<tr>
<td>Precision</td>
<td>87.0</td>
</tr>
<tr>
<td>Recall</td>
<td>93.0</td>
</tr>
<tr>
<td>$F_1$ score</td>
<td>90.2</td>
</tr>
</tbody>
</table>

Discussion

This algorithm can form a basis of cohort discovery for future studies regarding the health of transgender patients. One of the primary benefits of our classification model is that it performs well even with a small number of case patients. This classifier holds the potential to be useful in clinical decision support (such as alerting physicians to watch out for certain transgender-related health issues). However, it does not address the ethical issues that accompany the automated identification of transgender patients⁶, and is therefore not yet ready for use in clinical practice. We encourage further investigation in the hopes that care providers can better serve the unique health needs of transgender patients.

References

3. Lipton Z et al. Modeling Missing Data in Clinical Time Series with RNNs. PMLR. 2006;56:253-270.
Development and Validation of a Continuously Age-Adjusted Measure of Patient Condition for Hospitalized Children Using the Electronic Medical Record

Michael J. Rothman, PhD\textsuperscript{a}; Joseph J. Tepas, III, MD, FACS, FAAP\textsuperscript{b}; Andrew J. Nowalk, MD, PhD\textsuperscript{c}; James E. Levin, MD, PhD\textsuperscript{d}; Joan M. Rimar, RN, DNSc\textsuperscript{e}; Albert Marchetti, MD\textsuperscript{d}; and Allen L. Hsiao, MD, FAAP\textsuperscript{d,f}

Affiliations: \textsuperscript{a}PeraHealth, Inc., Charlotte, NC; \textsuperscript{b}University of Florida, Jacksonville, FL; \textsuperscript{c}University of Pittsburgh School of Medicine, Pittsburgh, PA; \textsuperscript{d}Yale New Haven Health System, New Haven, CT; \textsuperscript{e}Rutgers New Jersey Medical School, Newark, NJ; \textsuperscript{f}Yale School of Medicine, New Haven, CT

Introduction
Awareness of a hospitalized patient’s physiological status and the context of that status are keys to the effective delivery of quality care. For adults, the Rothman Index (RI)\textsuperscript{1}, a validated measure of patient condition, based on empirically derived univariate relationships between 1-year post-discharge mortality and each of 26 clinical measurements, which are available in the electronic medical record, assists clinicians by providing an estimate of physiological status and its trend over time. For pediatrics, there are no data from which a similar measure of patient condition may be derived. We report on the development and validation of a new methodology to leverage adult mortality data to generate estimates of risk and of patient condition for pediatrics, where the relationships between risk and several physiological parameters are continuous functions of the age of the patient.

Methods
Clinical data were extracted from electronic medical records (EMRs) at three pediatric hospitals: Children’s Hospital of Pittsburgh, Yale-New Haven Hospital and a 3\textsuperscript{rd} in the northeast, covering 105,470 inpatient visits over a 3-year period (2009-2013). IRB approvals were granted.

The RI input variable set was used as a starting point for the development of the pediatric Rothman Index (pRI). Age-dependence of continuous variables was determined graphically, by plotting mean values versus age. For variables determined to be age-dependent, polynomial functions of mean value and mean standard deviation versus age were constructed using data from two hospitals. Mean values and standard deviations for previously published corresponding adult risk curves, were separately estimated. Based on the “find the center of the channel” hypothesis (mean value equals lowest risk), univariate pediatric risk was then computed by applying a z-score transform to adult mean and standard deviation values based on polynomial pediatric mean and standard deviation functions. Multivariate pediatric risk is estimated as the sum of univariate risk. Other age adjustments for categorical variables were also employed.

Age-specific pediatric excess risk functions were compared to age-specific expert-derived functions and to in-hospital mortality. Odds ratios for in-hospital mortality were computed for each of eleven nursing assessments, which are part of the pRI and constitute a “head-to-toe” assessment which is a standard part of nursing protocol. AUC for 24-hour mortality and pRI scores prior to unplanned ICU transfers were computed, as were minimum visit pRIs.

Results
Age-adjusted risk functions correlated well with similar functions in two pediatric risk models, Bedside PEWS and PAWS\textsuperscript{2,3}(Fig. 1). Pediatric nursing assessment data correlated well with risk as measured by mortality odds ratios. AUC for pRI for 24-hour mortality was 0.93 (0.92, 0.94), 0.93 (0.93, 0.93) and 0.95 (0.95, 0.95) at the three pediatric hospitals. Average pRI scores
declined prior to unplanned transfers to the ICU (Fig. 1). Moreover, the minimum pRI score for the visit correlated with likelihood of unplanned transfer. A case study is shown (Fig. 1).

Discussion
We have demonstrated that each component of pRI, which is a heuristic model, is credible. Overall model performance as measured by AUC, a standard metric for acuity tools, is excellent, with similar results at each hospital. On a functional basis, the pRI shows potential as an early warning tool, declining prior to unplanned transfers to the ICU. While there are occurrences of sudden decompensation, often physiological deterioration resulting from a surgical or other treatment complication is a process which progresses over hours or days. Intervention to prevent an emergent condition is most effective if done early in that process. The pRI, which is integrated into major EMRs, provides a real-time estimate of patient condition and is presented to the clinician as a graph (see Fig. 1). Viewing an array of pRI graphs allows the clinician to survey many patients, allowing him/her to influence and prioritize the management of those patients. It is currently being used at pediatric hospitals for that purpose4, as well as for hospital-wide surveillance, and will be considered for telehealth applications.

REFERENCES

Figure 1 - Risk vs. heart rate at 6 months, 3, 8 and 15 years, compared to PEWS and PAWS. pRI vs. hours prior to unplanned transfer to ICU. 24-hour mortality vs. pRI at 2 hospitals. A 5-day pRI graph for a 2-week-old with congenital heart disease: A. stable, B. deterioration followed by ECMO, C. stabilization, D. improvement, E. septic shock.
Can Emergency Department Provider Notes Help Achieve More Dynamic Clinical Decision Support?

Justin F. Rousseau, MD, MMSc1,2, Ivan K. Ip, MD, MPH1, Ali S. Raja, MD, MBA, MPH1, Jeremiah D. Schuur, MD, MHS1, Ramin Khorasani, MD, MPH1

1Brigham and Women’s Hospital, Harvard Medical School, Boston, MA; 2Dell Medical School at The University of Texas at Austin, Austin, TX

Abstract
Retrospective observational data were reviewed from notes present prior to image ordering from 438 Emergency Department encounters with chief complaints of trauma, falls, and bicycle accidents and cervical-spine (C-spine) imaging. C-spine clinical decision support (CDS) was delivered in 26.7% of encounters. 59.2% of encounters with a note portion present prior to image ordering contained at least one CDS criterion. Concordance was lacking between polarity of attributes when present in both notes and CDS tool.

Introduction
Imaging clinical decision support (CDS) is an application of health information technology to inform clinical decision-making at the point of care regarding the need for imaging or the optimal diagnostic study based on the best available evidence1. Although clinical decision support (CDS) can be an effective health information technology tool to deliver timely advice at the bedside, many opportunities remain to improve CDS, including timing, sensitivity, and specificity of its delivery. One of the Ten Commandments of imaging CDS is to “respect ordering provider workflow,” and in particular to eliminate redundant data entry1. Currently, when imaging CDS is integrated within a computerized physician order entry (CPOE) system, clinicians enter patient data for CDS at the time of order entry in addition to and independent of data entered in the order entry module, leading to significant interruption of workflow as well as redundancy of documentation likely previously entered in the clinical note within the same electronic health record (EHR). Incomplete or conflicting information resulting from these redundancies may adversely impact quality of patient care, as well as the secondary use of data for purposes such as analyses of guideline adherence or research. Low sensitivity of CDS delivery may lead to missed opportunities to suggest appropriate investigations that might lead to the appropriate diagnosis. Further, CDS often exposes providers to clinical alerts with low specificity, resulting in alert fatigue2.

Cervical spine (C-spine) imaging can be guided by CDS that requires the entry of certain relevant attributes of the patient history and physical examination in typical EHR implementations. Our objective was to assess the frequency of CDS attributes present in emergency department (ED) provider notes at time of C-spine order entry to determine whether these might be a source of data which could be automatically pulled from the notes rather than entered again by the providers in the order entry process.

Methods
This Institutional Review Board-approved, retrospective study was performed in an adult quaternary academic hospital. We included all adult ED patients seen between 4/1/2013 and 9/30/2014 after falls, trauma, or bicycle accidents who underwent C-spine CT or x-ray. We excluded encounters where the trauma team was activated as these encounters involve immediate performance of trauma imaging by protocol, often before orders are placed, making CDS to inform the appropriateness of the study unnecessary. The NEXUS CDS rule identifies a set of five clinical criteria attributes which, when all are absent, indicates that the patient has a very low probability of cervical injury3. At the study institution, CDS based on NEXUS is delivered when a C-spine CT or 3-view plain film study is ordered and the structured indication of “trauma” is selected on the order requisition. CDS based on the Canadian C-Spine rules is delivered when a standard 6-view C-spine plain film study is ordered and the structured indication of “trauma” is selected on the order requisition4.

The primary outcome was the rate of CDS rule attributes identified in the clinical notes at the time of image order entry. Secondary outcome measures included rates of CDS rule delivery and rates of exclusion criteria from the CDS rules identified in note portion versions available at the time of image order entry. We examined concordance in CDS rule attributes identified in the clinical notes available at the time of image order entry compared to the attributes obtained via physician interaction with the CDS rule through the CPOE system using kappa statistic of agreement and
McNemar test of marginal homogeneity of the paired data. As attributes are present in the CDS tool in all encounters where CDS was delivered, concordance was calculated for individual attributes when CDS was delivered and an attribute was described as positive or negated in the note.

Results

We identified 3155 consecutive encounters for 2992 unique patients who presented with a chief complaint of fall, trauma, or bicycle accident where the trauma team was not activated during the 18-month study period. C-spine imaging (CT or plain film) was performed in 438/3155 (13.9%) of these encounters. A portion of the clinical note was submitted before imaging was ordered in 42% (184/438) of encounters. C-spine CDS was delivered in only 26.7% (117/438) of encounters. 59.2% (109/184) encounters with note portions submitted before image ordering had at least one positive CDS criterion identified, 34.8% (64/184) identified CDS exclusion criteria. Overall concordance of C-spine CDS attributes when present in both notes and CDS was 68.4% ($\kappa=0.35$; McNemar $p=0.23$).

Discussion

Information found in unstructured clinical notes present at the time of imaging order entry contained significant amounts of relevant CDS rule attributes and exclusion criteria that should have been entered by the physician into the CDS tool in the CPOE. Harvesting these CDS rule attributes and exclusion criteria from the EHR notes when available may 1) allow pre-population of CDS tools, freeing clinicians from unnecessary redundant data entry, and 2) allow suppression of CDS rules not applicable to the current patient, freeing clinicians from unnecessary visual and workflow interruptions, thus reducing alert fatigue. Our study was conducted in a single academic setting making its generalizability unclear, particularly where there is individual and institutional variation in diagnostic ordering practices and management of trauma cases. However, by bringing the best available evidence to the point of decision-making, a goal of CDS is to reduce such variability.

In this study, rates of delivery of the CDS tool when indicated were surprisingly low, suggesting that using data entered to a CDS tool alone would be insufficient to define a population of patients eligible for the rule. Finally, there was only fair agreement between CDS attributes documented in notes compared to those entered in the CDS tool in the same encounter, raising questions about the quality, sufficiency, veracity, and completeness of data entered into either source.

Conclusion

Clinician documentation in the EHR is an underutilized information resource that contains CDS attributes and exclusion criteria and is available in a significant percentage of encounters at the point of C-spine image ordering. This is a surprising finding in cases of C-spine imaging for trauma in the ED that do not activate the trauma protocol for immediate imaging, where a scarcity of documentation might be expected prior to image ordering. Future work will need to be done to evaluate the quality and quantity of clinical data documented in encounters that are less acute than cases of trauma in the ED. Future work will also need to be done in the area of text mining or natural language processing in order to automate the extraction of these attributes and exclusion criteria to realize a feasible application integrated into the EHR, allowing for pre-population of CDS and less data entry by providers.

References


Using mHealth to Monitor Asthma Symptoms Between Visits

Robert S. Rudin, PhD,1 Christopher H. Fanta, MD,2 Zachary Predmore, BA,1 Eyal Zimlichman, MD,3 Kevin W. Kron, BA,3 Maria Edelen, PhD,1 David W. Bates, MD2

1 RAND Corporation, Boston, MA; 2 Partners Healthcare, Boston, MA; 3 Sheba Medical Center, Israel

Introduction

Asthma affects more than 25 million individuals in the United States and imposes for many a burden of frequent respiratory symptoms and risk of asthmatic exacerbations. While research suggests that asthma is controllable if treatment is adjusted based on frequent monitoring of patients’ symptoms, such intensive symptom monitoring does not routinely happen. Previous research including 12 RCTs has involved complex interventions with multiple components, limited end user input, and inconsistent results. A recent review calls for better understanding of the contribution of intervention components. To begin to address these limitations, we developed a “minimum viable” asthma symptom monitoring smartphone application and clinical workflows needed to implement it in routine care. Our findings include preliminary critical components for implementing asthma symptom monitoring.

Methods

We combined principles of qualitative research, user-centered design, and “gamification” (i.e. game design elements) to iteratively understand user needs, develop and refine key use cases, develop prototypes, and create a fully-functional app and clinical workflows. We conducted 19 individual interviews or design sessions with patients (n=9, ages 21-74, 6 female, 2 African Americans, 1 Latino, 2 low-income, 2 section-8 housing residents) and clinicians (5 physicians, 2 nurses) practicing at an academic medical center. All patients were daily smartphone users. We conducted multiple design sessions with some patients.

Results

With only a modest sample of users, we achieved saturation of critical intervention components that met the stated needs of users, consistent with the goal of designing a “minimal viable” intervention. The intervention consists of 4 interrelated use cases:

1) Invitation. Patients are invited to use the app by their physicians, which is important for patient motivation.
2) Weekly symptom checks and notifications. Each week, patients receive a 5-item symptom questionnaire via the app. If symptoms are somewhat worse than the previous week or baseline, they are given the option to request a call from a nurse. For much worse symptoms, a request is sent automatically. If they complete several questionnaires consecutively, a smiley face icon appears, which is a surprisingly strong motivator for patients to use the app.
3) Patient review of symptoms. In the app, patients can view their self-reported symptoms in graphical format.
4) In-person visit. Physicians have access to their patients’ reported symptoms in the EHR for discussion during in-person visits.

Benefits to patients may include: helping decide when to call their provider; facilitating discussion and shared decision making; and reducing hospitalizations. Their only burden is the 5-item weekly questionnaire. Benefits to providers may include saving time discussing symptoms, and improved satisfaction. Provider organizations may need to pay nurses extra, but those costs may be offset by reduced visits and hospitalizations (in accountable care).

Discussion

Novel tools and clinical workflows will be necessary for providers to sustainably implement clinical guidelines and succeed under alternate risk-based payment methods. Combining qualitative and user-centered design methods to develop a minimum viable intervention can reveal critical components.

Conclusion

With a modest sample, we achieved saturation of user requirements and design for a minimal mHealth intervention that would facilitate asthma symptom monitoring between visits. Pilot testing is needed to further assess feasibility and impact. Such tools may be necessary for providers to succeed under emerging value-based payment models and may have a secondary benefit of producing large volumes of reliable patient-reported outcome data for research. [This work was supported by the Agency for Healthcare Research and Quality grant #1R21HS023960.]
Rationale and Design for the Duke Connected Care Predictive Modeling Pilot with a Medicare Shared Savings Program Population

Shelley A. Rusincovitch1, Ricardo Henao, PhD2, Michael Gao3, Lawrence Carin, PhD2, Ursula Rogers1,4, N. Benjamin Neely, MS5, Mary Schilder, MA4, Daniel Costello, MPA6, Eugenie Komives, MD6, Erich S. Huang, MD, PhD5,6

1Duke Clinical and Translational Science Institute, Durham, NC; 2Duke University, Durham, NC; 3Duke Institute for Health Innovation, Durham, NC; 4Duke Health Technology Solutions, Durham, NC; 5Duke Office of Research Informatics, Durham, NC; 6Duke Connected Care, Durham, NC; 7Duke Department of Biostatistics and Bioinformatics, Durham, NC

Introduction
Predictive modeling offers the opportunity to assess, manage, and intervene on risk in Accountable Care Organizations (ACOs)1-2. Although the deployment of machine learning and other predictive modeling tools has become well-established in many industries, these methods have only recently become more common in healthcare settings.

In this abstract, we discuss the rationale and design for a machine learning pilot undertaken by Duke Connected Care, an ACO participating in the Medicare Shared Savings Program (MSSP)3, in conjunction with a data science team of informaticists and machine learning experts. The result of the pilot is expected to be twofold: (1) development of a clinically-informed and actionable predictive model that can be assessed for potential implementation within Duke Health, (2) synthesis of experiences arising from an ACO’s engagement and collaboration with machine learning expertise.

Methods
The pilot targets inpatient admissions, complimenting earlier work by Duke Connected Care (DCC) in predicting the risk of readmissions. Predicting this initial hospitalization is broadly applicable to the MSSP population, and yields opportunities for intervention and prevention with the goal of achieving higher quality of care with lower cost.

This operational project is undertaken for patient care and its operational scope is carefully framed within appropriate data use parameters. The computational environment that we developed for the pilot is governed by mechanisms that include an honest broker function to manage secure transport of datasets, individual user agreements, and processes for access control and permissions. Links to external knowledge bases include application program interface (API) calls to the National Library of Medicine RxNorm database4. API calls never involve protected health information nor other patient attributes.

One of the key criteria used to assess the output of the machine learning model is actionability. Both the quantitative team and the larger cross-disciplinary stakeholder group will evaluate the ability of machine learning to identify predictive factors that contribute to higher risk of admissions, particularly those that care coordinators and providers can act upon to mitigate risk.

Results
Our machine learning methodology has foundations in earlier work with Electronic Health Record (EHR) data and is an extension of Deep Poisson Factor Analysis5. We have applied these methods to a new data source: claims data for the MSSP population served by Duke Connected Care. These data are provided by the Centers for Medicare & Medicare Services (CMS) in the ACO Operational System (ACO-OS) Claim and Claim Line Feed (CCLF) file format. These data are transactional and differ notably from the typical structure of research data files available through the CMS Research Data Assistance Center (ResDAC). The ACO-OS CCLF files are received monthly, and the population included in these files varies over time based on complex inclusion criteria and attribution.

A key component for the design of the pilot has been domain modeling to describe the source data and generating high-level concepts applicable across both claims and EHR data sources.
Table 1. A partial list of data domains modeled for feature extraction and consumed by the machine learning methods.

<table>
<thead>
<tr>
<th>Domain</th>
<th>Domain Definition</th>
<th>ACO-OS CCLF Claims Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Admitting diagnosis</td>
<td>Assigned with the information known when the patient is first admitted (prior to diagnostic testing and evaluation). These data may be symptom-oriented; for example, shortness of breath at admission may be later diagnosed as congestive heart failure through diagnostic evaluation. Aggregated into CCS diagnosis category¹.</td>
<td>Part A Claims Header File (CCLF1)</td>
</tr>
<tr>
<td>Discharge diagnoses</td>
<td>Assigned by medical coders after the conclusion of the encounter, and incorporating the results of diagnostic testing and provider evaluation. The primary diagnosis assignment is generally weighted by medical coding practices. Aggregated into CCS diagnosis category*.</td>
<td>Part A Claims Header File (CCLF1) and Part A Diagnosis Code File (CCLF5)</td>
</tr>
<tr>
<td>Procedures</td>
<td>The discreet medical interventions (such as surgical procedures) and execution of diagnostic testing (such as laboratory orders) delivered within a healthcare context. Assigned by medical coders based on facility and provider documentation of services rendered. Aggregated into CCS procedure category*.</td>
<td>Part A Procedure Code File (CCLF4)</td>
</tr>
<tr>
<td>Dispensed medications</td>
<td>Medication dispensed directly to a patient by a pharmacy; this is different than medications prescribed or administered within a healthcare facility. Codified in the highly granular National Drug Code (NDC) terminology, which specifies packaging and other attributes of the drug product. Derived into medication active ingredient and class from RxNorm**.</td>
<td>Part D File (CCLF7)</td>
</tr>
</tbody>
</table>

¹ Aggregated using the HCUP Clinical Classifications Software, a common-used schema that is used here to aggregate the individual ICD-9 and ICD-10 encounter-based codes into more meaningful, higher-level categories.

² The National Library of Medicine (NLM) RxNorm medication normalization naming platform is part of the open-source, publicly available Unified Medical Language System® (UMLS®). Here, the original NDC codes for each ingredient are mapped against RxNorm to retrieve the medication’s active ingredient and class listing from RxNorm.

Discussion
We anticipate that the organizing principles of performance and pragmatism will strongly influence our models’ eventual adoption and operationalization. One design consideration is evaluating the tradeoffs in timeliness and accuracy between modeling based on ACO claims versus EHR data, versus the combination of the two. Our current work has been based on Medicare claims data; as a next phase, we plan to combine these claims data with EHR data for the same patient population, and leverage supplemental data sources including geospatial data associated with patient addresses and other auxiliary sources.

Challenges for the pilot have included acquiring and deploying computational resources sufficient to support the high-performance data science within a typical enterprise IT environment, and the effort to assess the claims data structuring and nuances, especially attributes and concepts specific to the MSSP.

ACOs can help health systems achieve better outcomes by efficiently identifying and reducing health risks across populations, and we expect that the pilot model will more quickly and accurately identify high-risk individuals. This improved identification will create efficiency by shifting effort from risk assessment to actual care intervention and patient engagement.

References
Dr. Babel Fish: A Machine Translator to Simplify Providers’ Language

Tarek Sakakini¹, Renato F.L. Azevedo¹, Victor Sadauskas², Kuangxiao Gu¹, Yang Zhang¹, Suma Bhat, PhD¹, Dan Morrow, PhD¹, Mark Hasegawa-Johnson, PhD¹, Thomas S. Huang, PhD¹, Ann Willemsen-Dunlap²,³, Donald J. Halpin²,³, James Graumlich, MD²
¹UIUC, Urbana, IL, ²UICOMP, Peoria, IL, ³Jump Sim Center, Peoria, IL

Introduction

In the US, 37% of adults lack the health literacy needed to navigate the healthcare system.¹ Moreover, there exists a negative correlation between health literacy and age.¹ In other words, older adults, who are the largest demographic group interacting with the healthcare system, are often the least qualified. With low levels of health literacy resulting in worse health outcomes,² we find it necessary to reduce the gap between the health literacy of patients and the health literacy demands of the US healthcare system.

High literacy demands of the US healthcare system are due in part to the providers’ complicated usage of language. EHR systems normalize providers’ language through drop-down menus, with the provision to use free text when needed. The nature of the text occurring in the free text tends to be complex from a layperson’s standpoint for multiple reasons. One reason is abbreviations such as the term “RA”, which can stand for “Rheumatoid Arthritis”. Another reason is professional terms originating from Latin such as “prn” (pro re nata), which means “as required”. Also, professional terms such as “prophylaxis”, which means “prevention”, introduce complexity to the providers’ language. Not only do abbreviations increase the complexity of the used language, they also introduce ambiguity. For example, the abbreviation “RA” could mean “Rheumatoid Arthritis” or “Refractory Anemia” among other possibilities depending on the context it appears in. Borrowing from the field of Machine Translation (MT), we propose an automated system, which, to the best of our knowledge, is the first of its kind. By means of a unified framework, our system is designed to tackle both issues (complexity and ambiguity). It translates the multiple types of complicated terms to their simpler counterparts, so that patients with low health literacy can better understand their health information.

Method

Our system is best described as a 4-stage pipeline. The system is further illustrated in Figure 1.

1. Identifying complicated terms: We first identify the complicated terms in the sentence. Assuming complicated terms are infrequently used by laypeople, we consider all words but the 10000 most frequent words as complicated, after collecting frequencies on a corpus (GoogleNews) intended for laypeople. The low threshold ensures that all complicated terms are identified (high recall).

2. Generating candidate translations for complicated terms: Then, for each complicated term, we generate multiple surface forms using the Unified Medical Language System (UMLS).³ All atoms of the first 2 retrieved concepts using the UMLS search feature are considered candidate translations.

3. Generating all candidate sentence translations: Then, we generate all sentences corresponding to all combinations of translations.

4. Scoring candidate sentences: Finally, we score all candidate sentences using a 5-gram language model⁴ trained on a corpus (Gigaword) of texts geared towards the general public. The system outputs the sentence with the highest score. The language model ensures patient friendliness (comprehensibility) and accuracy of the output.

Results

We validate our method by comparing against 3 simpler but strong baseline automatic approaches evaluated for two output criteria: comprehensibility and accuracy. The first baseline takes a dictionary (Dict) approach by utilizing UMLS built-in search feature and the “preferred name” annotation to generate the translation of a word deterministically. This baseline replaces a complicated term by the “preferred name” of the first concept returned by the UMLS database when queried using the complicated term. The second baseline differs from our system in only the fourth stage where it chooses the candidate sentence with the maximum average word frequency (MaxFreq), with frequencies derived from a general domain corpus. The third baseline differs from our system only in the second
stage, whereby instead of the UMLS database, it utilizes state-of-the-art algorithms in NLP to map words into a vector space. Words with the same meaning will be closeby in the vector space. This third baseline generates translations by retrieving the 20 closest words in a vector space created using a word embeddings algorithm (WE).

To evaluate the 4 systems, we collect 150 sample medication instructions intended for patients from an EHR dataset. For each sample instruction, we generate the output of each of the 4 systems and evaluate comparatively. To evaluate accuracy, two physicians independently ranked (87.96% agreement level) all four outputs (anonymized) in terms of accurately depicting the provider’s intent. Our system outperformed the baselines with margins (number of times ranked higher - number of times ranked lower) of 62%, 50%, and 23% against Dict, MaxFreq, and WE respectively. Moreover, in 81% of the instances, our system precisely depicted the provider’s intent. As for evaluating comprehensibility, 4 laypeople (88.46% agreement level) ranked the input against our system (after anonymization). In 20% of the instances, our system increased comprehensibility, reduced it in 12% of the instances, and maintained it in the rest. Hence, results show that this is a promising direction towards simplifying medication instructions while also maintaining provider’s intent. Some ways in which we are addressing the equivocal results of our system in bringing about improved comprehension of medication instruction are mentioned in the Discussion section.

Discussion

We have presented an MT framework for simplifying providers’ medication instructions in EHR. This framework is expected to lower the health literacy requirements for patients in the US, and thus optimize provider-patient communication, enhance health outcomes, and reduce health costs. The performance gains over Dict and MaxFreq emphasize the significance of a language model and the MT framework of our system, whereas those over WE emphasize the benefits of using UMLS to generate our simple translation model. Ongoing experiments include, enhancements to the current MT framework through extending our translation model (unequal distribution of translation probabilities based on ranking in UMLS search results), improving the language model (collecting and training on a medicine-related corpus), and extending the decoding process (phrase-based instead of word-based). Moreover, current projects include speech extensions to Dr. Babel Fish making it an interactive medical Conversational Agent.

References

DeepPhe - A Natural Language Processing System for Extracting Cancer Phenotypes from Clinical Records

Guergana Savova 1,2, PhD, Eugene Tseytlin 3, MS, Sean Finan 1, MS, Melissa Castine 3, Tim Miller 1,2, PhD, Olga Medvedeva 3, David Harris 1, Harry Hochheiser 3, PhD, Chen Lin 1, MS, Girish Chavan 3, MS, Rebecca Jacobson 3,4, MD, MS
1Boston Children’s Hospital, Boston, MA; 2Harvard Medical School, Boston, MA; 3University of Pittsburgh Department of Biomedical Informatics, Pittsburgh, PA; 4University of Pittsburgh Cancer Institute, Pittsburgh, PA

Introduction: In this manuscript, we describe a novel double-pipeline system for extracting rich cancer phenotypes from sets of EMR documents, combining approaches of information extraction and summarization. We demonstrate the accuracy of the system in extracting key variables as measured against a human-annotated gold standard. Our work differs from previous efforts in that we move beyond entity mention recognition to episode- and patient-levels over the entire set of patient’s records (pathology, oncology, radiology notes etc.), and longitudinally from primary tumor to regional recurrence or metastasis. These tasks require sophisticated extraction techniques such as coreference resolution, relation extraction and temporal relation extraction to achieve reliable summarization. Our work also differs by generating a patient-level summarization across multiple documents and genres, informed by existing cancer phenotype rules and definitions. Project website: http://deepphe.healthnlp.org.

Methods: The DeepPhe system ingests multiple different clinical documents and optionally also discrete data, and outputs a single summary of the patient’s clinical phenotype. The system uses a novel double-pipeline design, combining a mention-annotation pipeline or pipeline 1 (an extension of cTAKES© (1); ctales.apache.org) a phenotype summarization pipeline or pipeline 2 and the DeepPhe domain ontology (2). The methods in each module range from pattern-matching to modern machine learning to knowledge engineering methods and are a result of our research efforts in the field of natural language processing. The system utilizes the engineering framework of Apache Unstructured Information Management Architecture (UIMA; uima.apache.org). A high level visualization of the system is shown in Fig. 1.

Fig 1: DeepPhe system architecture. Abbreviations/acronyms: cTAKES – Apache Clinical Text Analysis and Knowledge Extraction System; NLP – natural language processing; UIMA – Apache Unstructured Information Management Architecture; FHIR – fast health interoperability resources; Neo4J – graph database for visualization; tranSMART – phenotype-genotype analysis platform; http://transmartfoundation.org; DROOLS – rule-based inference engine; PLAY – application framework on which the visualization is built.

Results: Results are in Table 1 and 2. In addition, the performance of the coreference module is MUC 0.53, B3 0.47, CEAF 0.54 and CoNLL 0.53. MUC computes recall by counting the number of inter-cluster links in the gold that are found by the system. Precision is computed by reversing the gold and system outputs. B3 computes precision and recall scores for each mention, as the number of correct elements in the chain containing that mention, divided by the number of elements in the system chain or gold chain, respectively. CEAF computes an optimal alignment between the reference and gold mention sets, and compares that alignment to an alignment of gold outputs to itself (for recall) and system output to itself (for precision). The CoNLL score averages the F1 of those three metrics.
Table 1: Pipeline 1 evaluation results (mention-level). Numbers in brackets are inter-annotator agreement. BrCa - breast cancer; n/a – attribute not applicable

Table 2: Pipeline 2 evaluation results (Phenotype-level) for Cancer summary (A) and Tumor summary (B). Numbers in brackets are inter-annotator agreement. Summarization over 41 documents for two patients. cT – clinical tumor; cN – clinical lymph nodes involvement; cM – clinical metastasis; pT – pathology tumor; pN – pathology lymph node involvement; pM – pathology metastasis; n/a – no metric derived

Discussion: The initial evaluation of the system is quite promising. Results at both mention-level and phenotype-level approximate the IAA which is usually set as a system goal. Extraction of specific attributes in the phenotype pipeline are acceptable but require further refinement. Our study emphasizes even further the importance of research in key (albeit extremely challenging) areas of NLP and information extraction including word sense disambiguation, relation extraction (e.g. coreference, temporal and body location relations) as well as summarization. The DeepPhe software extends the existing state-of-the-art by providing a generalizable information extraction framework for cancer phenotypes. The development of platforms such as DeepPhe provide critical and missing infrastructure to serve the data-intensive needs of precision medicine and cancer surveillance. Future work will include (1) the development of a visualization interface to enable end-users to interact with the data, (2) the application of the system to other types of cancers, including melanoma, lung cancer, and ovarian neoplasms, (3) the expansion of extracted information to include clinical genomic observations, such as somatic variants (e.g. BRAF status), or gene rearrangements (e.g. EML4/ALK) and (4) inclusion of structured EMR data (e.g. synoptic pathology reports).

References
Novel Approaches to Identifying Rare Diseases Using Electronic Health Record data, Mabry Syndrome as an example

Prabhu R Shankar MD, MS1, Rachel Logan2, Allison Foley3, MS, Hong Li3 MD, PhD

1Dept. of Public Health Sciences, School of Medicine, UC Davis, Sacramento, CA
2 Emory Genetic Counseling Training Program and 3Division of Medical Genetics, Department of Human Genetics, School of Medicine, Emory University, Decatur, GA

Introduction: Rare diseases are uncommon individually, defined as affecting fewer than 200,000 people or 1 in 1,500 in USA (1 in 2000 in Europe). Collectively, however, these diseases affect millions of people worldwide. They can be chronic and debilitating and negatively impact the lifespan of affected individuals. When therapies are available reasonably good quality of life can be achieved, particularly with early diagnosis and intervention. The advent of Next-Generation Sequencing (NGS) has changed the landscape of rare genetic disease diagnosis and research. Advances in NGS, such as Whole Exome Sequencing (WES) and Whole Genome Sequencing (WGS) witnessed a rapid growth in identification of causative genes for rare diseases. By 2012, more than 180 novel genes had been identified by WES1 and new ones are continuously being added. This has spurred interest among genetics researchers to recruit and study patients and families with rare diseases. However, Identification of rare diseases is a challenge as many rare diseases are not recognized during routine patient care and consequently are not documented. Some of the reasons for missing rare diseases include nonspecific presentation during childhood, poorly defined or evolving case definitions, or lack of disease awareness among healthcare professionals. Further, the disease classification and coding systems that are currently used for documenting, reporting and billing purposes, are not granular enough to capture many rare diseases. All of these issues make it difficult to establish search strategies to define patient cohorts using current informatics tools on stored EHR data. Novel methods need to be developed to identify patients with rare diseases utilizing the rich clinical data that has accumulated over the past decade as part of increased implementation of EHRs. We define a new method using Mabry syndrome, a rare disease, as an example.

Mabry syndrome, also called Hyperphosphatasia with Mental Retardation Syndrome (HPMRS), is a group of inherited disorders caused by mutations in genes involved in the Glycosyl Phosphatidylinositol anchor (GPI) pathway. Clinical presentations are variable, but predominantly manifest with developmental delay, intellectual disability or seizure. A unique biomarker of the syndrome is consistently elevated Alkaline Phosphatase (ALP) levels. Diagnosis is often missed and under diagnosed due to 1) variable and non-specific clinical presentation 2) lack of knowledge about the unique chemical biomarker ALP 3) often misinterpret the elevated ALP level to a common effect of growing bone or liver dysfunction without further confirmation 4) genetically heterogeneous nature of Mabry with at least 8 genes being newly identified as disease-associated, and are not part of the currently available commercial seizure/epilepsy/intellectual disability gene testing panels2. A unifying genetic diagnosis of Mabry syndrome can provide family with prognostic information and precise risk assessment for family planning. Some individuals may benefit from a high-dose vitamin B6 treatment for seizures.

Methods: This is a retrospective study conducted using stored EHR data in a tertiary care children’s hospital. After appropriate IRB approval and following standard operating procedures, data Warehouse and data mart was queried from 2012 – 2015 for all patients who had an alkaline phosphatase levels more than 500 international units per liter (IU/L). For each of the patients, demographics, all the associated diagnoses, clinics visited, and relevant other investigations were extracted. Protocol was developed by the genetics study team for phenotypic definition of patients with Mabry Syndrome. Data was analyzed based on consistent elevation of alkaline phosphatase levels and refined iteratively based on associated neurological clinical features, such as intellectual disabilities, seizures and developmental delay, to arrive at a list of most probable patients with Mabry Syndrome. Premature babies, patients with bone/ liver diseases, cancers and medications that may elevate ALP and patients with organ transplants were excluded. Further, chart review of the refined list of patients was done by clinical research team to identify possible Mabry syndrome patients that met the clinical phenotypic definition, for further genetic studies.

Results: There were 2,823 patients with ALP levels at more than 500 IU/L seen at the Children’s Hospital during 2012 – 2015. After excluding patients with known hepatic or bone disease, extreme prematurity, cancer and recipients of organ transplant, and filtering out those who did not have any neurological presentations, 126 (4.4%) patients were left in the list. Further filtering with objective evidence of having had a neurology visit/ consultation with clinical documentation, and persistent elevation of ALP levels, there were 24 (0.85%) patients that were ear marked for manual chart review. Twelve patients (0.42%) met the clinical case definition and were further
considered by genetic study team for thorough clinical evaluation, targeted genetic and functional studies. Currently 6 patients have been recruited and biochemical, genetic and other tests have been conducted. Results will be presented and discussed.

**Discussion:** We describe a new method to identify patients with rare diseases using stored electronic clinical data, collected as part of routine patient care, adding to the growing list of secondary use of EHR data. Here we use sustained elevation of ALP levels, a unique biomarker for Mabry syndrome, along with major Mabry phenotypes (some represented with ICD codes) as part of an algorithm to process EHR data to zero in on patients that match the clinical Mabry syndrome case definition. As mentioned above, the traditional informatics tools to search for certain rare diseases, such as using structured diagnosis codes (ICD and SNOMED) and NLP to search clinical narrative free text data for any mention of the disease, may not yield results as those rare diseases might not have been captured in the clinical notes. This is because the diagnosis was missed due to lack of awareness about the rare disease or inadequacies of diagnosis coding system. A combination of biomarkers and clinical phenotypes could be used to develop algorithms to find patients that match rare disease case definitions. There is significant information content embedded in coded structured data such as diagnosis, laboratory, medications, procedures and specialty care clinic visits (ICD, SNOMED, LOINC, RXNORM, CPT, HCPCS, customized local EHR codes), that could be leveraged using data mining tools to match the case definitions of rare diseases. Additional phenotype features of the rare disease that are embedded as free text in the clinical notes may be amenable to NLP and could be added to the search criteria to streamline the data mining methods. Manual chart review is required to validate the disease descriptions that may match the rare disease case definition. Further, along with identifying potential rare disease patients that met the case definitions of the study protocol, patients with overlapping phenotypes that may be caused by defects in the genes that are part of the same metabolic pathway may also be discovered. For example, our refined list may have patients with Multiple Congenital Anomalies Hypotonia Seizures Syndrome (MCAHS) with defects in the genes encoding for Phosphatidylinositol-Glycan (PIG) proteins which are part of the Glycosyl-Phosphatidylinositol (GPI)-anchor biosynthesis pathway shared by both Mabry syndrome and MCAHS. The ability to identify patients with Mabry syndrome or other GPI-anchor defect diseases will allow us to conduct functional studies and investigate for further diagnostic biomarkers and disease pathway associations in the future.

The shortcoming of this method is that some patients that may have milder versions of the disease with less severe symptoms and signs that are compatible with normal life may not have been seen or investigated and documented by clinicians. Invariably, a certain proportion of the patient population may be missed with difficulty in estimating accurate incidence and prevalence. Additionally, not all rare diseases may have unique biomarkers to help identify them from other rare diseases that share phenotypic features. However, the same set of phenotypic features that are used to clinically define and differentiate a disease/syndrome from others could be used to carefully design algorithms, using a combination of structured and unstructured clinical data that may still yield foot prints of a particular disease. Thus, similar methods could be developed for other rare diseases. Alert systems could be set up in EHRs for prospective patient identification using methods and algorithms defined in this project.

Other approaches such as exome or genome analysis of all patients will likely be the way of the future for accurately identifying and diagnosing patients with rare diseases. Currently, the price for NGS (exome or genome analysis) is expensive and most of the insurance providers do not cover for untargeted next generation sequencing in clinical practice. Further, NGS are valuable for instances when a “pathogenic variant” as recommended by the ACMG guidelines is identified. However, NGS may result in many “Variants of Uncertain Significance” (VOUS) that will need functional analysis and supportive data to substantiate the pathogenicity of a VOUS to establish a definitive diagnosis. In addition, as an unintended consequence, NGS may identify pathogenic variants that cause other secondary conditions including adult onset neurodegenerative disorders as well as cancer. These possibilities with NGS need to be discussed with family members; will require extensive counseling and the genomic findings may have significant impact on the extended family. For these reasons, targeted NGS on phenotypically characterized patients is the current approach, at least till clinical utility of NGS in rare diseases is further established.

**References**


Introduction

Adolescence is a time when patients are approaching autonomy, both developmentally and legally. Yet they are still minors and are likely to encounter contradictions between situations in which they are treated as children and ones in which they are treated as adults. Being able to access their medical information may enable adolescents to take on a participatory role in their health care. However, federal policy, state law, and community norms are not consistent regarding adolescent healthcare and privacy. For example, in some regions and under some circumstances, adolescents may have consent and privacy rights similar to those of adults, with the right to make some, or all, of their own sensitive medical decisions privately. In other cases, parental notification is the norm, or guidance is unclear or lacking. In the absence of national guidelines, medical centers encounter serious challenges when developing policies about adolescent access to medical records via patient portals. The American Academy of Pediatrics has made recommendations, but these are not binding.

To explore diversity in adolescent privacy policies and identify common approaches, we are conducting a qualitative study with key informants from different types of medical organizations in different regions of the country. The main objective is to identify diversity in adolescent privacy features within the patient portal. Another objective is to enumerate the factors involved in making portal access decisions. A third objective is to identify the potential need for more formalized guidance and standards on privacy features within the patient portal.

Methods

We are conducting semi-structured interviews with chief medical information officers and other key informants involved in informatics operations. Purposive and snowball sampling is being conducted to identify representatives from medical organizations across the four census regions of the country and multiple organizational types (pediatric hospital, community health center, non-pediatric academic medical center, outpatient practice, public hospital, etc.). The semi-structured interview guide was developed with reference to a 2012 policy statement from the American Academy of Pediatrics (AAP) on ideal principles for the electronic health systems. The minimum sample size is 25, with representation of the four census regions of the country. However, in accordance with qualitative research best practices, we are conducting sampling and thematic analysis in tandem, and will adjust the target sample size upward if thematic saturation is not reached. This study was approved as minimal risk by the Weill Cornell IRB and determined to be exempt by the Georgia Tech IRB.

Results

To date, we have interviewed representatives from 26 medical centers, representing all four census regions of the United States. The most striking finding is the extreme variation in the policies regarding adolescent patient portal privacy. One policy type prohibits portal access completely during adolescence for both adolescents and parents. A second type of policy defaults to access for the adolescent only, with the parent losing access. In some of these, parents could be granted proxy access again, but the proxy rights ranged from complete access to highly limited access (for example, parents might be barred from viewing the record but be able to message with the clinician). Among those with adolescent access, the minimum age ranged from 10 to 14. However, many of the centers did not fit into either of these two policy types. For example, one granted default access for both the parent and the adolescent, and one gave default access for the parent of the young adolescent only until age 16, at which time there was no further access for anyone until 18.

Informants cited many different factors that went into the decision making of adolescent portal access. This study organized these decision-making factors into 8 underlying themes: (1) compliance with state and federal laws, (2) EHR capabilities, (3) accommodation of different types of patient needs, (4) prioritization and availability of resources, (5) insurance of patient safety/care quality and risk aversion, (6) balance of patient

Marianne Sharko, MD, Lauren Wilcox, PhD, Matthew K Hong, MS, Jessica S Ancker, MPH, PhD

1. Department of Healthcare Policy & Research, Division of Health Informatics, Weill Cornell Medical College, New York NY USA
2. School of Interactive Computing, Georgia Institute of Technology, Atlanta, GA USA
autonomy/empowerment and family shared decision making, (7) community expectations, and (8) tension between
teen privacy and parental preferences.

Our interviews helped highlight possible sources of assistance when making portal access decisions. Of
those surveyed, most agreed that guidelines would be helpful. One informant felt that guidelines would be
particularly helpful in complicated situations; such as patients in foster care or with limited mental capacity.
However, there were differing opinions on the form of those guidelines. Multiple informants suggested that policy
could be created with input from medical organizations, physicians, nurses, patient and parent representatives,
informaticians, legal experts, EHR vendor representatives, and administration.

One informant felt that federal guidelines would be essential in the age of health information exchange
when an adolescent’s health information is transferred electronically from one state to the next. However, there
was some concern that federal guidelines would force harmful changes to a current policy, which emphasized adolescent
privacy over parental access. Another informant felt limited by state law and thought that federal guidelines would
enable them to expand their policy to grant adolescents more control over their health information.

Our study identified other potential sources of assistance for medical centers. There is a need for more tech
innovations, such as increased granular control of portal information, flexibility of control over the information in
the portal, clearer identification of sensitive medical information, and a guarantee of privacy control. Another
informant addressed the need for more research-based evidence to evaluate the importance of adolescent portal
access. Importantly, there was also a call for more outreach and education to the general public on the importance of
these issues when understanding parental and adolescent portal access.

Discussion

The increased use of the EHR has great implications for the adolescent patient\textsuperscript{4}, with opportunities for
improved access to the medical provider and to the patient’s own medical information, thereby encouraging patient
autonomy. Multiple studies have found that privacy and confidentiality are vital to adolescent openness with their
physicians as well as adherence to care.\textsuperscript{3} There have been several policy statements regarding adolescent
confidentiality and privacy in the electronic health record. The AAP released a policy statement in 2012\textsuperscript{3} stating that
current health information systems lack the capability to allow for protection of the privacy and security of health
information for minors. The Society for Adolescent Health and Medicine stated in a position paper in 2014 that
“Protecting adolescent confidentiality is a shared responsibility and requires ongoing vigilance.”\textsuperscript{6}

However, there are many different issues that factor into a medical center’s decision on how to balance the
tension between patient autonomy and family shared decision making. Determining the best way to handle
adolescent patient portal access is challenging and involves the consideration of many issues. Therefore, it is not
surprising that there is striking variation in the way adolescent portal access is handled across the country.

Our study identified potential sources of assistance in this process, including improved education and
outreach, more evidence-based research evaluating adolescent portal access, and tech innovations providing more
granular control. Clearer guidelines could be helpful, and could result in more consistent approaches to portal
access. However, it is yet unclear who should determine these guidelines and how they should be implemented.

Acknowledgments: This study is supported in part by AHRQ K01 HS 021531 (PI: Ancker).

References
1. Hong MK, Wilcox L, Feustel C, Wasileski-Masker K, Olson TA, Simoneaux SF. Adolescent and Caregiver
use of a Tethered Personal Health Record System. AMIA Annual Symposium proceedings AMIA Symposium.
3. Blythe MJ, Del Beccaro MA. Standards for health information technology to ensure adolescent privacy.
4. Moreno MA, Ralston JD, Grossman DC. Adolescent access to online health services: perils and promise. J
5. Reddy DM, Fleming R, Swain C. Effect of mandatory parental notification on adolescent girls' use of
Assessing Usability of the D2Refine Platform for Harmonization and Standardization of Clinical Study Data Dictionaries

Deepak K. Sharma, MS, Kevin J. Peterson, MS, Guoqian Jiang, MD, PhD
Department of Health Sciences Research, Mayo Clinic, Rochester, MN

Introduction

D2Refine is a web-based platform developed for harmonization and standardization of clinical research study data elements. D2Refine extends OpenRefine (formerly Google Refine) and leverages its simple interface for representing data dictionaries. D2Refine introduces features of semantically binding study data elements to controlled terminology terms or common data elements. D2Refine enhances semantic interoperability of the clinical research study models, expressed by the data dictionaries, through their harmonization and serialization into standardized representation formats like Archetype Definition Language (ADL), Archetype Modeling Language (AML) or HL7 Fast Healthcare Interoperability Resources (FHIR) Profiles. We plan to assess the usability of the D2Refine features and it will help us understand user perceived usefulness of implemented, planned and desired functions and identify gaps. The objectives of this usability study is to not only quantitatively measure usability of D2Refine platform, but also as a tool to improve development efforts towards existing and planned D2Refine features.

Methods

TURF Usability Framework

The usability study is being designed and conducted using TURF EHR Usability Framework. The usability study includes four components of TURF Framework – Task, User, Representation and Function, to quantifiably assess how useful, usable and satisfying a system is. Task Analysis is to identify the steps (physical and mental) of carrying out operations by using a specific representation. It describes relationships and dependencies among various tasks. User Analysis identifies the users and characteristics of each type of users. Function Analysis is one of the most important components where knowledge of work domain is captured. Representing explicit and implementation-independent description of work domain is usually accomplished as work domain ontology. Representation Analysis is about figuring out the appropriate representation or interface to accomplish a given task efficiently.

Use Cases

D2Refine platform and similar tools should at least enable users to import data dictionaries, bind them and then export them into standardized formats. Based on these functionality requirements, we choose five use-cases of (1) import data dictionary, (2) configure the controlled terminologies for terms search including local terms, (3) search terms from selected code systems, (4) creating term bindings from the search result candidates and, then (5) export the standardized data dictionary to a standard format.

User and Function Analysis

D2Refine is designed for Clinical Study Developers and Clinical Element Modelers. Clinical Study Developers are the ones who would be creating new data dictionaries or editing existing ones to be used in designing new studies. The functionality is derived from the use cases and the steps involved in standardization of data dictionaries and translating them into standardized representations. The use cases require availability of terms from controlled terminologies, which users can configure through an interactive interface. The comparison with other solutions similar to D2Refine would help identify additional functions and hence augment the set of functions.

Task and Representation Analysis

Use cases guide the steps of the tasks to carry out the functions. We intend to articulate the tasks and their effect of the system by creating diagrams of state-transitions. The state-transition model will represent a particular state of the system at each step of a task. This will help us formalize and guide us to enumerate the user (or system) actions needed to accomplish a task in a given workflow. We augmented the domain ontology to represent the tasks.
Metrics

We choose two metrics of TURF Usability Framework to compute usefulness, calculated using Venn Diagrams of the domain model functions (see Figure 1), are:

1. **Within-Model Domain Function Saturation**: Ratio of domain functions in Designer Model to total functions in Designer Model. This is calculated as sum of numerators of the fractions in \( (A, D, E, \text{and } G) \) divided by sum of denominators of the fractions in \( (A, D, E, \text{and } G) \).

2. **Across-Model Domain Function Saturation**: Ratio of domain functions in Designer Model to domain functions in all three models (Designer, User and Activity). This is calculated as sum of numerators of the fractions in \( (A, D, E, \text{and } G) \) divided by the count of all domain functions.

Results and Discussion

The goal of the study is to gauge D2Refine’s effectiveness and usefulness using the TURF framework. We have developed the domain ontology for functions. The workflows and tasks are being added to this as we gear towards conducting the usability study. The first iteration of user analysis and the function analysis of D2Refine is completed - by capturing the functions and tasks in the Domain Ontology. A partial view of domain ontology\(^3\) [developed in RDF/OWL] representing the functional features of D2Refine is shown in Figure 2. The Usability study will compare D2Refine with few other platforms, e.g. OntoMaton\(^4\) and eleMap\(^5\), that provide similar functions and this would augment the domain ontology. The representation analysis of the study (in-progress) would evaluate how well D2Refine represents and aligns with user’s expectations. Using TURF’s Affordance ratings method, we plan to compare isomorphic representations (functionally equivalent) of tasks, which would help us identify gaps and value that D2Refine offers. Identified users would be invited to participate in the study and perform the tasks. The part of observations for each task would also be captured in domain ontology itself. We intend to compute and compare the metrics of usefulness for D2Refine leveraging RDF/OWL reasoning. This usability study would definitely guide us to evaluate our existing strategy and future development efforts.

Acknowledgement

This work was supported in part by funding from a NIH R01 GM105688 and a NCI U01 Project – caCDE-QA (U01 CA180940).

References

5. eleMap : A Platform to harmonize local phenotype data dictionaries: Vanderbilt University; 2010 [March 06, 2017]. Available from: https://victr.vanderbilt.edu/eleMAP.
Rule-Based Criteria for Recommending Algorithms and Visualization Techniques for Biomedical Data Analysis

Leslie Sim, Charlotte Cheng, Ajay Shah PhD, Srisairam Achuthan PhD

Research Informatics and Systems Division, Center for Informatics and Data Science, Beckman Research Institute, City of Hope, CA

Abstract

Data visualization and machine learning are increasing being adopted to analyze biomedical datasets in translational research. With the goal of guiding a researcher to relevant visualization methods and learning models, we have developed SPIRIT-SA, a comprehensive scientific analytics platform. Rule-based criteria derived from data provide visualization techniques and machine learning algorithms recommendations to a user enabling seamless biomedical data analysis.

Background and Introduction

Machine learning and data visualization are increasing being adopted to analyze biomedical datasets in translational research with applications spanning from disease diagnosis to hospital readmission rate prediction. With different types of datasets being utilized in a variety of biomedical applications and a number of visualization techniques and machine learning algorithms in literature, it can often be difficult to decide the best methods to analyze the data in order to yield interesting, actionable, and predictive results. With the goal of guiding a researcher to relevant visualization methods and learning models, we have developed a rule-based criteria derived from data that provide visualization techniques and machine learning algorithms recommendations to a user enabling seamless biomedical data analysis.

Rules derived from data such as data size, number of features, visualization context and machine learning task at hand can be useful to develop a recommendation system to select relevant visualization methods and machine learning models. Each machine learning algorithm has certain strengths and weaknesses which may be suitable for particular datasets, and thus, correctly selected machine learning algorithms can make analysis of datasets more useful. Visualization techniques differ depending on the type of data being analyzed as well as visualization task at hand. Therefore, utilizing the appropriate visualization technique depending on the data context is of utmost importance.

Methods

We have developed SPIRIT–SA, a comprehensive scientific analytics platform, as part of SPIRIT (Software Platform for Integrated Research Information and Transformation) to provide a rule-driven approach to simplify the data visualization method as well as the machine learning model selection problems. Pipeline Pilot⁴, a commercial data pipelining and scientific informatics platform with a large collection of components available for analyzing biomedical datasets is the underlying framework for our scientific analytics platform, SPIRIT–SA. Raw biomedical data sets such as those from a public source like UCI Machine Learning repository [1-2] can contain errors, missing data or outliers. Standard missing data imputation methods and Open Refine [3] are used for data clean up. An initial visualization of the processed data is made available using Orange [5]. This is followed by a summary of the data. One can either select a subset or all of the features to analyze the data. Visualization and machine learning algorithm selection rules are then applied. These rules are derived based on the data size, data visualization purpose, number of features to be analyzed, machine learning task and data properties. Visualization methods falling under categories such as: V1) Distribution, V2) Relationships, V3) Hierarchy, V4) Composition, V5) Geographic, V6) High Dimensional, V7) Time and V8) Comparison are used for selecting appropriate visualization techniques. Rules such as: M1) More than 50 samples present or not, M2) Labeled data present or not, M3) Predicting a categorical variable or not, M4) Accuracy is important or not are used for recommending appropriate machine learning algorithms. The rules we have mentioned here are a snippet of about a number of rules we have implemented in SPIRIT–SA. Finally,
a summary of all the results are captured with metrics numerically and graphically. Both, the visualization techniques as well as the machine learning algorithm rule based recommendations have been built using VisiRule [6]. The techniques and algorithms were implemented using Orange, Pipeline Pilot, Tableau, R and Matlab.

Results
As a proof of concept, we demonstrate our approach using two different biomedical datasets from the UCI Machine Learning Repository [1-2]. Additional illustrations related to bioinformatics datasets will be discussed. Dataset 1 (961 instances with 5 features) is used to predict the severity (outcome - binary response) of a mammographic mass in a patient. After pre-processing the data and selecting features of interest, histograms and box-plots available to us under visualization rule V1 can be selected to visualize the selected features across the two cohorts. Applying rules M1-M4 for this dataset we recommend that neural network, random forest and decision tree algorithms be used for the supervised learning task. Initial analysis of the prediction accuracy indicates that both neural network and decision tree models have the highest accuracy. The parameters of the random forest model had to be optimized to obtain better performance. Dataset 2 (345 instances and 6 features) is used to cluster male patients likely to develop liver disorder. After pre-processing the data and selecting all features, scatter plot matrix available under visualization rule V6 can be used to visualize the multi-dimensional dataset. Rules M1: More than 50 samples present. M2: Labeled data present, M3: Number of clusters known for this dataset leads us to recommend PAM (Partitioning Around Medoids), K-Means and Fanny clustering algorithms.

Conclusions
SPIRIT-SA can be a powerful end-to-end scientific analytics platform that has been developed to provide a seamless user experience for analyzing biomedical datasets. Key features that highlight the utility of this platform are the rules based approach for selecting visualization techniques and machine learning algorithms appropriate for a given dataset. A guided approach to choosing visualization methods and machine learning models make the tasks of prediction and trend analysis when analyzing data as seamless as possible for a user. The machine learning algorithm rules help a user determine which algorithms to implement on their dataset, whether to use machine learning, and when certain types of algorithms are most suitable. SPIRIT-SA will immensely help multiple translational research projects.

References
5. Orange. [http://orange.biolab.si/](http://orange.biolab.si/)
6. VisiRule. [http://www.visirule.co.uk/resources](http://www.visirule.co.uk/resources)

Learning Objective
After participating in this session, the learner should be better able to:

- Gain better insight to processing raw data.
- Selecting appropriate technique to visualize a given dataset.
- Choosing machine learning models to recommend for analyzing a given dataset.

Keywords
- Machine Learning, Data Processing, Data Visualization, Data Analysis Pipelines.
Personal Health Information Management in the Virtual Home: 
A New Taxonomy for a New Perspective

Catherine Arnott Smith, AMLS, MA, MSIS, PhD\textsuperscript{1}; Yuqi He, MLIS\textsuperscript{1}; Gail Casper, RN, PhD\textsuperscript{1}

\textsuperscript{1}University of Wisconsin-Madison, Wisconsin, USA

Introduction: Patient work “is situated in the larger context of life”\textsuperscript{1} and the home is where the majority of patient work occurs. Most studies of patient self-management focus solely on behavior, not on the tools used or the context in which the behaviors occur. There is evidence that individuals are more comfortable in familiar environments (e.g., one’s home) and as a result will perform tasks in a different way\textsuperscript{2}. To understand patient behavior holistically, outside of the clinic, independent of treatment, we must understand the spaces in which people live and perform their health-work.

Methods: This study utilized data from the vizHOME Project (AHRQ R01HS022548), the goal of which is to understand the impact of the home environment on Personal Health Information Management (PHIM). The vizHOME project defines PHIM as a broad range of behaviors including seeking health information on the web, applying information in decision-making, developing a system to remember appointments, being reminded by an object (a cue or reminder) to do something, and self-monitoring or self-tracking. The vizHOME team used advanced visualization tools (see 3) to create perfect 3D models of the interiors of 20 real houses representing four types of homes as characterized in the U.S. Census: detached, semi-detached, multi-unit and mobile (or manufactured) homes (https://www.eia.gov/emeu/efficiency/recs_2_table.htm). These homes had not been modified to accommodate any chronic illness or disability; participants were asked not to change or “prepare” their homes for scanning visits. These scans generated a reference set of 20 virtual households. By study design, in-depth interviews of household occupants were explicitly focused on information-intensive tasks representing medication management, self-monitoring, and general information management tasks - all real-life tasks deemed important by the participant. These interviews resulted in a catalog of personal health information management (PHIM) tasks performed in each of the 20 houses\textsuperscript{3}.

In the second phase of the study, six researchers with expertise in informatics, industrial engineering and nursing performed home assessments in 16 of the virtual homes—four of each U.S. census type-- in a six-sided immersive CAVE automatic virtual environment. Each house was viewed three times resulting in a total of 47 individual assessments (one assessment was discarded due to a CAVE malfunction). Each individual researcher identified items in each home that would be useful for PHIM; these items were captured electronically and by a note-taker. They resulted in a cumulative list of features, that is, both objects and spaces, further refined and normalized by consensus. Thus, a key product of this work is a taxonomy enumerating 71 PHIM-supporting “features” – including both objects and spaces - in the interior household environment.

The third phase of the vizHOME project enabled us to test the taxonomy we developed. Twenty lay people who had been told they have diabetes each performed two virtual home assessments. This resulted in a total of 40 individual assessments. The lay participants selected a total of 54 PHIM features, of which 51 matched the features selected by the experts, and three that were not chosen by the researchers: the bed, bathtub and glucometer. The original list of 71 features important to PHIM in the home forms the corpus upon which our taxonomy was built. These features range from things that are obviously healthcare-related – a pill organizer, reminder cards – as well the less obvious – a backpack (described by one professional as a tool used to carry diabetes supplies in the house) and whiteboard (used to record and share health information).

Our taxonomy is based on the Art and Architecture Thesaurus (AAT), a reference source for thesaurus builders used in different fields to facilitate classification and information retrieval. The AAT’s descriptors are drawn from hundreds of sources in the fine and applied arts.

Results: Sixty-one (86%) of the vizHOME household features were present as concepts in the AAT, either as exact matches or modified matches. Ten (14%) features are exclusively associated with healthcare activities. We found
that the ECRI Universal Medical Device Nomenclature System (UMDNS), a well-accepted source of vocabulary for describing home health equipment, was able to match all but 2 of those 10 features (see Table 1 below).

Table 1. Match results between vizHOME household features, AAT and UMDNS

| Step 1: Match results between vizHOME household features and AAT (Number = 71) |
|-----------------|-----------------|-----------------|
| Number | Percentage | Examples |
| Exact Match | 34 | 48% | Floor – Floors |
| Modifications | 27 | 38% | Backpack – Knapsacks; water dispenser - dispensers |
| No Match | 10 | 14% | Whiteboard; Inhaler; CPAP; TV and remote; |
| Total | 71 | 100% | |

| Step 2: Match results between AAT “no match” features and UMDNS (Number = 10) |
|-----------------|-----------------|
| Number | Percentage | Examples |
| Exact Match | 8 | 80% | Inhaler - inhaler; CPAP - CPAP; |
| No Match | 2 | 20% | Free weights; TV/ Remote |
| Total | 10 | 100% | |

Even in virtual worlds, space is an important dimension. Naming, and the sorting of names into meaningful categories, is how humans make semantic sense of those physical instantiations. We adopted AAT terminology to label our categories and sub-categories. Eight categories of the AAT hold the features in our vizHOME taxonomy: containers, furniture, storage spaces, architectural elements, information form, timepieces, equipment, and telecommunication. For the ninth category, we developed the category label “Durable Medical Equipment” to represent features such as CPAP and walker, terms not unsurprisingly missing from the AAT, but common terms used by third-party payers in the US healthcare system.

Discussion: To model health behaviors at home, a general taxonomy of household objects and spaces – as opposed to a purpose-built health taxonomy – is required. Existing knowledge representations of health-related objects and spaces, for example, the Unified Medical Language System (UMLS), focus on individual behavior, absent of context. For example, pillboxes are recognized by the UMLS, but only as an aspect of desirable behavior on the part of the patient: “Promoting medication adherence using pillbox” [C2364343]. Sofa is recognized only as part of the patient’s medical history, “exposure to sofa fire” [C2905436]; steps are something patients fall from [C0417023] and toilet is something a patient either does [C0562819] or does not [C0562820] use.

What makes a particular household feature “important” or “useful”, to domain experts no less than patients and members of their households, may be shaped by the interaction of the feature with its physical context. The feature’s location in space, the feature’s proximity to other features of the room, may be as important in personal health information management as the feature’s affordances. Our new taxonomy offers formal ways to characterize and understand both the virtual and actual home. This organized knowledge can give us insights on where home environment may influence health information management and behaviors. Ultimately, it may help inform the design and deployment of consumer-facing health information technologies due to its capacity for knowledge representation of home context.

References


Discordances between Patient-Reported Family History and Family Histories in EHR

Jung Hoon Son, MD\(^1\), Chunhua Weng, PhD\(^1\)

\(^1\)Columbia University Department of Biomedical Informatics, New York, NY, USA

Abstract

Reporting of data quality problems in family history data is difficult due to the complexity of family history data. Although statistical methods can be utilized to report data quality issues in family histories, visualization of the discordance between family histories documented in EHRs and patient-reported oncology family histories promises to be a more effective method and makes results more comprehensible.

Introduction

Family history information plays an important role in genomic medicine (1) and allows the calculation of risk for heritable conditions. Subsequently, efforts to supplant this traditionally clinician-curated data with patient-provided data has been gaining momentum (2). However, traditional measures of data quality are rather difficult to apply to family history data because of the heterogeneity and complexity of family history data (3). Statistical tools such as positive and negative agreement measurements (2) and kappa values may not convey the discrepancies intuitively; therefore, we aim to explore more informative ways to visualize the data quality of family history.

Methodology:

We obtained patient-reported surveys completed by patients with personal cancer histories from Columbia University Medical Center’s Data Base Shared Resources (DBSR) center, which maintains a repository of cancer-related clinical databases for our institution. The family history section elicited detailed extended positive and negative family history for diagnoses of various types of cancer. A manual chart review was conducted on 702 patients with the purpose of assessing data quality of the EHR data (Allscripts Touchworks) generated by primary care physicians and oncology specialists at large academic medical center, compared to the survey data. We assessed the documentation-completeness measure (4), counting the presence of data in the EHR that were semantically equivalent to the survey item response.

Results:

The survey data contained itemized family cancer history, consisting of each family member’s relationship to the patient, cancer type, living status, and age of cancer onset. A total of 5,970 family members were identified, with 1,038 family members reported to have current or past histories of cancer. EHR data were concordant on 536 of these family members. The top cancer types are shown in Figure 1. Positive (0.464) and negative agreements (0.989) of family history data were calculated as a reference. To further illustrate the discordance between the two data sets, the percentage of family members with a diagnosis of cancer was compared (Figure 2). In addition, we examined the discordance of first, second, and third degree relatives using the same plot, which is shown in Figure 3.

Conclusion:

Our findings show the patient-provided family history data from the survey consistently contains more complete family cancer history than the EHR data, demonstrated by the large discordances between the compared data sets. The major limitation of this study is that the evaluation was performed at a single hospital system, which does not include a dedicated cancer center. Clinicians only spend a small fraction of time obtaining and documenting family history in such settings. Nevertheless, our findings suggest that family cancer history data is lost or obtained at a cursory level during the clinical interaction and/or during the subsequent documentation into the EHR. Therefore, we conclude that family cancer history data from the EHR warrants additional data quality evaluation, particularly in cases where it is intended for secondary use. Furthermore, we hope to encourage development of more direct methods of integrating patient-generated family history data into patient medical records to improve documentation of family history.

References

Figure 1

Family History Data Stratified by Cancer Types in Survey and EHR

- **Breast**: 283
- **Colon**: 115
- **Lung**: 122
- **Prostate**: 40
- **Stomach**: 10
- **Ovarian**: 39
- **Melanoma**: 10
- **Uterine**: 30

**Legend**:
- EHR Data
- Survey Data

Figure 2

Percentage of family members with positive cancer history by relationship

<table>
<thead>
<tr>
<th>Relationship</th>
<th>EHR Data</th>
<th>Patient-Reported</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mothers</td>
<td>2.45%</td>
<td>3.11%</td>
</tr>
<tr>
<td>Fathers</td>
<td>2.60%</td>
<td>3.62%</td>
</tr>
<tr>
<td>Aunts</td>
<td>3.33%</td>
<td></td>
</tr>
<tr>
<td>Grandmothers</td>
<td>2.09%</td>
<td>3.21%</td>
</tr>
<tr>
<td>Grandfathers</td>
<td>0.52%</td>
<td></td>
</tr>
<tr>
<td>Sisters</td>
<td>1.11%</td>
<td>2.11%</td>
</tr>
<tr>
<td>Brothers</td>
<td>2.68%</td>
<td>2.61%</td>
</tr>
<tr>
<td>Children</td>
<td>0.20%</td>
<td>0.44%</td>
</tr>
</tbody>
</table>

Percentages denotes percentage within all family members identified. 'n' number refers to number of family members of the subgroup.

Figure 3

Percentage of family members with positive cancer history among relatives of different degrees

- **First-degree relatives**: 7.19%
- **Second-degree relatives**: 3.13%
- **Third-degree relatives**: 13.81%

Percentages denotes percentage within all family members identified. 'n' number refers to number of family members of the subgroup.
Implementing NLP-Based Point-Of-Care CDS Drives Adherence to Evidence-Based Clinical Pathways

Srinivasan Suresh, MD, MBA1; Andrew Buchert, MD1; Gabriella Butler, MSN, RN1; Amy Urban, DO1; Curtis Stratman, MBA2; Nancy Zimmerman, RN, BSN3; James Colbert, MD3,4

1Children’s Hospital of Pittsburgh of UPMC, Pittsburgh, PA; 2UPMC Enterprises, Pittsburgh, PA; 3medCPU, New York, NY; 4Harvard Medical School, Boston, MA

Introduction

Health care in the United States is largely delivered in systems of care that are complex, inefficient, error prone, and costly. There currently exists a high degree of variability in the delivery of care between providers and facilities. Other industries have been successful at reducing inefficiencies and eliminating waste through the standardization of processes. Clinical pathways allow for an opportunity to reduce variability in the delivery of health care.1 As part of their 10-yr strategic plan, Children’s Hospital of Pittsburgh (CHP) of UPMC is developing and implementing clinical pathways that provide consistent, safe, efficient, effective and timely care that yield positive patient outcomes while reducing the inappropriate use of unnecessary resources. The appendicitis pathway demonstrated measurable reduction in length of stay (LOS) and costs.

PowerPlans (order sets) were created in the Cerner electronic medical record application (EMR) to promote adherence to the pathways. The clinical goal is for clinicians to utilize these PowerPlans for all cases of appendicitis. Utilization of the appendicitis PowerPlans was initially high, however CHP observed a steady decrease, and by August 2016, fewer than half of appendicitis cases were managed using the appendicitis PowerPlans. The hospital decided to implement intelligent clinical decision support (CDS) technology with the goal of improving PowerPlan utilization.

Methods

CHP implemented an advanced CDS solution that can analyze both free text as well as structured data to identify patients eligible to receive care under the defined clinical pathway. To achieve these goals, the hospital partnered with medCPU, a company offering a CDS platform with natural language processing (NLP) capabilities. Using this technology, the appendicitis CDS module identifies patients with a diagnosis of appendicitis based on real-time analysis of free-text clinician documentation and radiology reports. Once a diagnosis of appendicitis is established by the CDS platform, a prompt is fired alerting the clinician to utilize the appropriate appendicitis PowerPlan if not already ordered.

Data elements are extracted in real-time as clinicians record them into the EMR using User Interface Automation (UIA), Document Object Model (DOM) parsing and Optical Character Recognition (OCR) technologies (collectively known as the “Reader”). This capability overcomes the deficiencies of HL7 and web service interfaces that typically allow only limited amounts of data to be outbounded, thus preventing latency and enabling real-time response. The Reader runs continuously in the background on the organization’s servers. Data is compiled and processed from an unstructured format utilizing a NLP engine that achieves high accuracy scores (both precision and recall) through an innovative clinical-context extraction technology.

Data were collected to measure the effect of this CDS implementation for patients with appendicitis. We measured and compared compliance with usage of the appendicitis admission and post-operative PowerPlans for the time periods before and after the implementation of CDS, as well as the resulting impact from increased adherence to the pathway on LOS for complicated and uncomplicated appendicitis.
Results

Table 1 shows the increase in utilization of the appendicitis PowerPlan after advanced CDS with NLP was implemented. Usage of the admission PowerPlan increased from 63% to 92%, while use of the post-operative PowerPlan increased from 62% to 92% for uncomplicated appendicitis and from 52% to 94% for complicated appendicitis. As a result of the improvement in use of the PowerPlans driving adherence to the pathway, we found a significant decrease in median LOS, with a 21% decrease for uncomplicated appendicitis and a 16% decrease for complicated appendicitis. The expected cost reductions were also observed.

Table 1. Impact of NLP-Enabled CDS on LOS and Adherence to PowerPlans for Patients with Appendicitis

<table>
<thead>
<tr>
<th>Metric</th>
<th>Pre-CDS (July ’16 - Nov ’16)</th>
<th>Post-CDS (Dec ’16 - Feb ’17)</th>
<th>p value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Median Length of Stay (Uncomplicated Appendicitis)</td>
<td>28 hours, N = 148, σ = 25.1</td>
<td>22 hours, N = 79, σ = 10.9</td>
<td>0.0027</td>
</tr>
<tr>
<td>Median Length of Stay (Complicated Appendicitis)</td>
<td>121 hours, N = 34, σ = 86.7</td>
<td>102 hours, N = 35, σ = 72.3</td>
<td>0.5176</td>
</tr>
<tr>
<td>Compliance with Admission PowerPlan Usage (Uncomplicated &amp; Complicated Appendicitis)</td>
<td>63%</td>
<td>92%</td>
<td></td>
</tr>
<tr>
<td>Compliance with Postoperative PowerPlan Usage (Uncomplicated Appendicitis)</td>
<td>62%</td>
<td>92%</td>
<td></td>
</tr>
<tr>
<td>Compliance with Postoperative PowerPlan Usage (Complicated Appendicitis)</td>
<td>52%</td>
<td>94%</td>
<td></td>
</tr>
</tbody>
</table>

We also assessed the impact of the NLP-enabled CDS platform on performance of the Cerner EMR and did not find any deleterious effects. A chart review performed on March 2, 2017 found the NLP engine performed with an accuracy of 95% in identifying patients who were not on the appropriate clinical pathway.

Discussion

CHP achieved improved adherence to an evidence-based clinical pathway for appendicitis, with resultant reduction in LOS through implementation of an advanced CDS platform that leverages NLP technology to extract a clinical diagnosis in real-time from both structured and unstructured data elements. CHP is planning to expanding the use of this novel technology to impact other hospital-wide improvement initiatives.

References

A Probabilistic Scoring Model to Evaluate Providers Regarding Continuity of Care from a Relational Perspective

Kazim Topuz, PhD¹, Scott Shepherd, DO¹, William Paiva, PhD¹, Dursun Delen, PhD¹, Krista Schumacher, PhD¹, Mehmet Yildirim, PhD², Shrieraam Sathyanarayanan, MS¹
¹Oklahoma State University, Center for Health Systems Innovation, Tulsa, OK
²Wichita State University, Industrial Engineering, Wichita, KS

Introduction

Studies have shown the importance of continuity of care (COC) in improving health care and clinical outcomes. Although COC lacks a universally accepted definition, it is generally agreed upon that COC consists of three interrelated components: informational, relational, and management continuity. Relational COC is defined as an established and ongoing relationship between a single physician and patient and is considered a key factor in the provision of quality primary care through the accumulation of medical knowledge and the development of a high level of trust between patients and providers. Reid et al. contend that chronological measures, e.g., sequence of care, can be valid indicators of continuity only if chronology is closely linked to a particularly continuity type. Therefore, we maintain that chronological measures can be valid indicators of at least one aspect of relational continuity, but only when taken as a whole.

Several approaches to measuring chronological continuity exist but focus on distinct aspects of chronology of care, such as concentration of care (e.g., density), provider dispersion, sequence of visits to same provider, and number of providers seen. Most COC measures rely on visit patterns, which use the patient as the unit of analysis and fail to probabilistically represent visit patterns in terms of providers. Using Markov decision processes, we developed an index comprised of four dimensions: usual provider care, dispersion among providers, number of providers seen, and sequential continuity using visit patterns. The index is based on a framework that unifies existing indices as well as shifts the focus of analysis to providers. Our model defines the chronology of care aspect of relational continuity as a product of density, dispersion, and sequence of visits and shows transitions among providers.

Methods

Two objectives guide this research: 1) develop a framework encompassing different COC indices to create a more robust COC model; and 2) apply the proposed model in a primary care setting to study the effects of different visit characteristics on model outcomes. The Markov model contains states corresponding to each provider, and each transition represents the set of all patients for a provider. Starting with a given provider for the first visit within a pre-determined timeframe, a patient may stay with the same provider or transition to a different provider the following visit. Some of these may return to their original provider on subsequent visits. All visit patterns are represented in the model. Any given patient-visit pattern can be expressed as a transition trajectory among various states over time. Every pair of states is connected with a pair of directed arcs representing whether patient transitions have historically occurred between corresponding providers in consecutive visits. Associated with each arc is a weight, the so-called transition probability, representing the relative frequency of such transitions. This yields a Markov chain describing patient visit progression over time (Figure1).

We calibrated the Markov model using HIPAA-compliant electronic medical records (EMR) extracted from the Cerner Health Facts data warehouse. Data came from the pediatric setting of a single teaching hospital in the Midwestern U.S. and consisted of primary care providers and resident physicians. After censoring patients with < 5 primary care visits and providers with < 40 service hours per week between 2005-2015, the final dataset consisted of 248,516 patients from infancy to 21 years of age (Table 1) and 253 providers.

Figure 1. Markov chain representing historical patient transitions (arcs) among providers (states)
Table 1. Patient Characteristics (n = 248,516)

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Number (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (years)</td>
<td></td>
</tr>
<tr>
<td>11 - 21</td>
<td>92,782 (37.3)</td>
</tr>
<tr>
<td>6 - 11</td>
<td>70,867 (28.5)</td>
</tr>
<tr>
<td>2 - 5</td>
<td>51,861 (20.9)</td>
</tr>
<tr>
<td>Fetus - 1</td>
<td>33,006 (13.3)</td>
</tr>
<tr>
<td>Race/Ethnicity</td>
<td></td>
</tr>
<tr>
<td>African American</td>
<td>46,550 (18.7)</td>
</tr>
<tr>
<td>Asian</td>
<td>5,236 (0.9)</td>
</tr>
<tr>
<td>Caucasian</td>
<td>181,851 (73.2)</td>
</tr>
<tr>
<td>Hispanic</td>
<td>9,238 (3.7)</td>
</tr>
<tr>
<td>Other</td>
<td>8,641 (3.5)</td>
</tr>
<tr>
<td>Sex</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>121,206 (48.8)</td>
</tr>
<tr>
<td>Female</td>
<td>127,310 (51.2)</td>
</tr>
</tbody>
</table>

The transition probability for each pair of states represents the relative frequency of transitions between associated providers in consecutive visits and is estimated using historical data on visit patterns during a given timeframe. An unbiased estimator for the transition probability from state  \( i \) to state  \( j \), denoted by  \( P_{ij} \), is calculated by the ratio of (1) all patient visits specific to provider \( i \) during the given period to (2) the subset of patients who visited provider \( j \) during their next visit. The proposed Markov-chain model can be used to assess visit characteristics for a patient, provider, or clinic. The provider score is calculated using the geometrically distributed sojourn time associated with each state of the Markov chain. The expected provider score, or sojourn time \( st \) for each provider, is calculated by \( st_k = (1 - p_{kk})^{-1} \), where \( k \) is the provider index, and \( P_{ii} \) is the transition probability from \( i \) to \( i \). Scores represent the number of patient visits with the same provider before seeing another.

Results

Figure 2 shows the distribution of provider scores. Most scores were between 2 and 3, with a mean of 2.8. Only 18 (out of 253) scores were greater than 5, while 30 were below 1.5.

![Box-plot](image)

Figure 2. Box-plot for provider score values (color points indicate scores greater than five)

Conclusion

An ongoing relationship with a provider is a critical factor in the overall quality of primary care delivery. This study addresses a gap in the literature by unifying various aspects of COC chronological continuity to create a more robust model of relational continuity that can be used at the provider level. The model can be extended by employing transitions between hospitals and/or clinics. Future studies are needed to test the framework in other settings and validate provider scores by comparing to other measures of relational COC and patient outcomes. A limitation is that while the model is presented as addressing relational continuity for ease of discussion, it is just one aspect of relationship quality and does not consider quality of the actual interactions.

References

Retrospective and Prospective Evaluations of the System for Hospital Adaptive Readmission Prediction and Management (SHARP) for All-Cause 30-Day Pediatric Readmission Prediction

Fuchiang (Rich) Tsui, PhD¹, Victor Ruiz, MS¹, Amie Barada, MS¹, Ye Ye, MS¹, Gabriella Butler, MSN, RN², Srinivasan Suresh, MD, MBA², Andrew Urbach, MD²

¹Tsui Lab, Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA
²Children’s Hospital of Pittsburgh of UPMC, Pittsburgh, PA

Abstract
Most existing approaches to predict 30-day hospital readmissions use ad-hoc risk factors for modeling. We have developed and evaluated the innovative real-time System for Hospital Adaptive Readmission Prediction and Management (SHARP), which estimates each inpatient’s risk of readmission from structured and unstructured electronic health record data. Both retrospective evaluation (AUC 0.81) and prospective evaluation of SHARP indicate that the system has the potential to allow care providers better identify patients with elevated risk of readmission and to facilitate better discharge planning and intervention for reducing readmission burden. SHARP is now integrated into Cerner® Millennium system at the Children’s Hospital of Pittsburgh of UPMC.

Introduction
Excess unplanned 30-day hospital readmissions are an indicator of quality of care and pose a financial burden to hospitals. Moreover, public reporting of high readmission rates may affect hospitals’ reputations. In October 2012, the Centers for Medicare and Medicaid Services (CMS) started enforcing reimbursement penalties for excess hospital readmissions. Existing approaches for identifying patients with readmission risk rely on manual screening or simple predictive algorithms using few risk factors; however, these approaches either require additional resources or have limited prediction performance.

Recent studies have identified interventions that have helped reduce readmission rates in adults and children. While successful interventions have been disease-specific, readmission risk is dependent on multiple, varying factors. Thus, researchers have highlighted the need for evaluation of interventions aimed at broader (all-cause) populations, and the use of risk assessments for guiding the use of specific interventions for patients with elevated readmission risk.

To address the unmet clinical need, we have developed a real-time 30-day all-cause readmission prediction system, the System for Hospital Adaptive Readmission Prediction and Management (SHARP), which estimates each pediatric inpatient’s risk of readmission from a variety of structured and unstructured electronic health record (EHR) data. In this study, we conducted both retrospective and prospective evaluations of SHARP.

Methods
We describe dataset and study design for retrospective evaluation and prospective evaluation in this section.

Dataset
We collected clinical and administrative data for all inpatient visits to the Children’s Hospital of Pittsburgh of UPMC (CHP) from January 1, 2007 to December 31, 2013. The data types we collected were comprised of two categories: structured and unstructured. The structured data types include demographics, hospital utilization (e.g., number of hospital visits, length of stay, etc), laboratory reports, medications, vital signs, ICD-9 diagnosis codes, Current Procedural Terminology (CPT) codes, Diagnosis-related group (DRG) codes, nurse assessments, intensive care unit (ICU) length of stay, and total length of stay. The unstructured data types are narrative clinical notes including history and physical examination (H&P) notes, progress notes, and discharge summaries. We used MedLEE² to extract medical concepts from free text clinical notes.

Study Design
Retrospective Study: We hypothesized that the use of a variety of linked patient data types could improve prediction performance. We used visits from 2007-2011 as training data to learn readmission-prediction models and visits from 2012-2013 as test data to evaluate the models’ predictive performance. We included data from patients aged 0 to 21 years that were admitted to inpatient units for any cause, and excluded visits from patients who died during admission or whose age was not recorded. We trained a series of naïve Bayes (NB) and K2 Bayesian models by first learning models with feature selection process from a subset of the available data types (Table 1, Group A) and then learning additional models by incrementally adding additional data source types (Table 1, Groups B and C). We did 10-fold
cross validation. Previously we tested several machine learning algorithms and found Bayesian networks in general reached best results. We used the area under the ROC curve (AUC) to evaluate predictive model performance.

**Prospective Study:** We deployed the SHARP system at the end of December 2016 at the Children’s Hospital of Pittsburgh of UPMC (CHP). Author G. Butler reviewed all discharged patients from non-intensive care units during Jan. 2017 and identified unplanned readmissions within the discharged patient cohort. An unplanned readmission was defined as a non-scheduled visit (e.g., any scheduled chemotherapy visits or EEG visits are considered planned readmissions). We measured the sensitivity, specificity, negative predictive value and positive predictive value for the high-risk flagged patients.

**Results**

Table 1 shows the performance of the models learned in the retrospective study. Increased performance was observed as additional data sources were included, i.e., as additional data sources were added, the AUC improved from 0.77 to 0.81 and from 0.76 to 0.80 in the NB and K2 models, respectively. Before deploying the naïve Bayes model to CHP, we identified a probability threshold $P_{\text{high\_risk}}$ based on test dataset and assigned a high-risk flag when $P(\text{readmission}|\text{Data}) \geq P_{\text{high\_risk}}$ based on acceptable sensitivity and specificity.

Table 1: Retrospective evaluation of 30-day all-cause pediatric readmission prediction using AUC.

<table>
<thead>
<tr>
<th>Data Sources</th>
<th>Number of Features [Naïve Bayes]</th>
<th>Number of Features [K2]</th>
<th>AUC (95% CI) [Naïve Bayes]</th>
<th>AUC (95% CI) [K2 Bayesian Network]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group A: Medications + Laboratory + Vitals + Clinical notes, ICU stay status + Nurse assessments</td>
<td>24 (=24+4)</td>
<td>27 (=27+1)</td>
<td>0.77 (0.76-0.78)</td>
<td>0.76 (0.76-0.77)</td>
</tr>
<tr>
<td>Group B: Group A + Demographic + hospital utilization (length of stay)</td>
<td>28 (=28+5)</td>
<td>28 (=28+5)</td>
<td>0.80 (0.79-0.81)</td>
<td>0.79 (0.78-0.80)</td>
</tr>
<tr>
<td>Group C: Group B + Diagnosis (ICD-9/10 codes, CPT, and DRG codes)</td>
<td>33 (=33+5)</td>
<td>30 (=30+2)</td>
<td>0.81 (0.80-0.81)</td>
<td>0.80 (0.79-0.80)</td>
</tr>
</tbody>
</table>

In the prospective evaluation, we reviewed 957 discharges. Of the discharges, 160 were readmitted within 30 days and 102 of the readmissions had been assigned a high-risk flag. The sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) for the high-risk flagged patients were 0.64, 0.76, 0.35 and 0.91 respectively.

**Limitation**

The findings identified in this study were based on one hospital setting. Additional experiments in different populations may be required to generalize the performance.

**Discussion and Conclusion**

We found that using multiple data types improved readmission prediction performance in the retrospective study. Moreover, the high-risk flag used in deployment at CHP demonstrated an appropriate combination of sensitivity and specificity, and a majority of readmissions were captured by the system. We conclude that SHARP has the potential to allow care providers better identify patients with elevated risk of readmission and to facilitate better discharge planning and intervention for reducing readmission burden.

**Acknowledgement**

SHARP is funded by the Coulter foundation, a local charity foundation, and the Children’s Hospital of Pittsburgh of UPMC.

**References**

Developing Computable Phenotypes of Pediatric Chronic Conditions in PEDSnet

Levon H. Utidjian, MD, MBI1,2, Ritu Khare, PhD1, Hanieh Razzaghi, MPH1, Amanda Dempsey, MD, PhD, MPH3, Michelle Denburg, MD, MSCE1,2, Christopher B. Forrest, MD, PhD1,2, L. Charles Bailey, MD, PhD1,2

1Children’s Hospital of Philadelphia, Philadelphia, PA; 2Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA; 3Children’s Hospital Colorado, Aurora, CO

Introduction

The ability to accurately identify patients with specific medical conditions is essential for constructing study cohorts. Traditional cohort construction often relies upon detailed chart abstraction or administrative datasets that are limited to billing data like diagnosis and procedure codes. Electronic health record (EHR) data provide a large number of clinical variables to allow for complex cohort construction. Computable phenotypes (CP) utilize multiple EHR data elements such as laboratory results, medications, and vital signs data to define cohorts of patients with specific conditions.1

In pediatrics, achieving adequate cohort sizes is a challenge given that most chronic diseases of childhood are rare. PEDSnet is a multi-institutional network that aggregates EHR data from eight of the nation’s largest children’s hospitals. The purpose of this study was to develop and validate CPs for Crohn’s Disease (CD), Glomerular Disease (GD), and Type 2 Diabetes Mellitus (T2DM) using a reproducible process applied to structured EHR data elements that balances sensitivity and specificity in order to support case-finding for observational research.

Methods

To develop the CPs, the study team worked with clinical content experts to generate clinical specifications for chronic conditions of interest, comprising combinations of diagnostic and procedure codes, laboratory results, and medications (Table 1). These clinical specifications were then converted to technical specifications that mapped clinical definitions to a common data model (OMOP CDM version 5). Once the technical specifications were complete, code was developed and executed against the aggregated network data.

Table 1. Computable phenotype definitions

<table>
<thead>
<tr>
<th>Condition</th>
<th>Algorithm</th>
<th>Exclusions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Crohn’s Disease</td>
<td>-One or more Crohn’s disease diagnosis code; AND</td>
<td>Patients with more instances of ulcerative colitis than Crohn’s disease diagnoses</td>
</tr>
<tr>
<td></td>
<td>-One or more medications used to treat Crohn’s disease (balsalazide, mesalamine, sulfasalazine, ciprofloxacin, levofloxacin, metronidazole, rifaximin, prednisone, budesonide, azathioprine, mercaptopurine, methotrexate, infliximab, adalimumab, certolizumab, natalizumab)</td>
<td></td>
</tr>
<tr>
<td>Glomerular Disease</td>
<td>-Two visits with a glomerular-specific diagnosis code; OR</td>
<td>Presence of Systemic Lupus Erythematosus diagnosis</td>
</tr>
<tr>
<td></td>
<td>-Two visits with broader renal diagnosis code; OR</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-One visit with the above diagnoses AND a code for a renal biopsy</td>
<td></td>
</tr>
<tr>
<td>Type 2 Diabetes</td>
<td>Presence of any American Diabetes Association’s criteria for T2DM:</td>
<td>Presence of Type 1 Diabetes diagnosis</td>
</tr>
<tr>
<td></td>
<td>-Fasting glucose ≥ 126 mg/dL; OR</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-Hemoglobin A1c (HbA1c) ≥ 6.5%; OR</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-2-hour oral glucose tolerance test (OGTT) ≥ 200 mg/dL; OR</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-Diagnosis code for Type 2 Diabetes</td>
<td></td>
</tr>
</tbody>
</table>

Internal validity of the CP algorithms was evaluated through a series of models that sought to test several hypotheses regarding associations between the conditions of interest and healthcare events and used to refine the algorithms. Univariable analysis using \( \chi^2 \) testing was done on each variable used in the algorithm to test the independent associations between each variable with outcomes of interest. Logistic regression modeling was used to estimate the
odds ratios adjusted for potential interactions. External validation of the algorithms was done by using the algorithms on cohorts with confirmed diagnoses for the condition of interest and examining precision and recall.

Results

The CD CP is presented in this abstract as an example of the development and validation of the CP. For all three CPs, univariable analyses uniformly produced significant associations between CP elements and disease-related outcomes (data not shown).

As a non-case comparator for the CD CP algorithm, we used a cohort of PEDSnet patients with visits to gastroenterology (GI) clinic and GI diagnoses (n = 1,060,127 patients). For logistic regression modeling, the exposures of interest were a continuous variable for the count of CD diagnoses (1, 2, 3, 4, 5+ occurrences) and binary value for any use of CD medications. The outcomes of interest were procedures for endoscopy or a GI fluoroscopic exam. The presence of CD diagnosis codes and medications were significantly associated with each outcome of interest (Table 2).

**Table 2. Odds ratios and 95% confidence intervals for the Crohn’s disease (CD) Computable Phenotype**

<table>
<thead>
<tr>
<th>Outcome</th>
<th>CD Diagnosis Code</th>
<th>CD Medications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Endoscopy</td>
<td>1.24 (1.21-1.27)</td>
<td>1.56 (1.54-1.58)</td>
</tr>
<tr>
<td>GI Radiographic Exam</td>
<td>3.80 (3.65-3.97)</td>
<td>3.39 (3.27-3.51)</td>
</tr>
</tbody>
</table>

External validation was done by comparing different models to a cohort of 1,713 patients with confirmed CD. Models consisting of different combinations of CD diagnoses and medications were compared against this cohort. As illustrated in Figure 1, a model requiring at least three CD diagnoses with any CD medications improved precision from 0.66 to 0.81, and achieved similar recall (0.99 vs 0.98) to a model requiring only a single diagnosis.

**Figure 1.** Comparison of models for Crohn’s disease phenotypes against a curated list of Crohn’s disease patients.

Conclusion

We have developed rule-based CPs for structured data with strong internal and external validity, using a reproducible design and validation process that maximizes return from available data before confirmation using resource-intensive chart review. Our CP algorithms demonstrate how incorporating clinical data elements like procedures, medications, and laboratory results can lead to models that are better predictors of case identification than diagnosis codes alone. By using a large aggregated EHR data set, our CP algorithms can also be used to develop multi-institutional study cohorts.

References

Clinical Workflow Visualization: Representation of clinician activity from location tracking data

Akshay Vankipuram, MS1, Stephen Traub, MD2, Vimla L. Patel, PhD1,3

1 Department of Biomedical Informatics, Arizona State University, Scottsdale, AZ; 2 Emergency Medicine, Mayo Clinic, Phoenix, AZ; 3 New York Academy of Medicine, New York City, NY

Introduction
The Health Information Technology for Economic and Clinical Health Act (HITECH) enacted as part of the American Recovery and Reinvestment Act (ARRA) of 2009 introduced incentives for healthcare organizations to adopt Electronic Health Records (EHR). This has led to a significant increase in EHR adoption and as of 2015, 96% of hospitals reportedly possessed certified EHR technology and 84% had adopted a basic EHR which was up from 9.4% in 2008. In less than a decade, EHR systems have transitioned from impacting a subset of hospitals to becoming nearly ubiquitous, and, subsequently, adding a new dimension and focal point to clinical workflow discussions. Shortly after the ARRA, the Center for Medicare and Medicaid services (CMS) released their guidelines for meaningful use of certified EHRs. This has required organizations to adopt techniques to quantify clinical workflow such as patient or clinician location tracking. However, the methods to present this information meaningfully to the end-users (clinical personnel), as feedback to impact behavior or as an overview of clinical workflow, are lacking. In this paper, we present a set of visualizations of workflow metrics developed using location tracking via sensor-based technology (Radio Frequency Identifier or RFID).

Methods
The data used for this study was collected at the Mayo clinic emergency department (ED) in Phoenix, AZ. The Mayo clinic ED uses a system developed by Versus® for RFID and Infrared (IR) based location tracking. The Versus system consists of RFID+IR receivers tracking 60 locations within the ED and tags given to each tracked physician. Being a proprietary technology, we had no control over data collection or tracking algorithms but we could pull the data from the system as needed. We then ordered the data temporally. The final dataset consisted of the following attributes: location, start time at location, and end time at location. Overall, this RFID dataset consisted of one month of data (November 2016) for five clinicians (6269 rows).

Results
Figure 1 shows the time spent by clinicians in each of the locations tracked. ‘Physician Workspace’ corresponds to the location in the ED where the EHR systems are housed. All the ancillary locations were combined into an attribute called ‘Other’ making it easier to visualize. Another technique to visualize relationships between entities is known as a force-directed graph. Figure 2 shows our use of the force-directed graphs to summarize workflow. The nodes (circles) represent locations and the relationships between them (links) are meant to represent movement of a tracked subject from location A to location B. The size of the circle represents the time spent in each location while the thickness of the links is used to code the strength of that relationship i.e. the number of times a tracked subject moved from location A to location B. The third type of summary visualizations we created were using hierarchical edge bundling. A Hierarchical edge bundling plot (Figure 3) was created as an alternate plot of ED locations. Hierarchical edge bundling places all the nodes/locations on the circumference of a circle to avoid any overlap, improving clarity. Figure 3 also shows the highlighted (in red) links to all the other nodes/locations from the workspace. Figure 4 is a radar plot showing the probabilities of the next location starting at the workspace (this can be created for all the locations as start
points). These graphs are useful both for immediate feedback and as part of a workflow visualization dashboard, or to create metrics that could be used in larger data sets combining RFID data with other data sources.

Discussion and Conclusion
We created visualizations that provide a way to convey summary and individual level information rapidly to the users. Clinicians can assess their total time spent at various locations as well as their movement behaviors. In addition, we can use the results of their analyses and visualizations to create metrics that can be used in more sophisticated workflow analysis techniques for example, probabilities of location can serve as metrics that can combine with location specific behaviors to create a better picture of clinical workflow. Additionally, we can use temporal information in sequence analysis techniques to contrast physician behavior or to assess bottlenecks during a specific time within a shift. We are also looking to combine RFID analysis with EHR log files (essentially a deeper analysis of the time spent in the workspace location which takes a majority of a clinician’s time as seen in Figure 1). This is both a first step towards that larger goal of creating quantifiable metrics and a means to allow clinicians to have feedback on their workflow.

Acknowledgements
This research is supported by grant #R01HS022670 from the Agency for Healthcare Research and Quality (AHRQ). The content is sole responsibility of the authors and does not necessarily represent the official views of the AHRQ. We also thank our Mayo clinic staff who supported our data collection efforts.

References
Using EHR and HIE data to identify patients’ need for services that address the social determinants of health

Joshua R. Vest, PhD, MPH\(^1,2\), Shaun J. Grannis, MD, MS, FAAFP\(^2\), Jennifer L. Williams, MPH\(^2\), Dawn P. Haut, MD, MPH\(^3\), Nir Menachemi, PhD, MPH\(^1,2\)

\(^1\)Indiana University Richard M. Fairbanks School of Public Health, Indianapolis, IN; \(^2\)Regenstrief Institute, Indianapolis, IN; \(^3\)Eskenazi Health, Indianapolis, IN

Introduction

Increasingly, health care organizations need to deliver services that address patients’ broader social, contextual, and environmental needs. Identifying patients in need of services to address the social determinants is a critical first step to key operational activities like effective resource planning and risk stratification. However, identifying the patients in need of such services is challenging due to incomplete and nonstandard documentation, variable billing practices, technology limitations, a lack of physician time, and potential patient response biases and sensitivities. A multi-modal and multi-information systems approach was used to identify primary care patients with documented needs for social determinants of health services.

Methods

We identified needs for services addressing the social determinant of health in a sample of 84,317 adult (>18 years) patients. All patients had at least one outpatient encounter at an Eskenazi Health’s federally qualified health center (FQHC) site between 2011 and 2016. Eskenazi Health serves the Indianapolis, IN metropolitan area with 10 FQHC sites. The services of interest were: social work, behavioral health, nutrition counseling, dental, respiratory therapy, financial planning, medical legal partnership assistance, patient navigation, and pharmacist consultation. Notably, the social determinant(s) of health services were offered on a co-located basis at the primary care sites during the study period, i.e. Eskenazi Health was not relying on referrals to external providers and agencies for these services.

Because many individuals who need services do not always receive services, we could not solely rely on receipt of services as a representative indicator of need. Therefore, we operationalized the concept of “need” as an indication that a health care professional judged that any of the aforementioned services were appropriate, recommended, or potentially beneficial to the patient. This definition extended indications of need beyond only instances of delivered or received services. In consultation with Eskenazi Health’s information technology staff, need for social determinants of health services was identified using the following information sources:

1) Billing codes – From G3, Eskenazi Health’s Regenstrief Institute developed EHR, and the Indiana Network for Patient Care, the area’s health information exchange (HIE), we identified ICD-9, ICD-10, and CPT procedure codes associated with behavioral health, dental, nutritionist, respiratory therapy, and financial assistance.

2) Appointments – From four different outpatient registration systems, we identified all kept, cancelled, and “no show” visits for dietitian, behavioral health, and respiratory therapy.

3) Orders & notes – From G3’s order entry system we searched provider orders for all nine of the social determinants of health services using keywords informed by a review of the literature and professional society’s documentation recommendations.\(^1-6\) Additionally, because the G3 progress notes contained a specific section for social worker documentation, we used Regenstrief’s NLP tools to identify instances of social worker contact with patients in the outpatient setting. Because it was routine practice for all inpatients to have a social work consult before discharge, we excluded any notes associated with inpatient admission periods.

In sum, documentation of “need” reflected any ordered, scheduled, or received services. For each service, we recorded where the documentation was found (i.e. billing, appointment, or orders) and created a summary binary indicator of any need.

We described the prevalence of the need for social determinants of health services, and where information was identified, using frequencies and percentages. To assess the level of agreement between information sources, we also calculated kappa coefficients for documented need of any service and for individual services (when possible).
Results

Overall, 54% of patients were in need of services to address a social determinant of health during the study period (Table 1). A total of 22% were in need of two or more services. Based on all available information sources, nutrition counseling services were the most common needed (33%) followed by behavioral health (19%), social work (13%), and dental (11%). All other social services were less common.

For the majority (61%) of patients, the need for social determinant of health services was only documented by one source. Procedure codes were a poor source of information regarding social determinants of health services. In addition, evidence for the need of social determinants of health services was inconsistently documented among the various information sources. The overall agreement between need for any social service as defined by appointments and as defined by orders & notes was only 0.30. Similar, low correlations existed per service as well. For example, the agreement on need of behavioral health services between appointment data and order & note data was 0.16.

Discussion

While previous research has documented the prevalence of social determinants of health risk factors, to the best of our knowledge, this analysis was the first attempt to quantify the need for social determinants of health services in primary care. An urban safety-net population demonstrated a frequent need for services to address social determinants of health. Health care organizations preparing for population health activities and payment reform need to identify approaches for both ensuring patient access to social determinants of health services as well as documenting services delivery. Identification of patients in need of such services required multiple information sources. Most often the need for services was recorded in unstructured formats. Organizations relying solely on procedures codes, and even appointment data, will likely underestimate needed services that directly address the social determinants of health. Moreover, clinical information systems will require better data capture and standardization before organizations can better identify patients in need of social services, be prepared as more non-clinical services become reimbursable, and to support effective referrals to a wide range of services in support of population health.

Table 1. Prevalence of social determinants of health services in an adult safety-net primary care population, by source of information, 2011-2016.

<table>
<thead>
<tr>
<th>Service</th>
<th>All sources</th>
<th>Billing codes</th>
<th>Appointments</th>
<th>Orders &amp; notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Social work</td>
<td>12.8</td>
<td>0.0</td>
<td>0.0</td>
<td>12.8</td>
</tr>
<tr>
<td>Behavioral health</td>
<td>18.5</td>
<td>0.0</td>
<td>6.6</td>
<td>14.4</td>
</tr>
<tr>
<td>Dietitian</td>
<td>32.6</td>
<td>0.0</td>
<td>23.1</td>
<td>23.0</td>
</tr>
<tr>
<td>Dental</td>
<td>10.8</td>
<td>0.1</td>
<td>7.0</td>
<td>4.4</td>
</tr>
<tr>
<td>Respiratory therapy</td>
<td>4.1</td>
<td>0.0</td>
<td>2.7</td>
<td>2.9</td>
</tr>
<tr>
<td>Financial counseling</td>
<td>6.4</td>
<td>0.0</td>
<td>5.4</td>
<td>1.0</td>
</tr>
<tr>
<td>Medical legal partnership</td>
<td>0.4</td>
<td>0.0</td>
<td>0.0</td>
<td>0.4</td>
</tr>
<tr>
<td>Patient navigation</td>
<td>0.5</td>
<td>0.0</td>
<td>0.5</td>
<td>0.0</td>
</tr>
<tr>
<td>Pharmacist consultation</td>
<td>0.2</td>
<td>0.0</td>
<td>0.0</td>
<td>0.2</td>
</tr>
<tr>
<td>Any social service</td>
<td>54.4</td>
<td>0.1</td>
<td>36.0</td>
<td>39.8</td>
</tr>
</tbody>
</table>

References

3. Richardson MR, Haley JA, Luther S, Berndt D. Using SAS® Text Miner 4.1 to create a term list for patients with PTSD within the VA.
A Conceptual Model of Personal Health Informatics for Chronic Illness

Lisa M. Vizer, PhD¹, Jordan Eschler, MS², Bonmi Koo, PhD¹, James D. Ralston³, Wanda Pratt, PhD²

¹University of North Carolina, Chapel Hill, NC; ²University of Washington, Seattle, WA; ³Kaiser Permanente Washington Health Research Institute, Seattle, WA

Introduction

Over half of Americans have a chronic illness—described by the World Health Organization as having long duration and resulting from “genetic, physiological, environmental and behaviors factors”—with 25% having more than one. It is also the leading cause of poor health, disability, and death, accounting for up to 86% of health care spending. Self-management is central to regulating chronic illness, but we give people limited guidance about how to self-manage, transferring much of the design of day-to-day management practices to people with chronic illness. Many personal informatics tools—such as continuous glucose monitors, activity trackers, heart rate monitors, and smartphone apps—promise to ease the burden of self-management. But people experience barriers to use around choosing and using tools and making sense of data. Despite barriers, people self-track to better manage chronic illness. To decrease the self-management burden in light of evolving informatics opportunities, we need to understand and design for self-management support. Here, we see an opportunity to improve models used to understand self-management tool use to (1) incorporate the iterative nature of self-management work and (2) append a social dimension to more accurately depict the context of self-management. This abstract presents analysis of chronic illness self-management practices of 63 expert and novice self-trackers resulting in the Conceptual Model of Personal Health Informatics (CoMPHI) depicting the process of chronic illness self-management in the person’s social context. Through insight into users’ practices, we as technology designers and implementers can improve the quality of chronic illness self-management support, better informing and engaging people with chronic illness.

Methods and Analysis

We collected and analyzed the processes described by people managing chronic illness from three cohorts:

- Video transcripts of 29 Quantified Self (QS) show and tell videos (www.quantifiedself.com) depicting experiences with managing chronic illness from a cohort of 23 people
- Interview transcripts from a cohort of 20 adults with Type 2 diabetes
- Interview transcripts from a cohort of 20 mothers of children with asthma

Videos were freely available and did not require IRB approval, and use of interview transcripts was approved by the Group Health IRB. Three authors coded transcripts in Atlas.ti using open coding as well as concepts from the Li et al.4 and Swan5 models and MacLeod et al.6; our analysis required the creation of additional concepts. We then reviewed the coded transcripts for themes, using an affinity diagramming method to iterate a final coding schema. We used the final schema to organize data and inform the CoMPHI.

Results

We constructed a Conceptual Model of Personal Health Informatics that is distinct from prior models of personal informatics and healthcare. The CoMPHI consists of two parts: Actors and Work. Actors are people involved in management: the Person with Chronic Illness, possible Carers, Community Members, and Clinicians. The Work in which those Actors engage includes: Information, Collection, Integration, Reflection, and Action. We found no specific order for work or transitions between types of work.

Discussion

Our analysis shows that self-management of a chronic illness is a social experience, the interpersonal nature of which is not adequately represented in current models. As a result, our primary contribution is an explicit description of the social sharing of health information while managing chronic illness. All people with chronic illness described instances of sharing the data collected, whether with health care professionals, to inform clinical actions, or with friends or colleagues, to synthesize feedback and maximize benefit from self-tracking activities. Although the Pew Internet Research Project’s survey on health tracking found that 12% of trackers track for someone else, no other model includes carers. We included carer as a main actor because he or she takes on a critical management role depending on the type of illness and needs of the person with chronic illness.
We have modified Li’s notion of stages to types of work because people with chronic illness continuously perform work in support of their health. We also changed Li’s Preparation stage to Information work.

Finally, we propose that the process is more fluid than implied by Li’s stage paradigm. The notion of work in the CoMPHI is also different from that in Swan’s model of personal health informatics. In that model, all the work rested with the individual, but we found that sharing work was key to successful self-management.

For people with chronic illness, effective self-management improves health outcomes, and new personal health informatics tools can further support self-management activities. We have an opportunity to understand how individuals achieve successful self-management, and use that knowledge to design more useful tools. Understanding how people with chronic illness and carers use tracking to successfully manage chronic illness informs the design of systems to support people outside of the clinic, where most self-management activities occur. As one person stated: “in chronic diseases, health is not created in healthcare.” She creates health when implementing therapies in life outside the clinic, and her tracking practice is critical. We recommend that designers consider interoperability, data richness, and shared work when designing health informatics tools. Most importantly, this work drives home the impossibility of extricating people with chronic illness from their social environment, even when describing their work. We must consider the social aspects of self-management when designing any health informatics tool. People employ technology to assist in managing their chronic illness, but that is only part of the puzzle. As another person said, “It’s not just technology, it’s people.”

Framework Illustration

![Figure 1: Conceptual Model of Personal Health Informatics illustrating social context.](image)

This project was supported in part by grant #R01HS021590 from the Agency for Healthcare Research & Quality.

References

The LOINC/RSNA Radiology Playbook: A unified terminology for radiology procedures

Daniel J. Vreeman, PT, DPT, MS\textsuperscript{1,2}, Kenneth C. Wang, MD, PhD\textsuperscript{3}, Chris Carr\textsuperscript{4}, Beverly Collins PhD\textsuperscript{4}, Swapna Abhyankar, MD\textsuperscript{1}, Jamalynne Deckard, MS\textsuperscript{1}, Clement J. McDonald, MD\textsuperscript{5}, Daniel Rubin, MD, MS\textsuperscript{6,7}, Curtis P. Langlotz, MD, PhD\textsuperscript{6}

\textsuperscript{1}Regenstrief Institute, Inc, Indianapolis, IN; \textsuperscript{2}Indiana University School of Medicine, Indianapolis, IN; \textsuperscript{3}VA Maryland Health Care System, Baltimore, MD; University of Maryland School of Medicine, Baltimore, MD; \textsuperscript{4}Radiological Society of North America, Chicago, Illinois; \textsuperscript{5}National Library of Medicine, National Institutes of Health, Bethesda, Maryland; \textsuperscript{6}Department of Radiology and \textsuperscript{7}Biomedical Data Science, Stanford University, Stanford, California

Introduction

Despite advances in biomedical imaging technology and growing adoption of electronic health records (EHRs)\textsuperscript{1}, imaging information often lacks interoperability across systems.\textsuperscript{2} Currently, each radiology facility typically still creates its own nomenclature to identify its imaging procedures,\textsuperscript{3} which hinders data exchange and aggregation.

To address these problems, both the Regenstrief Institute, Inc. and the Radiological Society of North America (RSNA) initially undertook separate efforts to create standardized radiology procedure names. Regenstrief expanded and enriched LOINC\textsuperscript{4} to include radiology terms.\textsuperscript{5} RSNA developed the RadLex\textsuperscript{5} Playbook\textsuperscript{5}, an outgrowth of its RadiLex terminology that focuses specifically on procedures names. The organizations were on parallel paths, but shared common perspectives. With funding from the National Institute of Biomedical Imaging and Bioengineering (NIBIB), in 2013 the organizations began to unify their standards. The goal of this joint effort is to provide a common information model and unified terminology with a single governance process for radiology orders and results. Here we describe our unification approach for the jointly developed LOINC/RSNA Radiology Playbook.

Methods

In 2011, a subcommittee of the RSNA RadLex project began the RadLex Playbook. RadLex is a comprehensive lexicon of radiology terms for standardized indexing and retrieval of radiology information resources.\textsuperscript{6} The RadLex Playbook\textsuperscript{5} aims to provide a standard system for naming radiology procedures, based on the elements which define an imaging exam, and has two components: 1) the model used to generate names (the grammar), and 2) the names generated from that grammar. The RadLex Playbook grammar describes how to create pre-coordinated terms, with each Playbook term being comprised of several RadLex terms. Each unique combination of RadLex terms makes a unique Playbook term and is given a unique identifier (the RadLex Playbook ID, or RPID). Thus, for each RPID there is a corresponding set of associated RadLex clinical terms.

LOINC is a freely available international standard for health measurements, observations, and documents. Today, LOINC is used in more than 166 countries by many kinds of organizations. It is a national standard in more than 30 countries, including the U.S.\textsuperscript{6} LOINC began representing radiology procedures in 2000 (Version 1.0O). The current release (version 2.61) contains more than 5,600 active radiology terms. LOINC names radiology procedures within the six major, and up to four minor, attributes of the LOINC concept model. Within these LOINC attributes, some specialized conventions were developed for radiology. Prior analyses\textsuperscript{5} have demonstrated that LOINC has good coverage (91-92\%) of the procedure codes found in local radiology systems.

In 2013, the two organizations agreed to cooperate on a single standard for radiology procedure names, bringing together the globally-recognized terminology experience of Regenstrief and the preeminent domain expertise of the RSNA. The unified, jointly developed terminology is called the LOINC/RSNA Radiology Playbook.

The organizations agreed that LOINC codes would be the primary identifiers for procedures in the unified terminology. We would map RadLex Playbook identifiers to corresponding LOINC codes, but new RadLex Playbook identifiers would not be created in the future. The organizations developed a unification plan with these objectives: 1) develop a unified model for radiology procedure names that represents the attributes with an extensible set of values, 2) transform existing LOINC procedure codes into the unified model representation, 3)
create a mapping between all the attribute values used in the unified model as coded in LOINC (i.e., LOINC Parts) and their equivalent concepts in RadLex, 4) create a mapping between the existing procedure codes in the Core Playbook (a subset of the RadLex Playbook) and the corresponding codes in LOINC, 5) develop a single integrated governance process for managing the unified terminology, and 6) publicly distribute the terminology artifacts.

Results

The first phase of the unification spanned two years ending September 2015. We created a unified model and governance process for the collaboration, and implemented the unified model and mappings for computed tomography (CT). We defined a release format, including mappings between LOINC and the RSNA Playbook at the procedure code level, and connections between procedure terms and their attribute values expressed as LOINC Parts and RadLex IDs. In December 2015, we published the first version along with a Users’ Guide.

The second phase spans from October 2015 through September 2017. We expanded the modeling and mapping, and have iteratively improved the unified model and its application. The current version (June 2017) contains unified modeling of 5,500+ terms covering all imaging modalities. Table 1 shows the attribute values from the unified model for an example term. Other example names include [42274-1] CT Abdomen and Pelvis WO and W contrast IV, [39077-3] XR Shoulder AP and transthoracic, and [82800-4] PET+CT Heart W contrast IV.

Table 1. Example of attributes and mappings for LOINC (36244-2) MR Prostate W contrast IV

<table>
<thead>
<tr>
<th>Attribute</th>
<th>LOINC Part</th>
<th>LOINC Part Name</th>
<th>RID</th>
<th>RadLex Display Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Region Imaged</td>
<td>LP199998-8</td>
<td>Pelvis</td>
<td>RID2507</td>
<td>Pelvis</td>
</tr>
<tr>
<td>Imaging Focus</td>
<td>LP200001-8</td>
<td>Prostate</td>
<td>RID343</td>
<td>Prostate</td>
</tr>
<tr>
<td>Modality type</td>
<td>LP206549-0</td>
<td>MR</td>
<td>RID10312</td>
<td>Magnetic resonance imaging</td>
</tr>
<tr>
<td>Timing</td>
<td>LP200088-5</td>
<td>W</td>
<td>RID49853</td>
<td>W</td>
</tr>
<tr>
<td>Pharmaceutical.Substance given</td>
<td>LP200085-1</td>
<td>Contrast</td>
<td>RID11582</td>
<td>Contrast agent</td>
</tr>
<tr>
<td>Pharmaceutical.Route</td>
<td>LP200078-6</td>
<td>IV</td>
<td>RID11160</td>
<td>Intravenous</td>
</tr>
</tbody>
</table>

For these 5,500+ terms, there are over 37,000 attribute value relationships linking more than 830 LOINC Parts to RadLex clinical terms. Before Phase II ends, we will complete modeling of the last remaining content: modality-agnostic interventional procedures. As Phase II ends, the collaboration will transition into an ongoing, sustained activity that is jointly maintained and governed by the two organizations. Users can requests additions to the terminology through the existing LOINC submission process.8

Discussion

The Regenstrief Institute and the RSNA have created a unified terminology standard for radiology procedures that builds on the strengths of LOINC and the RadLex Playbook. The approach to developing this freely available standard is novel because it represents a true unification, rather than independent efforts tenuously joined by an ongoing mapping effort. The Office of the National Coordinator for Health IT lists the unified LOINC content as the standard for imaging procedures in its 2017 Interoperability Standards Advisory. As the remaining content is modeled, the LOINC/RSNA Radiology Playbook will become a complete resource of imaging procedures.

Conclusion

The LOINC/RSNA Radiology Playbook provides a universal terminology standard for radiology orders and results.

References

4. Vreeman DJ, McDonald CJ. Automated mapping of local radiology terms to LOINC. Proc AMIA Symp. 2005;769–773. PMID: 16779144
Impact of Clinician Experience on Machine Learned Clinical Order Patterns

Jason K. Wang,1 Alejandro Schuler,2 Nigam H. Shah, MBBS, PhD,2 Jonathan H. Chen, MD, PhD3

1Mathematical & Computational Science Program, Stanford University, Stanford, CA, USA; 2Center for Biomedical Informatics Research, Stanford University, Stanford, CA, USA; 3Department of Medicine, Stanford University, Stanford, CA, USA

Introduction

Clinical decision-support (CDS) aims to reinforce best-practices by distributing knowledge-based content through order sets, templates, alerts, and prognosis scoring systems. The current standard for developing and distributing content is from the top-down (e.g. human-authored order sets curated by hospital committees). As such, one of the “grand challenges” in CDS is to automatically generate content by data-mining clinical data sources from the bottom-up. We previously developed a data-driven recommender system for clinical orders (e.g. lab tests, medications, imaging) drawing inspiration from Netflix and Amazon.com’s “customers who bought A also bought B” system. Our engine dynamically generates order recommendations based on real-world clinical practice patterns represented in electronic health record (EHR) data. However, effective medical decision making and patient outcomes may be compromised if patterns are learned from less experienced fellows, residents, and medical students administering care. This phenomenon gives rise to the concern over learning indiscriminately from the “wisdom of the crowd” when the crowd consists of both experienced attending physicians and trainees. In this study, we investigate how clinical orders learned from board-certified physicians alone versus physician-trainee teams can influence the quality of learned clinical order patterns.

Methods

We extracted deidentified patient data from the (Epic) electronic medical record for inpatient hospitalizations in 2013 at Stanford University Medical Center via the STRIDE clinical data warehouse. Using patient provider information, we prepared two cohorts of general medicine services: 1) patients seen by a private attending-only service (n=1774) and 2) patients cared for by physician-trainee teams through the university teaching service (n=3404). To minimize biases due to confounding features, we conducted propensity score matching to balance the two patient cohorts. Using demographic data, initial vital signs recorded before the onset of care, and existing diagnoses as covariates, we applied a logistic regression model to compute the probability (p) of each patient’s assignment to the attending-only cohort. The propensity score is then defined as the logit function \( \log \frac{p}{1-p} \). Caliper matching on the propensity score at a threshold of 0.25 resulted in balanced cohorts of 1530 patients each, corresponding to >1.7 million and > 2.3 million recorded clinical data elements for the attending-only and teaching service cohorts, respectively.

Given a dataset of patient encounters, our previously described clinical order recommender algorithm counts co-occurrences for all clinical item pairs occurring within 24 hours by association rule episode mining. These counts are then used to populate 2x2 contingency tables to compute association statistics such as baseline prevalence, positive predictive value (PPV), relative risk (RR), and P-value for each pair of clinical items. For a given query item (e.g. admission diagnosis), we generate a list of clinical item suggestions ranked by a specified association statistic. We trained two distinct models either using patient encounters from the attending-only or the teaching service cohorts. We generated a predicted order list ranked by PPV from each model for admission diagnoses including pneumonia (ICD9: 486) and gastrointestinal (GI) hemorrhage (ICD9: 578.9). To develop an external reference standard, we curated “gold standard order sets” based on clinical practice guidelines available from the National Guideline Clearinghouse (www.guideline.gov) and PubMed. Orders were included in the reference if they were mentioned or heavily implied as appropriate to consider in the guideline text for general hospital admissions for the respective diagnosis. PPV-ranked predicted order lists were compared against guideline reference orders by receiver operating characteristic (ROC) analysis.

Results

The post-matching standardized difference in means and p-values computed using two-sample t-tests were < 0.1 and > 0.15, respectively, across all covariates, demonstrating a statistically insignificant difference between balanced
attending-only and teaching service patient cohorts. Table 1 shows an example of the clinical orders generated by our recommender engine for an admission diagnosis of pneumonia trained separately on each patient cohort. Figure 1 shows plots generated for two admission diagnoses: pneumonia and GI hemorrhage. Each plot reports 3 order lists compared against the external reference standard: attending-only predicted orders, teaching service predicted orders, and a pre-authored order set curated by Stanford Hospital. Pre-authored order sets have no inherent ranking to convey relative importance and are thus depicted as a single discrete point on the ROC curve. Area-under-curve (AUC) is reported as c-statistics with 95% confidence intervals empirically estimated by bootstrap sampling items with replacement 1000 times.

**Table 1.** Top clinical order associations for pneumonia generated from attending-only and teaching service trained models sorted by P-value calculated by Yates’ chi-squared statistic, alongside association statistics and 95% confidence intervals.

**Figure 1.** ROC plots for pneumonia (left) and GI hemorrhage (right). Each plot compares a hospital-authored order set and automated predictions from attending-only and teaching service models against the external reference standard.

**Discussion**

While the specific association statistics vary, the overall top clinical order recommendations generated from association models trained on attending-only vs. teaching service cohorts yield comparable c-statistics (AUC) against the external reference standard. This supports the argument that aggregating clinical order patterns of less experienced clinicians (e.g. resident trainees) and more experienced providers (e.g. private attendings) will converge towards comparable top results as both cohorts share the common end goal of patient care. Although propensity score matching was conducted to minimize the influence of confounding covariates recorded in EHR data, study limitations include results influenced by undocumented biases and the necessity of interpreting practice guidelines to define reference standards. When trainees administer care, they may simply follow university attending directions or hospital-approved order sets. Thus, a potential risk of interpretation is recapitulation of existing order sets, though usage rates are relatively low (15% of all orders) and clinical items derived from order sets do not significantly change our results when excluded from analysis. This study provides insights into the consistency of clinical order patterns learned from historical data and how their quality is influenced by underlying practitioner experience.

**References**

Interactive Medical Word Sense Disambiguation with Instance and Feature Labeling

Yue Wang, MS¹, Kai Zheng, PhD², Hua Xu, PhD³, Qiaozhu Mei, PhD¹,¹
¹Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI; ²Department of Informatics, University of California, Irvine, CA; ³School of Biomedical Informatics, University of Texas Health Science Center, Houston, TX; ⁴School of Information, University of Michigan, Ann Arbor, MI

Introduction

Medical documents contain many ambiguous words. Word sense disambiguation (WSD), i.e. assigning the appropriate meaning to an ambiguous word in context, is a critical step for many medical natural language processing applications, such as named entity extraction and computer-assisted review. Previous works have proposed many approaches for medical WSD, including unsupervised learning, supervised learning, semi-supervised learning, active learning, and interactive search and classification. However, these approaches still lack the ability to learn from domain knowledge, and take relatively long time to reach a reasonable performance, which is known as the “cold-start” problem. In this study, we design a novel interactive learning algorithm that directly incorporates expert knowledge into the WSD model training process. We consider two types of expert knowledge in the WSD task: prior knowledge and reasoning process. The expanded form of many ambiguous abbreviations are documented in medical ontologies, and the expert may know contextual words of a particular sense before looking at any instance, both of which are prior knowledge. Upon seeing an instance containing an ambiguous word, the expert can pinpoint the words and phrases that support his decision on the word sense, which is the reasoning process hidden behind the label. In the new interactive learning process, an expert can express knowledge as the association between senses and textual patterns, and the machine learning algorithm can directly learn from such association. Experiments on one biomedical literature corpus and two clinical notes corpora show that the proposed algorithm makes better use of human efforts in training WSD models than all previous approaches, achieving the state-of-the-art performance with the least effort from domain experts.

Methods

The proposed algorithm has an interactive learning component and two computational components (Figure 1A).

1) Labeling instances and features. In the labeling process, a domain expert can come up with informative contextual words, label instances, and highlight informative text snippets as “rationales” behind the label decision. As a result, the expert provides two types of supervision: instance labels and feature labels. An instance label is assigned to an individual instance containing an ambiguous word, as in traditional machine learning. A feature label is assigned to a word, a phrase, or a textual pattern, all of which can be viewed as features. When the expert provides informative words for a label (sense), these words are explicitly associated with the label. When the expert highlights words in an instance, the highlighted words are implicitly associated with the label of that instance.

2) Learning from labeled instances and features. The machine learning module takes in both instance labels and feature labels from (1) to train a WSD classifier. At each iteration, labeled features may contain arbitrary textual patterns (e.g. n-grams) that are hard to know beforehand. We start with a set of base features (unigrams) and dynamically expand the feature set as new features are provided/highlighted. We use logistic regression with a linear kernel as the WSD classifier. We train the classifier by adding a new term to the original logistic regression loss function, such that any instance containing a labeled feature is more likely to bear that label. A labeled feature may carry substantially more information than a labeled instance if the feature appears in many instances.

3) Selecting instances. An instance selector picks up new instances from the unlabeled pool, presents them to the expert, and asks for labels. We use the margin active learning algorithm to select uncertain instances. To evaluate the new algorithm, we compare it with several baseline methods, including random sampling, margin active learning¹, and ReQ-ReC expert². We use two settings of the proposed algorithm: providing labeled features and highlighting informative features in labeled instances. To make head-to-head comparison with a previous work², we use the same evaluation settings. We simulate an expert by leveraging benchmark medical WSD corpora, including MSH (biomedical literature), UMN, and VUH (clinical notes). Labeled instances are sense-tagged examples in these corpora, and labeled features are n-grams (n = 1, 2, 3) with high information gain for an ambiguous word. In all experiments, the expert provides only one labeled feature at the beginning; all subsequent labels are instance labels. To make realistic assumption on feature labeling, we simulate experts that provide or highlight the k-th best feature.
Results

Figure 1B, 1C, and 1D show the learning curves of different algorithms on three WSD corpora. The new interactive learning algorithm outperforms strong baseline methods. On biomedical literature WSD corpus (MSH), the new algorithm achieves 90% accuracy with 15 labels, saving 40% of labeling effort compared to active learning. WSD in clinical notes is more difficult. On UMN corpus, the new algorithm achieves 90% accuracy with 15 labels, saving 35% of labeling effort compared to active learning. On VUH corpus, the new algorithm saves labels at the beginning.

Discussion

The proposed approach effectively handles the “cold-start” problem in active learning. Active learning works best when the model has a reasonably good “understanding” of the problem space so that the selected instances are the most informative. At the beginning, the model trained on very few labeled instances can perform poorly. In the new learning process, human experts can kick off training by providing domain knowledge, giving the model a warm start.

References

Real-Time Queries of Millions of Patients Using Probabilistic Sketches

Griffin M Weber, MD, PhD\textsuperscript{1,2}
\textsuperscript{1}Harvard Medical School, Boston, MA;
\textsuperscript{2}Beth Israel Deaconess Medical Center, Boston, MA

Abstract

Large clinical research databases provide investigators access to millions of patients. However, queries often take minutes or even hours to complete. Although this is sufficient for many use cases, it creates a barrier to true exploratory analysis that would be possible with a real-time system. In this study, we show how “streaming algorithms”, which transform Big Data into small probabilistic data structures called “sketches”, can produce sub-second query times with less than 1% error.

Introduction

Sketching algorithms, which are also called “streaming” or “sub-linear space” algorithms, represent big data as small probabilistic data structures. They are extremely efficient, requiring only a small amount of data storage, and producing fast accurate estimates of cardinality (number of distinct items), quantiles, frequency distributions, and other properties of a dataset. They were first invented in the 1980s and are widely used today by search engines like Google, internet routers, and many other applications [Flajolet 1985, Heule 2013]. However, they have had little attention so far from the biomedical community, other than in a few genomics use cases.

An important aspect of sketching algorithms is that the data only have to be read once. In other words, the data are treated as a stream of elements that are processed one at a time and then discarded. A simple example of how a sketch can be used to determine the number of distinct items in a stream is as follows: For each element in the stream, use a hash function to convert it into a random value, \(H\), between 0 and 1. As each element is processed, keep track of the smallest hash value you see: \(S = \min(H)\). The only thing that needs to be stored is the one value \(S\) (the sketch), regardless of the size of the stream. The more distinct items in the stream, the more likely it is to have some element that gets hashed to a very small value. In other words, on average, the smaller \(S\) is, the larger the cardinality. Precisely, the number of distinct items can be estimated as \(N = (1/S) - 1\). This is illustrated in Figure 1. With two items, on average, one will get hashed to value of about \(S = 1/3\); with three items, \(S = 1/4\); and, with ten items, \(S = 1/10\).

![Figure 1. Cardinality Estimates Using a Random Hash Function.](image)

The problem with this particular algorithm is that although \(S\), on average, accurately estimates the cardinality, the variance is so large that the estimate is not very useful. For example, an element in a small dataset can always get hashed to an extremely low value by random chance, causing the cardinality to be grossly overestimated. The solution is to use multiple hash functions and keep track of the smallest value from each hash function in the sketch. The average of those values provides a much better estimate. A popular algorithm, called HyperLogLog (HLL), introduced an efficient way of doing this, where only a small number of values are needed to generate a highly accurate cardinality estimate [Flajolet 2007]. For example, only 64 integers results in just a 10% expected error. Larger sketches can be used for smaller error if needed.

A nice feature of this approach to estimating cardinality is that the result is the same if all the data exist in one dataset/stream or if the data are distributed across many datasets/streams. In the distributed version of the algorithm, a sketch is created for each dataset. The sketches are combined by determining the smallest \(S\) across all sketches. That smallest \(S\) can be used to estimate the cardinality of the entire distributed dataset.

Several enhancements and extensions to HLL have been proposed. For example, Google modified the algorithm to support high cardinalities (HyperLogLog++) [Heule 2013]. AdRoll showed how HLL combined with another streaming algorithm MinHash can generate accurate estimates the cardinality of the union and intersection of multiple datasets (in a manner similar to how distributed HLL sketches are combined) [Pascoe 2013].

This presentation will provide a brief overview of these algorithms, show how they can be adapted for queries of clinical research databases, and give specific examples using a large nationwide administrative claims database.
Methods
Informatics for Integrating Biology & the Bedside (i2b2) is an example of a widely used software platform for query and analysis of clinical repositories [Murphy 2010]. i2b2 users can define an arbitrary Boolean query (e.g., inclusion and exclusion criteria) through a web-based interface, and the software will return the number of distinct patients who match the query. In i2b2, data (diagnoses, medications, laboratory tests, etc.) are primarily stored in a single indexed fact table, which can have billions of records at large institutions. The records include a concept code (e.g., ICD-9 diagnosis), a patient number, an observation date, and optional metadata, such as a laboratory test result.

To create a probabilistic representation of the i2b2 fact table, all the facts associated with a given concept, like hypertension, can be thought of as a stream of patients. HLL is used to hash the patient numbers into a small sketch. That sketch can then be used to estimate the number of distinct patients with that concept. Similarly, a sketch is made for every medical concept. A typical i2b2 instance might have a few hundred thousand concepts; though the amount of storage needed for the sketch is so small that hundreds of thousands of them are still small compared to the size of the fact table. Boolean queries are implemented as unions and intersections of the sketches. Dates and laboratory test results are stored as additional data structures linked to the concept sketches.

The approach was applied to a large claims database. First, 3 billion diagnoses (64 thousand separate ICD-9 and ICD-10 diagnosis codes) for 51 million members were loaded into an i2b2-like fact table in a Microsoft SQL Server 2012 database. Unlike a regular i2b2 fact table, non-essential columns and indexes were removed, the table was compressed, and a columnstore index (a special index for large tables was added). This basically created an optimized version of the traditional way clinical data are stored. Next, the table was transformed into probabilistic sketches, with each sketch containing $2^{15}=32,768$ values. In practice, smaller sketches are often used, but this creates a worst-case scenario in terms of performance and reduces the expected error to under 1% relative to the actual number of patients who match a query.

Results
The optimized diagnosis fact table contained 3 billion rows and used 160 GB of disk space. The sketch version contained 3.8 million rows and used 1.2 GB of disk space. The equivalent query can run orders of magnitude faster as sketches. Two examples: (1) The query “Hypertension (ICD9 401.9) OR Diabetes (ICD9 250.00)” returns 8,293,530 distinct members in 6.6 seconds in the fact table compared to 8,309,356 (0.19% error) in 33 milliseconds as sketches. (2) The query “Hypertension AND Diabetes” returns 2,131,928 members in 7.5 seconds compared to 2,150,878 (0.89% error) in 100 milliseconds as sketches. More complex queries in a standard i2b2 fact table can require far more time to run, while queries using sketches remain fast because they scale linearly with respect to the number of concepts in the query (not the number of patients).

Conclusion
Research queries run on probabilistic representations of large clinical data repositories can scale to tens of millions of patients, while retaining real-time sub-second performance. The method was demonstrated using a single nationwide claims database. However, the distributed version of the algorithm could perform equally well as a privacy-preserving approach to queries in a federated network of multiple hospitals or payors since only the sketches, not the actual lists of patients, would need to be shared.

References

This work was supported by NIH Big Data to Knowledge (BD2K) grants U01CA198934 and U54HG007963.
Clinical code set engineering for reusing EHR data for research: A review
Richard Williams, MA Cantab1,2, Evangelos Kontopantelis, MSc, PhD1, Iain Buchan, MD, FACMI1,2, Niels Peek, MSc, PhD1,2
1MRC Health eResearch Centre; 2NIHR Greater Manchester Primary Care Patient Safety Translational Research Centre, University of Manchester, Manchester, UK

Introduction
The rapid adoption of electronic health records (EHRs) is creating unprecedented opportunities for studying clinical medicine, health services, and population health1–2. Adoption of EHRs has more than doubled in the US over the last decade and is expected approach 100% in 20203. UK primary care providers switched to using EHRs in the 1990s, and this has made large EHR databases available for research4–5. The re-use of EHR data for research provides large datasets with long follow-up times against low costs6–8. However, there are considerable methodological challenges because raw EHR data are not "research ready": they need to be transformed before they can be meaningfully used for research9. Most studies that re-use EHR data for research focus on the structured part of the EHR, using clinical codes to extract meaningful information from the raw EHR data. Clinical code set construction is the process of assembling a set of clinical codes that represent a single clinical concept such as a diagnosis, a procedure, an observation or a medication. Once constructed the set is then used to query and extract data from an EHR database, for use in further analysis. These sets of codes, variously referred to as "code lists", "clinical code lists", "code sets" and "value sets", are one of the main building blocks for creating the database queries. It is frequently an early step in research, and arguably a hazardous one, as errors introduced here of missing or wrongly specified codes could result in selection biases that propagate throughout subsequent analyses having a major impact10. However, several reviews have shown that code sets are rarely included in applied papers11–12, let alone the process by which the code sets have been constructed. The construction process is perhaps of greater significance than the set itself as the process could, and should, be scrutinised by reviewers, and can form the basis for other researchers to reuse their methods rather than taking a crude set on trust. We aimed to review and compare methods and tools for managing (constructing, validating, sharing, and reusing) sets of clinical codes and to develop recommendations.

Methods
A PubMed search was performed in August 2016 by the lead author for papers with a substantial methodological component using the terms: "code list(s)", "code set(s)", "value set(s)", "list of codes", "set of codes". The list was supplemented with papers identified by searching citations of relevant material, via snowball sampling using PubMed and Google Scholar. Papers were excluded if they were not related to clinical code sets or not written in English. In total, 507 papers were discovered from the initial searches with 502 screened after 5 duplicates were removed. 464 papers were rejected for lack of relevance (435 from the title and 29 from the abstract). The snowball sampling commenced from the 38 full papers read and in turn this identified 31 additional papers (initial list of 183 papers was reduced by removing 26 duplicates, 109 rejected from the title and 17 rejected from the abstract). Out of the 69 full papers reviewed, 39 were rejected as not relevant. This review is based on the remaining 30 papers.

Results
Although differences existed between the methods described, common themes emerged. A popular approach was to reuse an existing code set (n=21) from: a previous study (n=5); a national clinical quality management scheme such as the Value Set Authority Centre (VSAC) (n=11); or both (n=4). The reused set was almost always updated or extended (n=20). Authors reported some specific strategies for searching for relevant codes: exploiting the hierarchical nature of coding terminologies (n=23); preparing an initial list of synonyms to search for (n=20); and employing an iterative approach after preliminary searches (n=13). The putative sets were usually reviewed (n=26), mostly with clinician input (n=20), before definitive use. There were calls for openness and sharing of code sets and code set management methods (n=14), with some giving suggestions or platforms for sharing (n=8). The need for sensitivity analysis (n=19) and caution due to the dynamic nature of code sets (n=13) were also mentioned. Seven papers described software to support the selection of code sets and a further two suggested features for such tools.

Discussion
The process of constructing clinical code sets is time-consuming and error-prone. This review has identified and analysed the code selection methods that are commonly reported. However, despite the existence of relevant software
tools, their use is seldom reported, suggesting they are underused. Potential barriers to their uptake might be: lack of awareness of their existence; ignorance of their necessity; or deficiencies in the tool themselves, either in functionality or that they are time consuming to use. Researchers frequently modify existing code sets which is likely time consuming and error-prone. Future work is needed to reduce this effort. We propose the following recommendations: the objects reviewed in this paper should be called “clinical code sets” or simply “code sets”; journals should insist that code sets for retrospective studies on routinely collected data are published alongside the article; sensitivity analyses with multiple codes sets should be performed; and a clinician should help to create the initial list of synonyms and review the final code set. Code set construction tools should: be open source, publicly available and easily accessible; have minimal setup to facilitate widespread adoption; make use of the hierarchy of clinical dictionaries to assist in the searching and selection of codes; facilitate the initial construction of a list of synonyms; facilitate an iterative process by suggesting additional synonyms based on codes discovered by searching the code hierarchy that are not found themselves directly; facilitate the various stages of the review process from the selection of the initial list of synonyms, to the review of the included and excluded codes; make this process as simple and quick as possible; facilitate and encourage the creation of multiple sets of codes for sensitivity analyses; record metadata such as: the initial list of synonyms, the excluded codes, the purpose for the set and the author; and facilitate the reuse, validation and sharing of codes sets, not simply their construction. Code set sharing platforms should: be discoverable, maintained indefinitely and support versioning; and support the storage of metadata alongside the code set. Online code repositories such as github.com and bitbucket.org could be used to store code sets and their associated metadata.

Conclusion

Research using healthcare databases could be improved through the further development, more widespread use and routine reporting of the methods by which clinical codes were selected.

Acknowledgements: Funded by the National Institute for Health Research Greater Manchester Primary Care Patient Safety Translational Research Centre (NIHR GM PSTRC). The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health. MRC Health eResearch Centre grant MR/K006665/1 supported the time and facilities of EK, IB and NP.

References

Development and Evaluation of a Novel User Interface for Reviewing Clinical Microbiology Results

Adam Wright, PhD\textsuperscript{1,2,3}, Pamela M. Neri, MS\textsuperscript{3}, Skye Aaron\textsuperscript{1}, Thu-Trang T. Hickman, MPH\textsuperscript{1}, Francine L. Maloney, MPH\textsuperscript{1}, Daniel A. Solomon, MD\textsuperscript{3}, Dustin McEvoy\textsuperscript{3}, Angela Ai\textsuperscript{1}, Kevin Kron\textsuperscript{3}, Gianna Zuccotti, MD\textsuperscript{1,2,3}

\textsuperscript{1}Brigham & Women’s Hospital, Boston, MA; \textsuperscript{2}Harvard Medical School, Boston, MA; \textsuperscript{3}Partners HealthCare, Boston, MA; \textsuperscript{4}Ariadne Labs at Brigham and Women’s Hospital and Harvard Chan School of Public Health, Boston, MA; \textsuperscript{5}Mount Auburn Hospital, Cambridge, MA

\textbf{Introduction:} Microbiology laboratory data is essential to the diagnosis and treatment of infectious diseases. Accurate, timely, and complete microbiology results guide selection of antimicrobial therapy, enable detection of disease outbreaks, and can help prevent the development and spread of antibiotic-resistant microbes.\textsuperscript{[1]} However, compared with other types of laboratory data, microbiology data are unusually complex and can be cumbersome to review and interpret.\textsuperscript{[2]} Integrating structured microbiology results into EHRs has been hindered by the variability of test results, their inherent temporal nature and mix of qualitative and quantitative findings and the numerous data elements that can be returned for each test. As such, microbiology results are often loaded into EHRs as unstructured documents, with limited or no automated capabilities for sorting and filtering results. We sought to develop a new review tool to improve the ease and accuracy of microbiology results review.

\textbf{Methods:}Clinicians were observed using existing microbiology results tools and informally interviewed to determine areas in which existing tools were lacking. Based on this user input, a new tool was developed that reorganizes infectious disease and microbiology laboratory data by time and organism. A scenario-based evaluation of the new tool was conducted, with comparison to existing legacy tools and using a balanced block design to control for scenario difficulty and learning effects. For each scenario, accuracy, completion time and ease-of-use (using the Single Ease Question) were assessed. Additionally, participants were surveyed using the System Usability Scale to evaluate the usability of the new tool.

\textbf{Results:} The average time-on-task to complete all 8 scenarios decreased from 45.3 minutes for the legacy tools to 27.1 minutes for the new tool (P < 0.0001). Users made a total of 41 errors with the legacy tools and 19 errors using the new tool (P = 0.0068). The average Single Ease Question (SEQ) score was 5.65 (out of 7) for the new tool, compared to 3.78 for the old tool (P < 0.0001). As shown in Table 1, the average time-on-task, accuracy, and SEQ score stayed the same or improved for each metric in all 8 scenarios. The overall System Usability Scale score was 88 for the new tool, which is “Excellent.”

\textbf{Table 1.} Comparison between the legacy tools and the MV are shown for three metrics: time-on-task, accuracy, and SEQ score. For each metric, a pooled analysis across all scenarios is also shown in the last row. The SEQ score was measured in two parts for scenario 3, and these scores are shown separately.

<table>
<thead>
<tr>
<th>Task</th>
<th>Average Time-On-Task (Minutes)</th>
<th>Average Accuracy (Out of 100%)</th>
<th>Average SEQ Score (1=Very Difficult, 7=Very Easy)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Legacy Tool</td>
<td>Micro Viewer</td>
<td>Legacy Tool</td>
</tr>
<tr>
<td>Scenario 1</td>
<td>4.06</td>
<td>3.27</td>
<td>100%</td>
</tr>
<tr>
<td>Scenario 2</td>
<td>8.08</td>
<td>5.12</td>
<td>100%</td>
</tr>
<tr>
<td>Scenario 3</td>
<td>8.05</td>
<td>4.39***</td>
<td>63%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scenario 4</td>
<td>4.68</td>
<td>2.74**</td>
<td>100%</td>
</tr>
<tr>
<td>Scenario 5</td>
<td>5.44</td>
<td>2.16***</td>
<td>94%</td>
</tr>
<tr>
<td>Scenario</td>
<td>Score</td>
<td>p-Value</td>
<td>Error Rate</td>
</tr>
<tr>
<td>-------------</td>
<td>-------</td>
<td>---------</td>
<td>------------</td>
</tr>
<tr>
<td>Scenario 6</td>
<td>4.89</td>
<td>2.24***</td>
<td>75%</td>
</tr>
<tr>
<td>Scenario 7</td>
<td>6.10</td>
<td>4.97</td>
<td>63%</td>
</tr>
<tr>
<td>Scenario 8</td>
<td>3.96</td>
<td>2.24*</td>
<td>71%</td>
</tr>
<tr>
<td>Combined</td>
<td>45.26</td>
<td>27.13***</td>
<td>82%</td>
</tr>
</tbody>
</table>

*Difference between tools is significantly non-zero at: *p < 0.05, **p < 0.01, ***p < 0.001*

**Discussion:** Our results strongly suggest that the new tool represents a substantial improvement compared to existing tools. Compared to legacy tools, users were more satisfied with the new tool, completed clinical scenarios much faster, and made fewer errors. The design and results of our study also have an important implication for system architectures. We built the new tool entirely outside of our legacy EHR systems and used web services to fetch coded microbiology and infectious disease laboratory data from our clinical data repository. We also used web services to support user authentication and access control, and to enable patient search and patient list functionality. Our ability to do this work was enabled by Partners’ significant investment in service-oriented architecture[3] and provides a model for how clinical and informatics teams can innovative alongside a commercial EHR. After the completion of the study, the tool was rebuilt to enterprise standards and deployed throughout Partners healthcare.

**Conclusion:** The new tool reduces time required to review microbiology results, increases accuracy, and is considered useful and usable by infectious disease clinicians.

**Figure 1:** Specimen view in microbiology result viewer

**References:**
Detecting Body Location Modifiers of Disorders in Clinical Texts via Sequence Labeling

Jun Xu, Ph.D., Yonghui Wu, Ph.D., Yaoyun Zhang, Ph.D., Hua Xu, Ph.D.
School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX

INTRODUCTION
Body location is an important type of modifier for disorders and it is critical for clinical natural language processing (NLP) systems to capture this information. It is a typical relation extraction task in NLP research. However, body location and disorder pairs could be complicated in clinical text. As shown in Figure 1, the sentence contains two disorders: ‘intra cranial hemorrhage’ and ‘fractures’ and two body location modifiers: ‘intra cranial’ and ‘head’. An NLP system needs to determine that ‘intra cranial’ is an inside body location modifier (BLM) for ‘intra cranial hemorrhage’ and ‘head’ is the BLM for ‘fractures’. If we assume that disorder entities are given, traditional methods for identifying BLM relations typically consist of two steps: 1) recognition of all body location entities (BLE) and 2) classification of possible BLM relations. In this study, we propose a new relation extraction method for identifying BLMs for disorders, which recognizes body location entities and classifies their relations with the target disorder in one step, via a sequence-labeling model. We conducted several experiments to compare our sequence labeling-based approaches with traditional two-step relation extraction methods and our results show that sequence labeling-based method achieved much better performance for identifying the BLMs of disorders than traditional methods.

MATERIAL AND METHOD
Dataset We used the ShARe corpus developed by the organizers of the SemEval 2015 challenge task 14. It consists of 531 de-identified notes. We used the training set for model development and the development set for evaluation. The training set contains 6,402 BLMs and the development set contains 3628 BLMs.

Traditional relation extraction approach (baseline): The traditional method consists of two steps to identify BLMs. 1) BLEs recognition: it is a typical NER task to recognize body location entities and we collected four anatomy-related lexicon resources: directional terms, anatomy abbreviations from the training set, body location terms from WordNet, and anatomy lexicon from UMLS. 2) Classification of BLM-disorder pairs: it can be further divided into two tasks: candidate BLM-disorder pair generation and classification. Candidate BLM-disorder pairs were generated according to heuristic rules (e.g., either within one sentence or look at the previous sentence if no BLE exists in the current sentence) and then labeled as positive or negative based on the gold standard. To classify candidate BLM-disorder pairs, we developed two methods: one was a rule-based system and the other was a machine learning-based system using the Support Vector Machines (SVMs) algorithm.

Relation extraction by sequence-labeling (RESL) approach: In our new approach, we treat BLM-disorder detection as one sequence labeling task. This is done by a new transformation schema. For each disorder entity in a sentence, we will generate a sequence labeled sample. As shown in Figure 1, the sentence has two disorders so that it will generate two labeled sentences as samples: one for the target disorder ‘intra cranial hemorrhage’ and the other for the target disorder ‘fractures’. For a given target disorder, only body locations that are modifiers of the target disorder will be labeled as ‘B/I’. For example, in the sample generated for ‘intra cranial hemorrhage’, ‘head’ is not labeled as ‘B’. After we transform each disorder (with its sentence) to samples with BIO labels, we train a single sequence-labeling model for all disorders. We use the following features: 1) Disorders (target and non-target); 2) All position, direction, discourse, context and relation information between all body location and disorder entities, which have been used for traditional approach model building; 3) The output semantic labels recognized by the dictionary lookup program and 4) Other features including word representation features learned from the unlabeled corpus. We evaluated two sequence labeling algorithms: Conditional Random Fields (CRFs) and the Structured Support Vector Machines (SSVMs), to recognize the BLMs for the target disorder.

Evaluation Following the evaluation guideline by the SemEval 2015 challenge, we used the standard precision (P), recall (R) and F1-measure (F1) under both strict and relaxed criteria. We also adopted per-disorder accuracy (Acc) to evaluate the ability of identifying the BLM for a given disorder, defined as: \( Acc = \frac{N_{\text{correct.predict}}}{N} \), where, \( N \) = total number of gold standard disorders; \( N_{\text{correct.predict}} \) = number of disorders which BLM is predicted correctly.
RESULTS
Table 1 shows the performances for our new approaches and the traditional approaches for identifying BLMs of disorders in the ShARE corpus. The RESL approach achieved much higher performance than baseline methods. RESL-SSVMs achieved the highest performance, which was over 9% higher than the SVM-based traditional method (F1 0.8213 vs. 0.7285) under the strict evaluation.

DISCUSSION and CONCLUSION
In this study, we propose a novel method for a relation extraction task – detecting BLMs of disorder. Through a new transformation schema, we combined the entity recognition and relation classification into a single sequence labeling model. Our evaluation shows that the new approach achieved much better performance than the traditional methods on the ShARE corpus. We think the performance gain may come from the combined features for both entity recognition and relation classification in one model.

References

Figure 1. An illustration of sequence labeling model training: 2 disorders correlated with their BLMs.

Table 1. The results achieved by different approaches.

<table>
<thead>
<tr>
<th>Method</th>
<th>Strict/Relax</th>
<th>Acc</th>
<th>P</th>
<th>R</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline (rule)</td>
<td>S</td>
<td>0.784</td>
<td>0.705</td>
<td>0.662</td>
<td>0.683</td>
</tr>
<tr>
<td></td>
<td>R</td>
<td>0.830</td>
<td>0.786</td>
<td>0.738</td>
<td>0.761</td>
</tr>
<tr>
<td>Baseline (SVMs)</td>
<td>S</td>
<td>0.792</td>
<td>0.774</td>
<td>0.705</td>
<td>0.728</td>
</tr>
<tr>
<td></td>
<td>R</td>
<td>0.868</td>
<td>0.861</td>
<td>0.785</td>
<td>0.821</td>
</tr>
<tr>
<td>RESL (SSVMs)</td>
<td>S</td>
<td>0.847</td>
<td>0.836</td>
<td>0.807</td>
<td>0.821</td>
</tr>
<tr>
<td></td>
<td>R</td>
<td>0.903</td>
<td>0.890</td>
<td>0.859</td>
<td>0.874</td>
</tr>
<tr>
<td>SESL (CRFs)</td>
<td>S</td>
<td>0.845</td>
<td>0.833</td>
<td>0.798</td>
<td>0.817</td>
</tr>
<tr>
<td></td>
<td>R</td>
<td>0.899</td>
<td>0.888</td>
<td>0.845</td>
<td>0.866</td>
</tr>
</tbody>
</table>
Using Logistic Regression to Verify Completeness of Electronic Health Records for Infant Mortality Analysis

Luge Yang, PhD¹, Jose D. Posada, MS¹, Lingyun Shi, MS¹, Fan Mi, MS¹, Fuchiang (Rich) Tsui, PhD¹

¹Tsui Lab, Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA

Introduction
Completeness is a frequently used concept in the field of data quality assessment. Traditional completeness concept addresses counts of missing values. When one seeks to predict a response of interest such as mortality, a newly defined concept on completeness refers to whether the dataset contains sufficient information or risk factors for the predictive task. However, the predictive completeness is not readily observable in the sense that we can't tell the risk factors without running model on dataset. Traditional measures are not able to assess implicit predictive ability of dataset because data required are those that can contribute to prediction. Further, predictiv completeness assessment fully based on physical presence of risk factors in dataset is limited, especially when research interest is associated with more than handful non-ad-hoc risk factors.

It is known that the risk factors of infant mortality include but not limited to a wide range of categories such as antepartum complications, maternal diseases, maternal demographic characteristics, maternal reproductive history, infant birth injuries, infant delivery characteristics, labor and delivery complications, infant diseases, maternal behavior, health care quality and access, and environmental exposures. Therefore, for researches aimed to predict infant mortality, it is imperative to assess data completeness before initiate further data analyses. We propose an assessment method on predictive completeness, and we verify our method on infant mortality by using Magee Obstetric Medical and Infant Database (MOMI) in Allegheny County, which is generated and maintained by the University of Pittsburgh Medical Center (UPMC) Magee-Womens Hospital (Magee). The database includes demographic and medical information for all deliveries at Magee. The datasets retrieved from the MOMI database contains 128,023 records corresponding to 117,929 deliveries of total 88,780 patients between January 1, 2002 and December 31, 2014.

Methods
MOMI dataset has no flags for infant deaths. To identify dead cases in MOMI, we linked 1,021 infant death cases in Allegheny County between 2003 and 2013 from the Allegheny County Department of Health Service (DHS) with MOMI dataset. MOMI records in 2002 and 2014 were removed from this study so that its time range was in accordance with that of DHS data. We applied direct matching using mother’s SSN, mother’s full name, mother’s date of birth and delivery date, which labeled total 496 death cases in MOMI. Further, both stillbirths and fetal deaths were excluded from MOMI dataset. In addition, we removed some features from analyses due to their data quality issues. Final dataset contained 117 predictors including 6 numeric variables and 111 categorical variables. Moreover, DHS provided causes of deaths data for 496 dead cases where death causes comprised 115 diseases. We found the dataset contained 35 causes of deaths that resulted in 281 death cases.

We modeled probability of event of infant death by logistic regression with all 117 variables as predictors. Then we ranked the p-values of regression coefficients of 117 predictors in an increasing order. We listed significant factors (p-value < 0.1) and compare them with 35 causes of deaths.

Results
15 variables in MOMI dataset are identified to be significantly related with infant death (p-value < 0.1, Table 1). The data shows the majority of deaths of infants born at Magee were due to respiratory diseases of infant, short gestation and low birth weight, genetic defects, maternal complications of pregnancy and complications of placenta cord. For assessment on predictive completeness, we found 10 risk factors among 15 are perfectly matched or closely related with 14 causes of deaths, which accounted for 84% of 281 death cases. Associations of other 5 significant risk factors with causes of deaths were not clear. Besides, the value of odds-ratio indicates an association between infant death, and each risk factor is in agreement with common knowledge.
Table 1. Significant risk factors identified in MOMI dataset by logistic regression.

<table>
<thead>
<tr>
<th>Variable Description</th>
<th>Variable Value</th>
<th>Odds Ratio</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Post-asphyxia complications</td>
<td>0 = No; 1 = Yes</td>
<td>18.4193</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>Pulmonary hypoplasia</td>
<td>0 = No; 1 = Yes</td>
<td>13.5814</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>Pneumothorax or air leaks</td>
<td>0 = No; 1 = Yes</td>
<td>25.1284</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>Uterine rupture</td>
<td>0 = No; 1 = Yes</td>
<td>14.9783</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>Obstetrical estimate of gestational age (weeks) at delivery</td>
<td>Normal range is 20-45 week</td>
<td>0.7091</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>Inborn error of metabolism</td>
<td>0 = None; 1 = Mineral</td>
<td>6.7639</td>
<td>0.0062</td>
</tr>
<tr>
<td>Maternal collagen vascular disease</td>
<td>0 = None; 1 = Other</td>
<td>11.5352</td>
<td>0.0333</td>
</tr>
<tr>
<td>Total number of pregnancies</td>
<td>Number includes the current pregnancy</td>
<td>1.0652</td>
<td>0.0475</td>
</tr>
<tr>
<td>Meconium stained amniotic fluid without fetal distress</td>
<td>0 = No; 1 = Yes</td>
<td>1.4665</td>
<td>0.0495</td>
</tr>
<tr>
<td>Maternal postpartum complications=1</td>
<td>0 = None; 1 = Pulmonary complication</td>
<td>3.0994</td>
<td>0.0604</td>
</tr>
<tr>
<td>Maternal thrombocytopenia=1</td>
<td>0 = None; 1 = Gestational</td>
<td>16.2794</td>
<td>0.0636</td>
</tr>
<tr>
<td>Placental pathology noted by Obstetrician or Neonatologist</td>
<td>0 = None; 1 = Placenta accreta, increta, or percreta with hem</td>
<td>2.3250</td>
<td>0.0649</td>
</tr>
<tr>
<td>Maternal hemoglobinopathy</td>
<td>0 = None; 1 = Other</td>
<td>1.9646</td>
<td>0.0736</td>
</tr>
<tr>
<td>Maternal postpartum complications</td>
<td>0 = None; 1 = Hemorrhage</td>
<td>0.5922</td>
<td>0.0765</td>
</tr>
<tr>
<td>Maternal upper gastrointestinal disease</td>
<td>0 = None; 1 = Gastric bypass</td>
<td>0.5023</td>
<td>0.0902</td>
</tr>
</tbody>
</table>

Discussion and Conclusions

Whether a dataset contains sufficient information to predict research interests is the most complex definition of completeness. In this sense, current assessment methods such as simple incidence rate, standardized incidence ratios (SIRs), incidence rate ratios (IRRs), and kappa statistic are not able to measure completeness especially when research interest is associated with many risk factors. We proposed a logistic regression method to examine significance and magnitude of effects of all available features in dataset on research interest. We picked significant risk factors and compared them with known causes of deaths. Our method provides a quantitative, comprehensive yet straightforward data completeness assessment in terms of predictive ability.

Acknowledgement

The project is funded by the Richard King Mellon Foundation. We thank Mr. Hoah-Der Su for de-identifying the datasets in this study.

References

High-throughput Phenotyping via Denoised Normal Mixture Transformation

Sheng Yu1,2 PhD, Yumeng Ma3, Jessica Gronsbell4, Katherine P. Liao5 MD, Tianrun Cai6 MD, Ashwin N. Ananthakrishnan7 MD, Vivian S. Gainer8, Susanne E. Churchill9 MD, Peter Szolovits10 PhD, Shawn N. Murphy11,8 MD, Isaac S. Kohane9 MD, Tianxi Cai4 ScD.
1Center for Statistical Science, Tsinghua University, Beijing, China; 2Department of Industrial Engineering, Tsinghua University, Beijing, China; 3Department of Mathematical Sciences, Tsinghua University, Beijing, China; 4Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA; 5Division of Rheumatology, Brigham and Women’s Hospital, Boston, MA, USA; 6Department of Radiology, Brigham and Women’s Hospital, Boston, MA, USA; 7Division of Gastroenterology, Massachusetts General Hospital, Boston, MA, USA; 8Research IS and Computing, Partners HealthCare, Charlestown, MA, USA; 9Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA; 10Computer Science and Artificial Intelligence Laboratory (CSAIL), Massachusetts Institute of Technology, Cambridge, MA, USA; 11Department of Neurology, Massachusetts General Hospital, Boston, MA, USA.

Introduction: EHR-based phenotyping is the classification process that infers whether a patient has a specific phenotype, based on the information recorded in his/her electronic health records (EHR). Such extracted phenotypic information can be used to reproduce results of previous genomic studies based on conventional cohorts[1,2], and it is driving new studies as well, such as the phenome-wide association study (PheWAS)[3–5]. The phenotyping process requires an algorithm that aggregates “features”, such as codes or identified terms from the EHR, including diagnosis, symptoms, medications, labs etc. that are related to the phenotype, and the coefficients associated with the features are conventionally determined by model fitting using expert annotated training samples. Designing the features and annotating the samples require tremendous expert involvement and are the major bottlenecks of the development of phenotyping algorithms, which typically takes many months. Previous studies have successfully automated feature designing by leveraging online knowledge sources and unannotated EHR data[6,7]. In this talk, we present PheNorm, a technique to generate phenotyping algorithms without using expert-annotated samples.

Methods: We use a patient’s count of the target phenotype’s ICD-9 codes $x_{ICD}$, the count of the phenotype’s positive mentions in the narrative notes $x_{NLP}$, the count of notes $x_{note}$, and other natural language processing (NLP) features automatically selected by the SAFE method[7] to form a working data set. A derived feature $x_{ICD,NLP} = x_{ICD} + x_{NLP}$ is added. All features are logarithm-transformed. A multiple of the note count is subtracted from the features to make them resemble normal mixture distributions, where the multiplier is determined by minimizing the difference between the empirical distribution and the normal mixture fitted by the EM algorithm. For $x_{ICD}$, $x_{NLP}$, and $x_{ICD,NLP}$, the two modes in the normal mixture represent patients with and without the phenotype. The modified features, excluding the note count, further go through a denoising process using dropout training, where a corrupted version of the features are used to predict the uncorrupted normalized $x_{ICD}$, $x_{NLP}$, or $x_{ICD,NLP}$. The denoising process yields a formula that scores the probability of the phenotype.
Results: We evaluated the accuracy of PheNorm by AUC, using the annotated samples for the phenotypes coronary artery disease, rheumatoid arthritis, Crohn’s disease, and ulcerative colitis that were presented in Yu et al.[7] The AUC of $x_{ICD}$ was 0.84, 0.87, 0.82, and 0.81, respectively for the four phenotypes, and the AUC of $x_{NLP}$ was 0.84, 0.90, 0.91, and 0.90, respectively. Based on $x_{ICD\&NLP}$, PheNorm achieved AUC 0.90, 0.94, 0.95, and 0.94 for the 4 phenotypes, respectively. For comparison, the AUC of algorithms trained with gold-standard labels was 0.90, 0.92, 0.94, 0.94, respectively, when the training sample size $N = 100$ and 0.91, 0.93, 0.94, and 0.95, respectively, when $N = 200$. These sample sizes were chosen as references because current large-scale phenotyping efforts (e.g. 10 phenotypes at a time) rarely have the bandwidth to offer more than 200 gold-standard labels for training for each phenotype, and it shows the potential of our method to streamline phenotyping without compromising the accuracy of a supervised approach.

Conclusion: The key characteristic of high-throughput phenotyping is accurate algorithm generation without the intensive involvement of clinical experts. PheNorm has achieved this by exploiting the underlying joint distribution of the automatically curated features, thus avoiding expert annotation. By fully automating the generation of accurate phenotyping algorithms, PheNorm demonstrates the capacity for EHR-driven annotations to scale to the next level – phenotypic big data.

REFERENCE

Predicting Non-Small Cell Lung Cancer Diagnosis and Prognosis by Fully Automated Microscopic Pathology Image Features

Kun-Hsing Yu, MD, PhD1,2,3, Ce Zhang, PhD4, Gerald J. Berry, MD5, Russ B. Altman, MD, PhD2, Christopher Ré4, PhD, Daniel L. Rubin, MD, MS2, Michael Snyder, PhD3

1 Department of Biomedical Informatics, Harvard Medical School, Boston, MA; 2 Biomedical Informatics Program; 3 Department of Genetics; 4 Department of Computer Science; 5 Department of Pathology, Stanford University, Stanford, CA

Abstract

Lung cancer is the most prevalent cancer worldwide, and histopathological assessment is indispensable for its diagnosis. However, human evaluation of pathology slides cannot accurately predict patients’ prognoses. In this study, we obtained 2,186 hematoxylin and eosin stained histopathology whole-slide images of lung adenocarcinoma (AC) and squamous cell carcinoma (SCC) patients from The Cancer Genome Atlas (TCGA) and 294 additional images from the Stanford Tissue Microarray (TMA) Database. We extracted 9,879 quantitative image features and used regularized machine-learning methods to classify cancer types (AUC>0.73) and to distinguish shorter-term survivors from longer-term survivors with stage I AC (P<0.003) or SCC (P=0.023) in the TCGA test set. We validated the survival prediction framework with the TMA cohort (P<0.036 for both tumor types). Our results suggest that automatically-derived image features can predict the prognosis of lung cancer patients and thereby contribute to precision oncology. Our methods are extensible to histopathology images of other organs.

Introduction

Lung cancer causes more than 1.4 million deaths annually. Visual evaluation of the microscopic histopathology slides is indispensable to establishing the diagnosis and defines the types and subtypes of lung cancers1. However, qualitative evaluation of histopathology patterns alone is insufficient for predicting the survival outcomes of patients with lung adenocarcinoma (AC) or lung squamous cell carcinoma (SCC)2,3, and even the best-characterized histopathology features only achieve modest agreements among experienced pathologists4. In this study, we aimed to improve the diagnostic and prognostic prediction of lung AC and SCC patients through machine learning methods5.

Materials and Methods

2,186 whole-slide hematoxylin and eosin (H&E)-stained histopathology images were obtained from The Cancer Genome Atlas (TCGA), which included samples from 515 lung AC patients and 502 lung SCC patients. A segmentation and feature extraction pipeline for 9,879 features (including the size, shape, and texture of cells and their nuclei) was designed using CellProfiler6. Seven supervised machine learning approaches (naïve Bayes, support vector machines with Gaussian, linear, and polynomial kernels, bagging, random forest with conditional inference trees, and Breiman’s random forest) were used for diagnostic classification. Elastic net-Cox proportional hazards models were employed for survival prediction. A held-out test set from TCGA was utilized to evaluate the classification performance. For survival prediction, patients in the test set were classified into good or poor prognostic groups, and the log-rank test was employed to examine the difference between groups. To ensure the extensibility of the developed methods, tissue microarray (TMA) images of 227 lung AC and 67 lung SCC patients from Stanford Department of Pathology were used to validate the diagnostic and prognostic prediction methods. Detailed methods and source codes are available at http://www.nature.com/ncomms/2016/160816/ncomms12474/full/ncomms12474.html.

Results

We extracted 9,879 image features and used them to build machine-learning models for diagnosis and prognosis prediction. Our classifiers identified images of lung AC from those of adjacent dense benign tissue (e.g. inflammation, atelectasis, or lymphocytic infiltration in the absence of tumor cells) with area under the receiver operator characteristics curves (AUC) 0.85 in the test set, classified lung SCC from adjacent dense benign tissue (test set AUC = 0.88), and distinguished AC from SCC (test set AUC > 0.73 for most classifiers; Figure 1A). Our diagnostic classification method was validated in the TMA dataset (test AUC > 0.73; Figure 1B), and the performance of the top classifiers did not differ much (ANOVA test P-value ≥ 0.08). In addition, we built survival models based on the extracted image features and successfully distinguished shorter-term survivors from longer-term survivors with AC (log-rank test P-value < 0.003) and SCC (log-rank test P-value = 0.023). Our machine-learning workflow was validated in the TMA cohort (log-rank test P-value < 0.036; Figure 2).
Figure 1. Quantitative image features successfully distinguished histopathology images of lung AC from those of lung SCC. (A) ROC curves for classifying the two malignancies in the TCGA test set. (B) The results were validated in the TMA cohort.

Figure 2. Quantitative image features predicted the survival outcomes of lung cancer patients. The survival prediction methods were validated in the TMA datasets for both (A) stage I AC and (B) SCC.

Discussion
To our knowledge, this is the first study to predict lung cancer patients’ diagnoses and prognoses by fully-automated quantitative histopathology analyses. Previously, the vast amount of information contained in whole-slide pathology images has posed a great computational challenge to researchers. In this study, we designed an automated workflow that identified thousands of objective features from the images, built and evaluated machine-learning classifiers to predict the tumor types and prognosis of lung cancer patients, and validated our results in an independent cohort. The top features associated with survival outcomes included Zernike shape features of the nuclei and nuclei texture features, demonstrating that the nuanced tumor nuclei patterns are important determinants of patient prognosis.

Conclusion
In summary, we demonstrated that histopathology image classifiers based on quantitative features successfully predicted tumor types and survival outcomes of patients with non-small cell lung cancer. This capability was superior to the current practice utilized by pathologists. Our methods could facilitate prognostic prediction based on the routinely collected H&E stained slides, thereby contributing to precision oncology and enhance the quality of care.

References
Criteria2Query: Automatically Transforming Clinical Research Eligibility Criteria Text to OMOP Common Data Model (CDM)-based Cohort Queries

Chi Yuan, MS1,4, Patrick B. Ryan, PhD1,2,3, Yixuan Guo, MA1, Peng Jin, MS1, Kang Tian, MA1, Chunhua Weng, PhD1

1Department of Biomedical Informatics, Columbia University, New York, NY, USA; 2Observational Health Data Sciences and Informatics, New York, NY, USA; 3Janssen Research & Development, LLC, Titusville, NJ, USA; 4School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, Jiangsu, P.R. China

Introduction

Patient recruitment has been a persistent barrier to clinical and translational research1. Over 50% of studies fail due to difficulties in recruitment. The universal adoption of electronic health records makes it possible to query distributed, large clinical databases for prescreening potentially eligible patients for performing feasibility assessment or cohort identification for clinical studies. However, interpretations of clinical research eligibility criteria may vary from site to site or from person to person, which can lead to incompatible or mismatched patient queries and compromise the integrity of multi-site studies. In addition, encoding of eligibility criteria into database queries often involves laborious human effort, which is costly, not scalable, and often prohibitive for clinicians and researchers, who do not necessarily understand informatics and know how to map criteria concepts to their data representations. This study aims to build a natural language processing system to transform eligibility criteria text into standards-based cohort identification queries that are sharable and configurable by end users, who can then focus their energy on feasibility assessment and iterative refinement of the criteria based on dynamic feedback with patient counts during eligibility criteria design.

Methods

The system Criteria2Query consists of five modules (Figure 1): i.e., text parsing, criteria filtering, terminology standards-based concept mapping, automatic query formulation, and query execution with dynamic feedback generation for users. Criteria parsing is supported by an open-source machine learning parser, EliIE2, which outputs XML representations for recognized entities and their temporal relationships using the Observational Medical Outcomes Partnership (OMOP) common data model (CDM) V5. Criteria (e.g., “willing to sign the consent” and “able to walk 5 miles on treadmill”) that cannot be queried within EHR for prescreening purposes are filtered out based on heuristics or empirical knowledge. On this basis, an initial event, which is needed to establish the target population, is selected automatically while allowing for user configuration. Concepts and their relationships in retained criteria are then mapped to the Observational Health Data Sciences and Informatics (OHDSI) controlled clinical vocabularies to obtain concept IDs for each concept and to create structured cohort definitions, which are further translated into cohort queries in JSON or SQL formats using a public OHDSI Web API (https://github.com/OHDSI/WebAPI). These queries can be executed against any OMOP CDM-based patient database to prescreen potentially eligible patients. We evaluated the usefulness and cost-effectiveness of Criteria2Query by documenting the time saved by using Criteria2Query when translating clinical trial eligibility criteria into queries for a professional clinical database query technician.

Results

Example queries and their results are shown in Figure 2. Criteria2Query enables the generation of executable cohort queries that can be shared across different CDM-based databases to return counts of distinct patients satisfying the criteria. The system enables a closed feedback loop from criteria text to patient counts and allows iterative criteria refinement based on data feedback. This feedback can inform clinical study designers to author more realistic criteria for their research. Criteria2Query can be used as a standalone system to generate SQL queries from text or used as a middleware for different clinical decision support systems based on OMOP CDM. According to our preliminary evaluation with one technician, who performs criteria-to-query translation in our institution, in general an experienced user needs > 40 minutes to create CDM cohort queries for clinical studies. This time includes defining concept sets and identifying the condition event, the observational event and several other events. The use of Criteria2Query noticeably reduced cohort generation time to be an average of 15 minutes per study, with nearly 60% reduction of time.

To successfully perform clinical research of sufficient power, clinician researchers need to recruit an adequate number of patients meeting the eligibility criteria of their research. Clinicians may not have the information about how many
patients meet their requirements in clinical data warehouse, and may not know the number of the candidates when they are designing the criteria. Criteria2Query empowers clinicians and researchers to interrogate the patient databases directly and provides real-time feedback about the number of potentially eligible patients so that they can further fine tune the criteria in a user-friendly interface iteratively. Criteria2Query removes the need for clinicians and researchers to understand SQL and allow them to better focus on iterative feasibility assessment and criteria optimization.

Conclusions
Criteria2Query demonstrates early promise for improving the efficiency for cohort identification for EHR-based comparative effectiveness studies and empower researchers or clinicians to identify patient cohorts without mastering knowledge of patient database or clinical terminology. Future studies are warranted to test the correctness, usability and cost-effectiveness of this system at a larger scale and in a more systematic manner.

Acknowledgements
This study is sponsored by grant R01 LM009886 (Bridging the semantic gap between research eligibility criteria and clinical data; PI: Weng). We thank Dr. Maurine Tong for her editing of the paper.

References

Figure 1. The behind-the-scene parsing pipeline that enables iterative eligibility criteria design.

Figure 2. The front-end user interface of the Criteria2Query system.
Revision of Order Sets Using Experts’ Knowledge and Data-Driven Evidence

Yiye Zhang PhD¹, Richard Trepp MD²,³, Jorge Luna MS, MPH, MPhil², David K. Vawdrey PhD²,³, Victoria L. Tiase MSN, RN-BC²
¹Weill Cornell Medical College, New York, NY; ²NewYork-Presbyterian Hospital, New York, NY; ³Columbia University Medical Center, New York, NY

Introduction
Order sets are a core component of Computerized Provider Order Entry (CPOE) systems and serve to facilitate more efficient and effective order entry¹. Order sets allow providers to select multiple relevant orders with a few clicks by presenting appropriately grouped medical orders, thus increasing the efficiency of ordering. They also serve to reduce variation in the ordering process and encourage compliance with best practices². Order sets are designed based on the intended functions and ordering times of their content items¹. For example, an ‘AM Lab Order Set’ is used at our institution to order laboratory tests to be collected the next morning. Items in the order set can be further customized with default settings according to clinical relevance, recommended practices, and frequency of usage.

One significant concern raised with CPOE is the burden of excessive mouse clicks in placing orders, which potentially lead to undue workload for providers and even medical errors². As an alternative to using order sets, providers can place orders a la carte; however, a la carte items often require even more clicks than order sets, and thus contribute to increased workload. An emerging approach to this challenge is to develop and optimize order sets based on data-driven evidence²,³. With data-driven order sets, though, it may be difficult to incorporate experts’ knowledge and policy-driven standards that are not identifiable from data. Taking our institution’s ‘AM Lab Order Set’ as a case-study, we investigated the workload associated with the usage of this order set. We simulated 5 scenarios to examine approaches for order set revision that incorporates data-driven evidence, experts’ knowledge, and policy-driven standards, and compared the number of mouse clicks required to use the order set after each simulated revision.

Methods
We obtained patient characteristics and order-related information from September 2014 to September 2015 from an electronic health record (EHR) system (Allscripts Healthcare Solutions, Inc.) at our institution. The ‘AM Lab Order Set’ contained the 12 order items shown in Table 1. It was used nearly 200,000 times during the one-year study period. Data included 1,065,401 order placements for 27,419 patients who had at least one order placed via the ‘AM Lab Order Set’. This order set was used 6.7 times (SD=10.55), on average, over an average length of stay of 4.8 (SD =10.31) days during the study period. Variables include de-identified ID, order name, order set name (or a la carte), time stamp of order placement, Diagnosis-Related Group (DRG), and patient demographic information. All items in the ‘AM Lab Order Set’ were unselected by default (i.e., default-OFF) in the CPOE user interface.

Table 1. Order items contained in ‘AM Lab Order Set’ and their usage statistics.

<table>
<thead>
<tr>
<th>Category</th>
<th>Order</th>
<th>Number of times used</th>
<th>Number of unique patient</th>
<th>% Unique patient</th>
</tr>
</thead>
<tbody>
<tr>
<td>Labs</td>
<td>Complete Blood Count (CBC)</td>
<td>28,286</td>
<td>10,867</td>
<td>39.6%</td>
</tr>
<tr>
<td></td>
<td>CBC with Differential</td>
<td>129,913</td>
<td>24,165</td>
<td>88.1%</td>
</tr>
<tr>
<td></td>
<td>Liver Function Panel</td>
<td>52,171</td>
<td>13,506</td>
<td>49.3%</td>
</tr>
<tr>
<td></td>
<td>Basic Metabolic Panel</td>
<td>166,781</td>
<td>27,143</td>
<td>99.0%</td>
</tr>
<tr>
<td></td>
<td>Phosphorus</td>
<td>73,127</td>
<td>17,537</td>
<td>64.0%</td>
</tr>
<tr>
<td></td>
<td>Troponin I</td>
<td>3,286</td>
<td>2,477</td>
<td>9.0%</td>
</tr>
<tr>
<td></td>
<td>Type and Screen</td>
<td>28,255</td>
<td>10,930</td>
<td>39.9%</td>
</tr>
<tr>
<td>Labs – Coag Panel</td>
<td>PT/INR</td>
<td>72,577</td>
<td>16,903</td>
<td>61.6%</td>
</tr>
<tr>
<td></td>
<td>Activated Partial Thromboplastin Time (APTT)</td>
<td>66,182</td>
<td>16,173</td>
<td>59.0%</td>
</tr>
<tr>
<td>Labs – Cardiology Adult</td>
<td>Magnesium</td>
<td>158,963</td>
<td>26,707</td>
<td>97.4%</td>
</tr>
<tr>
<td></td>
<td>ECG 12 Lead.</td>
<td>15,292</td>
<td>4,710</td>
<td>17.2%</td>
</tr>
<tr>
<td>Radiology</td>
<td>Port Chest 1 View</td>
<td>5,782</td>
<td>2,248</td>
<td>8.2%</td>
</tr>
</tbody>
</table>

232
Click cost (CC) is defined as the number of mouse clicks associated with performing a task\(^2\). For this analysis, to compute CC from data, we estimated that it required three clicks to open an order set, and one click to select a default-OFF item or to de-select a default-ON item. We also estimated that it required three clicks to place an *a la carte* order. Estimates were based on the specific EHR interface at the study site. Table 2 describes seven simulated scenarios for revision of the ‘AM Lab Order Set’. In M1, orders that were defaulted to ON are ‘basic metabolic panel’, ‘CBC with differential’, and ‘magnesium.’ The six most common orders in M3 include ‘b/p, hr, rr and temp’, ‘general nursing’, ‘glucose whole blood poc’, ‘magnesium sulfate inj’, ‘calcium, serum’, and ‘heparin inj’. Furthermore, we applied simulations (M4 to M7) that combine scenarios M1 to M3.

<table>
<thead>
<tr>
<th>Simulation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>M0: Status Quo</td>
<td>Compute CC under the status quo content of ‘AM Lab Order Set’, including all <em>a la carte</em> orders placed 10 minutes within the ‘AM Lab Order Set’</td>
</tr>
<tr>
<td>M1: Default</td>
<td>Default orders from ‘AM Lab Order Set’ that have been used by at least 80% of the patients to ON</td>
</tr>
<tr>
<td>M2: Expert Knowledge</td>
<td>Revised ‘AM Lab Order Set’ based on team of experts’ knowledge</td>
</tr>
<tr>
<td>M3: Data-driven Evidence</td>
<td>Add 6 most common out of 1,148 laboratory orders placed within 10 minute proximity of ‘AM Lab Order Set’</td>
</tr>
<tr>
<td>M4</td>
<td>M2 + M3</td>
</tr>
<tr>
<td>M5</td>
<td>M1 + M2 + M3</td>
</tr>
<tr>
<td>M6</td>
<td>M4 + policy-driven standards</td>
</tr>
</tbody>
</table>

**Results**

Figure 1 shows the average CC per patient across M0 to M6. The lowest CC was obtained in M4 when we revised using experts’ knowledge with the data-driven evidence of most commonly co-occurred orders with ‘AM Lab Order Set.’ The reduction was 65,806 clicks for all patients in a year for this one order set. M6 incorporates M4 with policy-driven standards, which noted that while ‘general nursing’ is a common order, it should be replaced by a specific nursing order. M6 had the second lowest CC, which shows promise in its future implementation, and we also expect to see a larger effect of M6 in reducing CC in future data. Adding defaults in M1 resulted in the third lowest CC, performing better than revising orders based on experts’ knowledge or data-driven evidence alone. The difference in the average CC between M0 and M4 is significant with a p-value <0.00 and a confidence interval of [0.78, 4.22] using T-test.

![Figure 1](image.png)

**Discussion**

This study demonstrated that when revising order sets, incorporating both experts’ knowledge and data-driven evidence may result in the most efficient ordering process for providers. Future work should examine order sets of varying complexities and designed for disparate purposes to generalize findings, as well as relations between over-utilization of services and order set settings such as defaults. While data-driven evidence in this study was based simply on frequency of orders, future work will also employ more advanced data analytics methods such as clustering algorithms to identify optimal order set assignments in reducing the click costs.

**References**

Subtyping Parkinson’s Disease with Recurrent Neural Network Models

Xi Zhang, PhD¹, Jian Liang, MS², Cao Xiao, PhD³,
Yize Zhao, PhD¹, Fei Wang, PhD¹

¹Department of Healthcare Policy and Research. Weill Cornell Medical College. New York, USA
²Department of Automation. Tsinghua University. Beijing, China
³Center for Computational Health IBM T. J. Watson Research Center. New York, USA

Abstract

Objectives: To discover the cohesive progressive subtypes of Parkinson’s Disease (PD) patients using comprehensive data provided by the Parkinson's Progression Markers Initiative (PPMI). Methods: To explore the sequentiality of the patient records, we propose to leverage the Long-Short Term Memory (LSTM) model, which is a variant of Recurrent Neural Network that received extensive attentions in natural language processing in recent years. In particular, we first concatenate the records for each patient into a sequence according to their associated timestamps. Then we train an LSTM model on top of those sequences and embed them into a latent homogeneous sequence space. Patient similarity will be evaluated on those latent sequence embeddings and finally the progressive subtypes were identified through clustering with the learned patient similarities. Results: We performed evaluation using 6-years PPMI data including 15,798 patient records with our proposed methodology. Eventually three PD subtypes were identified. Statistical testings were conducted to identify the discriminative factors for each subtype for characterization purpose. The first subtype was characterized by moderate function decay on motor ability and stable cognitive ability. Patients in the second subtype had mild function decay on both motor and non-motor ability. The third subtype was characterized by rapid progression on both motor and non-motor ability.

Methods

As Parkinson's Disease has been demonstrated large clinical heterogeneity, identifying subtypes may facilitate the further research on underlying etiologies and designing appropriate therapies. Because of the various challenges on working with heterogeneous patient data (e.g., sparsity, heterogeneity, noisiness, etc.), learning proper representation for each patient is important. This paper presents a novel framework based on Long-Short Term Memory (LSTM) recurrent neural network to represent patients by incorporating patient records from multiple sources. We divided the feature in patient records into two parts: inputs or targets. Targets are a set of determinant variables that are closely related to the progression of PD shown in previous clinical studies. Other variables were regarded as input features. The dimension of input features and targets were 319 and 82, respectively.

Before training the LSTM with inputs and targets, we first performed some pre-processing on the raw patient records. Most of the patient features are temporal with different value types. Some of them are continuous while the others are categorical. We transformed the categorical features into one-hot representation. For instance, if one variable contains three values: red, blue, and green, we encoded the red as 001, blue as 010, and green as 100 respectively. There could also be lots of missings in patient records, we just use the simple Last Occurrence Carry Forward (LOCF) to impute them. If the missing value is at the beginning, we will carry next value backwards. Otherwise if the all values for the entire feature are missing, we just impute them with average feature values over all patients. Once the sequential input and target vectors are prepared, the neural network can be trained on top of them. In LSTM, the representation vector $h_t$ (t is the index for timestamp) can be computed by following equations:

$$i_t = \sigma(W_i[x_t, h_{t-1}] + b_i); \quad f_t = \sigma(W_f[x_t, h_{t-1}] + b_f); \quad o_t = \sigma(W_o[x_t, h_{t-1}] + b_o)$$

$$c_t = \tanh(W_c[x_t, h_{t-1}] + b_c); \quad h_t = o_t \cdot \tanh(c_t)$$

where $\sigma(x)$ is the logistic sigmoid function, $i$, $f$, $o$ and $c$ are respectively the input gate, forget gate, output gate, cell state. When training LSTM, we use square loss for continuous valued targets and logistic loss for binary valued targets. After the sequential embeddings $\{h_t\}$ are learned, we use Dynamic Time Warping (DTW) to measure the similarities between pairwise patient embeddings and t-SNE to visualize the patients as data objects in two-dimensional space, where the pairwise patient similarities are preserved. Finally, Kmeans was deployed on the t-SNE results to get the patient clusters as subtypes.

The data set we used in our study is from the Parkinson’s Progression Markers Initiative (PPMI) study organized by Michael J. Fox Foundation. PPMI is an ongoing observational, international, multi-source study that has various Parkinson’s disease progression markers including demographics, clinical, imaging and biospecimen. The data contained archives of enrolled subjects from June 1, 2010 to June 1, 2016, and consisted of 15,798 heterogeneous records of 683
subjects including 466 PD patients and 217 healthy controls. On average, each patient had approximately 23 records. To build an LSTM model, the data were randomly split to 60% training, 20% validation, and 20% testing.

Results

One difficulty for evaluation of disease subtyping results is that there is no ground truth. Therefore, visualization provides an effective way of perceiving the results. Figure 1 shows the t-SNE visualization of the patients in two-dimensional space based on the DTW similarities, from which we can clearly observe three relatively well-separated and compact patient groups. To characterize each subtype, we also conducted rigorous statistical testing for each of the features in patient records. The testing is done with respect to both the value mean at the baseline and follow-up for each feature. Discriminative variables with \( p \)-value \(<0.05\) were identified and shown in Table 1. At baseline, the first subtype of 201 patients with average age 58.79 years old was characterized by the mild motor stage of H&Y and MDS-UPDRS Part I-II, the slightest non-motor symptoms of cognitive impairment, depression, anxiety among three subtypes. The motor severity of the second subtype of 107 patients with average age 61.93 was similar with the first subtype. However, several non-motor manifestations such as cognitive, depression, and anxiety of subtype II were worse than subtype I. As for subtype III of 158 patients with average age 65.32, almost all of the motor and non-motor symptoms were severer than the two former subtypes. After a 6-year follow-up, we conducted the statistical tests on the difference of mean values between the start timestamp and end timestamp, and discovered discriminant variables of disease progression. By comparing the baseline and follow-up feature values, we can observe the following: the first subtype was with moderate function decay on motor ability and stable cognitive ability; the second subtype had mild function decay on both motor and non-motor ability; and the third subtype suffered rapid disease progression on both motor and non-motor ability.

Table 1 Group characteristics of patients in the three subtypes

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Subtype I (N = 201)</th>
<th>Subtype II (N = 107)</th>
<th>Subtype III (N=158)</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Baseline</td>
<td>Follow-up</td>
<td>Baseline</td>
<td>Follow-up</td>
</tr>
<tr>
<td>Age onset</td>
<td>58.79(9.53)</td>
<td>61.93(9.05)</td>
<td>65.32(8.86)</td>
<td>(&lt;0.0001^a)</td>
</tr>
<tr>
<td>Hoehn and Yahr Stage</td>
<td>1.44(0.5)</td>
<td>1.81(0.48)</td>
<td>1.52(0.52)</td>
<td>1.66(0.51)</td>
</tr>
<tr>
<td>MDS-UPDRS Part I</td>
<td>4.28(2.99)</td>
<td>6.92(4.58)</td>
<td>6.38(4.88)</td>
<td>7.26(5.26)</td>
</tr>
<tr>
<td>MDS-UPDRS Part II</td>
<td>4.52(3.23)</td>
<td>7.48(4.85)</td>
<td>5.58(4.41)</td>
<td>6.85(4.41)</td>
</tr>
<tr>
<td>MDS-UPDRS Part III</td>
<td>18.34(7.9)</td>
<td>22.39(11.86)</td>
<td>19.99(9.04)</td>
<td>23.18(9.92)</td>
</tr>
<tr>
<td>Montreal Cognitive Assessment</td>
<td>27.75(2.01)</td>
<td>27.98(1.86)</td>
<td>27.26(2.42)</td>
<td>27.09(2.4)</td>
</tr>
<tr>
<td>Geriatric Depression Scale</td>
<td>5.11(1.43)</td>
<td>5.2(1.31)</td>
<td>5.2(1.17)</td>
<td>5.31(1.28)</td>
</tr>
<tr>
<td>State Trait Anxiety Inventory</td>
<td>61.84(15.85)</td>
<td>59.52(16.07)</td>
<td>62.14(17.96)</td>
<td>61.89(18.15)</td>
</tr>
<tr>
<td>DaTScan</td>
<td>1.43(0.55)</td>
<td>1.23(0.55)</td>
<td>1.60 (0.64)</td>
<td>3.05(0.62)</td>
</tr>
</tbody>
</table>

*Chi-square test; *Kruskal-Wallis H-test. *Abbreviations: MDS-UPDRS: Movement Disorders Society–revised Unified Parkinson's Disease Rating Scale;

References

Assessing Electronic Health Record Readability

Jiaping Zheng¹, Hong Yu, PhD²

¹College of Information and Computer Sciences, University of Massachusetts Amherst, MA; ²Department of Quantitative Health Sciences, University of Massachusetts Medical School, Worcester, MA

Introduction

Patient engagement¹ has emerged as an important component of strategies to improve health care. A growing body of evidence has accumulated on better health outcomes and care experiences associated with higher engagement². Electronic Health Record (EHR) systems with patient engagement capabilities have also enjoyed tremendous growth in usage recently³. However, access to EHR by itself is not sufficient to motivate patients to be involved because of the complex and technical nature of the EHR. Furthermore, many patients have limited health literacy and are not proficient in comprehending and acting on health information⁴. Integrating tools that can match clinical narratives within patients’ literacy capabilities into EHR may make them more accessible.

In this study, we explore methods to automatically assess readability levels of clinical narratives in EHR. Numerous readability metrics have been developed to assess the a document’s grade level. For example, Flesch-Kincaid Grade Level (FKGL)⁵ predicts grade level using average sentence and word length. This metric and many others rely on the assumption that longer words and sentences are more difficult. However it may not hold for EHR narratives as sentences are usually short and abbreviations are common. Our previous work⁶ has demonstrated that these formulas did not show adequate correlation with readers’ perceived difficulty, and were not appropriate to assess the readability of EHR notes. In this work, we considered measuring readability as a ranking task, where the relative difficulty of two documents are compared. Our method can be applied to both EHR narratives and a Personal Health Record system.

Method

Amazon Mechanical Turk (AMT) users were recruited to rate 20 random document pairs based on their perceived difficulty. They are requested to rate the readability of the documents on a scale from 1 (easy) to 10 (difficult). The document pair was of similar length (±50 tokens) and comparable difficulty according to FKGL (±0.5). The sources of the documents included Wikipedia (denoted as “wiki”) and de-identified EHR notes (denoted as “med”). Three common diseases were selected: cancer, diabetes, and hypertension. Wikipedia documents were randomly selected from the respective disease category, EHR notes using ICD-9 codes. Document statistics are shown in Table 1.

<table>
<thead>
<tr>
<th></th>
<th>wiki</th>
<th>med</th>
</tr>
</thead>
<tbody>
<tr>
<td>cancer</td>
<td>166 documents</td>
<td>113 documents</td>
</tr>
<tr>
<td>diabetes</td>
<td>58 documents</td>
<td>133 documents</td>
</tr>
<tr>
<td>hypertension</td>
<td>84 documents</td>
<td>208 documents</td>
</tr>
</tbody>
</table>

We learned a support vector machine model from the pairwise comparisons from AMT user’s ratings. In this setting, one example is generated when two documents were assigned different scores by a user. Score differences by different users were not used to generate examples since the difference may be due to the users’ background knowledge. A model was then learned using SVM²³⁶. We employed several types of features, including readability formula based, word frequency, and neural language models. We included word and sentence length, and proportion of polysyllabic words from readability formulas. We also included word frequency obtained from Wikipedia and EHR notes as features, since common words are likely to be easier to understand. Additional features included document length in words and sentences. Long documents requires more cognitive processing to comprehend, which might translate to higher perceived difficulty. Lastly, we captured language patterns using word embeddings learned from a combination of Wikipedia documents and de-identified EHR notes.
Results

We split the annotated data three ways (70%/10%/20%) into training, development, and test sets. The three disease topics were stratified. Hyperparameters were optimized on the development set. We evaluated our system using Kendall’s coefficient of concordance $W$, a statistic that measures the agreement between rankings. This metric ranges from 0 to 1. The best agreement between our system and the AMT annotators on the test set was 0.65. As a comparison, we show the distributions of $W$ between two AMT users in Figure 1. Our system’s concordance to a human annotator is comparable to the average between two AMT users’. Existing machine learning based systems were usually designed around classification. They are often limited to a few pre-defined labels or require corpus labeled at distinct levels.

![Figure 1: Histogram of Kendall's W between AMT users.](image1.png)

To highlight the range of difficulty perceptions by AMT users, Figure 2 shows the maximum difference in ratings assigned by AMT users on documents that were rated by at least two users. The mean difference is 3.6, suggesting that perceptions of difficulty vary considerably among users.

![Figure 2: Max difference in AMT ratings on a document.](image2.png)

Conclusion

Patient’s access to EHR notes has increased dramatically according to national statistics. However, actively engaging patients remains challenging. Assessing EHR notes’ readability may make them more accessible. We developed a new machine learning based method toward this end. We collected document difficulty ratings from AMT users on EHR notes and Wikipedia articles. Our method learned a ranking model from comparisons of the difficulty judgments. Our experiments showed that its concordance with a human user was similar to the concordance between different human annotators. This method can potentially be personalized to better accommodate for users’ background knowledge.

Acknowledgments

This work was supported in part by the Award 1I01HX001457 from the United States Department of Veterans Affairs Health Services Research and Development Program Investigator Initiated Research. The contents do not represent the views of the U.S. Department of Veterans Affairs or the United States Government.

References

Acute Kidney Injury in the Intensive Care Unit: Timing and Detection

Leming Zhou, PhD1, Valerie Watzlaf, PhD1, Christopher Keener, BS2, Dilhari DeAlmeida, PhD1, Rebecca DeSensi, BS2, Emily Foldes, MS2, Gilles Clermont, MD1, John A. Kellum, MD2

1Department of Health Information Management, University of Pittsburgh, Pittsburgh, PA; 2Department of Critical Care Medicine, University of Pittsburgh, Pittsburgh, PA

Introduction

Acute kidney injury (AKI) is common during critical illness occurring in approximately 30-60% of patients and is associated with a case fatality rate of approximately 60%-70%1-3. With no known unifying mechanism and unsatisfactory treatment, AKI is a devastating illness4. Serum creatinine (sCr) is used as one of the biomarkers (along with urine output) in the standardized AKI definition recommended by the Kidney Disease Improving Global Outcomes (KDIGO) guideline5.

However, sCr is a delayed signal of AKI and is not monitored continuously. AKI may occur in between measurements of sCr. In other words, AKI staging based on sCr lags behind actual renal damage. This delay may directly impact patient outcomes. Therefore, in this work, we sought to determine the time delay between the theoretical start of stage 2 AKI according to the sCr trajectory and the actual detection of stage 2 AKI from measured sCr values, and the impact of this delay to the patient outcomes, such as cost, length of stay, and mortality. The results may lead to a more proper way of managing AKI cases at the early stage. Here the stage of AKI is determined by the widely validated KDIGO criteria5.

Methods

This retrospective, observational cohort study used information from the HiDenIC database. Clinical and administrative data, including demographic, procedural, diagnostic, laboratory, pharmacologic and billing information on all patients admitted to one of eight intensive care units (ICUs) at the University of Pittsburgh Medical Center (UPMC) from 2000 to 2008 are contained in HiDenIC. The source population includes more than 54,000 patients. We excluded patients with severe existing kidney disease (sCr > 3.5 mg/dl), including those on chronic hemodialysis or kidney transplant recipients. From 41,821 patients 18 years of age or older, we identified 1,014 patients who developed stage 2 or 3 AKI based on the KDIGO criteria after ICU admission. These patients also had a baseline sCr value and a sufficient number of in-hospital creatinine measurements available to reliably fit a regression line to determine the AKI event time and slope of the creatinine kinetics. Once the creatinine kinetics were determined and the AKI event time was identified for each patient, we further excluded patients whose projected AKI event occurred prior to hospital admission, whose creatinine values fluctuated heavily during hospitalization, or with gaps > 96 hrs between sCr measurements. The final study population included 362 patients and almost all of them (98.9%) were admitted to UPMC Presbyterian.

All sCr trajectories were visually examined. It was noticed that only the last three sCrs (before or at the detection of stage 2 AKI) reflected the trajectory of AKI and all other sCr measures before the start of AKI typically fluctuated around the baseline. Linear regression precisely shows the trajectory of these three sCr values (as demonstrated in Figure 1). The slope from the linear regression is a measure of the rate of renal functional loss and probably is proportional to the degree of injury. The time windows from the theoretical onset of AKI to the detection of stage 2 AKI were determined according to this linear regression. The times of onset of AKI and theoretical stage 2 AKI were calculated. The time difference between this theoretical stage 2 AKI and the actual detection of stage 2 AKI is the diagnosis delay (Dx Delay in Figure 1). By definition, before the start of AKI, time gaps between sCr measurements do not have direct impact on the timeliness of AKI detection. Only the last two gaps (sCr measure gap 1 and sCr measure gap 2, especially sCr measure gap 2 shown in Figure 1) are closely related to the diagnosis delay. The average of these two sCr measure gaps is the mean gap. The correlation coefficients between diagnosis delay and slope, and between slope and mean gap were calculated.
Total hospital cost associated with the hospital visit was determined by first converting all charges recorded during
the visit to cost using the appropriate cost-to-charge ratio as defined for the specific cost center and year of the charge.
All resulting costs related to the visit were summed and a total cost calculated. The costs were inflated to 2013 US
Dollars using the Consumer Price Index for Medical Care Services from the US Bureau of Labor Statistics. Hospital
and ICU Length of Stay (LOS) were determined by calculating the number of days between admission and discharge.

We used logistic regression to examine the association between mortality and the slope of the sCr kinetics and
diagnosis delay. Mortality outcomes included hospital mortality, as well as mortality at 30-days and 1-year following
ICU admission. Models were adjusted for patient age, severity of illness factors, comorbidities and surgical
admissions. In addition, we assessed resource use and factors that contribute to hospital costs, including the hospital
and ICU LOS and renal replacement therapy (RRT) use. Models for LOS were created for patients who were
discharged from the hospital alive in order to eliminate the competing risk of death prior to discharge. Zero-Truncated
Negative Binomial Regression models adjusted for baseline illness and admission type (surgical or medical) were
created to predict the relationship between hospital and ICU LOS and the slope of the creatinine kinetics and diagnosis
delay. Adjusted logistic regression models were also created to determine the association between slope and detection
delay and in-hospital RRT use. We created multiple regression models of the log-transformed costs to determine the
association between the slope and the diagnosis delay and total hospital costs.

**Results**

The minimum value of the mean gap was 3.35 hours and the maximum value of this variable was 71.5 hours (mean =
20.5 and SD = 8.85). The median time window between the start of AKI to stage 2 AKI was 41 hours (min = 6.43,
max = 184.73, mean = 54.05, SD = 29.91). The median diagnostic delay was 2.25 hours (min = 0, max = 34, mean =
6.74, SD = 6.45). The correlation coefficient between slope and diagnosis delay was -0.070 (P = 0.18). In other words,
the rate of AKI development was not correlated with the diagnosis delay. The correlation coefficient between slope
and mean gap is -0.587 (P < 0.001), which means the faster the AKI developed (greater slope), the more frequent the
sCr was measured. The logistic regression models did not show a significant association between hospital mortality
(30-days and 1-year mortality after ICU admission) and slope or diagnosis delay (P = 0.49, OR = 0.98 for 30-day
mortality and P = 0.27, OR = 0.98 for 1-year mortality, respectively). Similar results were seen for the relationship
between hospital LOS, ICU LOS, RRT use, and slope or diagnosis delay. There was no significant association between
hospital costs and the slope or diagnosis delay. However, we did notice that steeper slopes (more rapidly worsening
renal function) were generally associated with increased mortality, hospital cost, and RRT use.

**Discussion**

Using existing knowledge concerning creatinine kinetics, changes in serum creatinine over time, and the time-stamped
reporting of creatinine, we determined the relationship between diagnosis delay (slope or rate of sCr change during
AKI development) and patient outcomes. The results indicate that clinicians already adjust the frequency of sCr
measurement according to the speed of AKI development, thus limiting diagnosis delay. This delay therefore was not
associated with a significant difference in patient outcomes. If simple computational modeling could be introduced to
forecast the development of stage 2 AKI, physicians might introduce interventions (such as fluid administration,
discontinuation of ongoing nephrotoxic drugs, or administration of drugs that reduce the nephrotoxicity) before the
occurrence of stage 2 AKI, which may in turn produce lower mortality rate and costs. However, our results indicate
that more frequent monitoring with sCr would not produce a meaningful improvement in the timeliness of AKI
diagnosis. Other methods to predict AKI before it occurs will be needed to improve outcomes.

**References**

1. Bagshaw SM, George C, Dinu I, Bellomo R. A multi-centre evaluation of the RIFLE criteria for early acute kidney
2. Ostermann M, Chang RW. Acute kidney injury in the intensive care unit according to RIFLE. Crit Care Med.
5. KDIGO. KDIGO Clinical Practice Guideline for Acute Kidney Injury, Section 2: AKI Definition. Kidney Int
Bridging the Gap Between Direct Patient Outreach in the Community and Clinical Information Access: Next Step in Effective Nationwide Clinical Trials

Lindsay P. Zimmerman, MPH1, Shazia Sathar, MS2, Charon Gladfelter, MPH1, Faraz S. Ahmad, MD1, Alejandra Onate, BA1, Lindsey L. Cook, MPH1, Shelly Sital, MPH3, Jasmin Phua2, Paris Davis, PhD, MBA4, Helen Margellos-Anast, MPH5, David Meltzer, MD, PhD6, Tamar Polonsky, MD, MS6, Raj Shah, MD7, Satyender Goel, PhD1, Abel N. Kho, MD, MS1

1Northwestern University, Feinberg School of Medicine, Chicago, Illinois; 2Medical Research Analytics and Informatics Alliance, Chicago, IL; 3The Chicago Community Trust, Chicago, IL; 4PASTORS4PCOR, Chicago, IL; 5Sinai Urban Health Institute, Chicago, IL; 6The University of Chicago, Chicago, IL; 7Rush University Medical Center, Chicago, IL

Introduction

Traditionally, clinical trials are conducted within the confines of academic medical centers, hospital systems, or institutions where clinical care is delivered. This is a longstanding process due to the effectiveness of research within a clinical setting related to regular contact with patients, access to clinical data, and research teams to execute data collection and analysis. Inadvertently, the current approach neglects patients who fall outside of the clinical system and may otherwise be the most eligible for clinical trials. We present a novel approach to recruit eligible patients directly from the community, while also leveraging affiliated health systems to extract clinical data for community participants. This approach will be tested in a PCORI funded, nationwide trial, ADAPTABLE.

ADAPTABLE (Aspirin Dosing: A Patient-centric Trial Assessing Benefits and Long-Term Effectiveness) trial is the first demonstration project conducted through PCORNet1. PCORNet: The National Patient-Centered Network is purposed to build a “network of networks” of various healthcare institutions and stakeholder groups and build partnerships to collect and use data for improved comparative-effectiveness research2,3. ADAPTABLE not only seeks to answer a clinical question, but to assess whether efficient and low-cost clinical trials can be conducted during “real-world” medical care1. ADAPTABLE presents a new model for clinical trials, aiming to minimize the burden of research activities on patients, clinicians, and institutions, by using existing EHR data and patient-reported outcomes to answer clinical questions1. This three-year pragmatic clinical trial is designed to compare the effectiveness of two aspirin doses, low (81 mg/day) vs. high (325 mg/day), in preventing myocardial infarction and stroke among 20,000 individuals with coronary heart disease1.

Methods

Northwestern University is the lead CAPriCORN (Chicago Area Patient Centered Outcomes Research Network) site for the ADAPTABLE trial, among four additional sites, University of Chicago, Rush University Medical Center, NorthShore University HealthSystem, and Cook County Health & Hospitals System. Each site maintains their own pool of eligible patients for ADAPTABLE. The community-based recruitment strategy is unique to the ADAPTABLE national network, proposed and tested by the CAPriCORN system. No other participating Clinical Data Research Networks (CDRNs) are engaging the community in ADAPTABLE recruitment at this time. The community-based recruitment strategy incorporates two community institutions into this existing workflow, the Sinai Urban Health Institute (SUHI, https://www.sinai.org/content/sinai-urban-health-institute-0) and PASTORS4PCOR (P4P, http://www.southlandmhn.com/pastors4pcor/). The community-based recruitment workflow is depicted in Figure 1.

Figure 1. Overall community-based recruitment workflow and participating organizations.
To be eligible for the ADAPTABLE trial, potential participants identified at the community site must be linked to a CAPriCORN institution for future follow-up queries of the participant’s clinical data by the ADAPTABLE Team. The linkage not only allows for assessment of whether the participant receives care at a CAPriCORN site, but also validates eligibility for patient-reported answers to the initial screening questionnaire by comparing these data to clinical data at CAPriCORN sites. Potential study participants are directly enrolled at community events, and the eligibility is assessed and demographics are collected using REDCap (Research Electronic Data Capture, https://projectredcap.org/about/). Medical Research Analytics and Informatics Alliance (MRAIA), serves as CAPriCORN’s data hub and honest broker, and assists in the de-identified patient linkage. For matching patients, the “Golden Ticket” is assigned and credited to community partners. Various stages in the community-based recruitment are described in Figure 2.

**Results/Discussion**

The CAPriCORN approach to the ADAPTABLE trial is leading the way in demonstrating the potential for new, low-touch, clinical trial methods. Recruiting participants directly from the community is a primary step in building more efficient and thorough patient recruitment methods for clinical effectiveness research. Through our initiative, we developed a community-based recruitment strategy to connect two community institutions, SUHI and P4P, to clinical sites within the CAPriCORN ADAPTABLE trial system. We created a workflow incorporating tools for data collection, linkage, and tracking. The CAPriCORN recruitment goal for ADAPTABLE is to consent and randomize 2,800 participants over the 1.5-year enrollment period, ending January 2018. Of the total enrollment goal, 30% are anticipated to come from the community. The methods described above will be used to reach this recruitment target. By the end of year one, we anticipate to have a demonstrable patient population recruited from the community. We will present learnings from implementation and actual recruitment results at AMIA 2017 Annual Symposium in November 2017.

**Acknowledgements**

We would like to acknowledge the following individuals for their contributions to this project: Madeleine Shalowitz, Doriane Miller, Randall Doubet-King, Ellen Becker, Regina Greer-Smith, Jovana Ljuboje, James Fischer, and Mary Jane Welch.

**References**

1. ADAPTABLE, the Aspirin Study – A Patient-Centered Trial. Secondary ADAPTABLE, the Aspirin Study – A Patient-Centered Trial. http://theaspirinstudy.org/.
The Impact of Human Microbiome on Precision Medicine

Alexander V. Alekseyenko, PhD¹, Lita M. Proctor, PhD², Paul Carlson, PhD³, Scott A. Jackson, PhD⁴, Connie Chen, PhD⁵

¹Program for Human Microbiome Research, The Biomedical Informatics Center, Medical University of South Carolina, Charleston, SC; ²Trans-NIH Microbiome Working Group, National Institutes of Health, Bethesda, MD; ³Food and Drug Administration, Silver Spring, MD; ⁴Complex Microbial Systems Group, Biosystems and Biomaterials Division, National Institute of Standards and Technology, Gaithersburg, MD; ⁵The Health and Environmental Sciences Institute, Washington, D.C.

Abstract

The human microbiome consists of the organisms that accompany us at all body sites our entire life and contribute to the health and disease processes. Actionable associations of the host microbiota with health conditions span many health areas, such as oral and dental health, inflammatory bowel disease, metabolic syndrome, and other systemic conditions, and cancers. In addition, the value of microbiome for diagnostic and prognostic tasks is becoming more prominent. As such, microbiome is becoming a vehicle for precision medicine research. The underlying data enabling microbiome research are high-throughput, complex, and heterogeneous. Extensive informatics expertise is required at all translational steps (basic, clinical, consumer and public health) to understand and utilize these data for precision medicine. The purpose of this panel is to discuss the representative precision medicine efforts in the microbiome domain at academic, public, and federal institutions. Our goal is to understand the breadth of deployed expertise in these efforts and achieve roadmap for alignment of informatics efforts to maximize the broad positive impact on precision medicine.

Panel description

Microbiome is a term used to describe the collection of all microorganisms that colonize our bodies. In their numbers the bacteria are at parity with the mammalian cells in our bodies. Functionally, our microbial ‘second genome’ contributes upwards to 100 times the functional genomic diversity. Whereas understanding the human genome has received enormous amount of attention in the last 15 years, the impact of host associated microbiome genomes (the metagenome) and their function in human health and disease is still in very early stages of being conceptualized. The recent surge in microbiome science has been afforded by increasing availability of technologies, such as high-throughput DNA sequencing, mass spectrometry, automated cell sorting, etc. Most notably, of course, the ability to sequence hundreds of individual microbiomes in a single run of a sequencing instrument has resulted in a perfect storm of microbial community profiling datasets addressing a vast array of biomedical questions. In addition to being a vehicle for hypothesis generation in basic research, increasingly, microbiome is being utilized for translational tasks. Specific examples include hospital associated infection management and more recently building personalized precision medicine approaches for clinical decision support.

The complexity of the microbiome itself and the sophistication of the microbiome assay technologies necessitates involvement of quantitative scientists in many branches of knowledge including informatics. Many data challenges are still outstanding in the microbiome arena. For example, there is currently a gap in standards and standard reference resources for microbial identification. Standardization of microbiome measurement is still to be achieved. This results in lack of comparability of outcomes of different experimental groups. Poor reproducibility of results is a major barrier to translating the findings into useful application. Another translational roadblock is the fact that the experimental microbiome data is typically far removed from the health record data. A standard for integration of microbiome assay results into clinical reports does not currently exist. Finally, microbiome research is maturing to include true epidemiological scale studies involving thousands of subjects. This brings the field to novel challenges all too familiar to medical informatics community, including data standardization, management, integration, security, etc.

This is the third panel on microbiome research presented to the biomedical informatics community. The first panel on “S09: The Human Microbiome: Informatics Challenges and Opportunities” has taken place at the AMIA Annual Symposium 2016 in Chicago and provided a high-level overview of the challenges in the microbiome research where informatics can make a difference. The main goal has been to introduce the concepts in microbiome to the medical informatics community and invite collaboration among AMIA members. The panel has been attended by over 70 individuals including the 2017 AMIA Annual Symposium keynote speaker, Jessica Richman. The panel has been...
followed by an informal social, ‘Probiotic breakfast’ attended by about half dozen of AMIA members. The second panel on “IIT02: Tools for Accessible and Reproducible Computing in Microbiome Research” at AMIA Summits 2017 will specifically address implementation science in microbiome research and focus on tools that enable reproducible computing and lower the bar for biomedical investigators to perform their own analyses.

This goal of the proposed panel is to highlight translational microbiome research at academic institutions, public, and federal agencies. Specific focus will be on how this research links to precision medicine, as well as the ways that this research will have a lasting formative impact on human health. The secondary goal of the panel is to strengthen links between government agencies, public organizations, and medical informatics community. A likely outcome of this will be standing communication and collaboration between these communities.

**Brief description of each panelist’s presentation**

- **Dr. Alekseyenko**, “Microbiome as a platform for precision medicine at population level”
  This presentation will focus on our efforts at MUSC to measure microbiomes of our entire in-patient population and to utilize these data for development precision medicine approaches to triage and intervention in common clinical conditions, growing new branches of our learning health system.

- **Dr. Proctor**, “NIH Human Microbiome Project and Trans-NIH Microbiome Working Group”
  This presentation will highlight the funding history of the microbiome at NIH including conception, execution, and results of HMP and current state of the HMP 2.0: the Integrative Human Microbiome Project. The presentation will highlight future ways that NIH is supporting translational microbiome research as it relates to precision medicine.

- **Dr. Carlson**, “Microbiome research at Food and Drug Administration”
  This presentation will discuss research related to the microbiome that is ongoing across FDA centers. Projects focused the microbiome in infectious disease resistance, food safety, antibiotic utilization, nanoparticles, and tobacco use will be discussed.

- **Dr. Jackson**, “On the state of standardization of microbiome measurements”
  In 2016, the NIST has launched an initiative on standardization of microbiome measurements. This presentation will focus on the challenges and issues involved. The potential roles for informatics in this process will be discussed.

- **Dr. Chen**, “Emerging issues subcommittee on the role of microbiome in human health”
  Microbiome varies across individuals and is documented to interact with drugs and environmental chemicals. This is a specific example of factors beyond genetics that defines personalized response to such exposures. This presentation will discuss a new effort by HESI to foster collaboration on microbiome in toxicity, disease and drug interactions.

**Relevance of topic and anticipated audience:**

The White House precision medicine initiative, which is likely to shape the research funding policy in the next few years, specifically calls microbiome an important characteristic for Precision Medicine. In 2016, the White House Office of Science and Technology Policy has announced the Human Microbiome Initiative “to foster integrated study of microbiome across different ecosystems”. The initiative calls for supporting interdisciplinary research in microbiome and development of platform technologies to enable such research. With changing administration, the importance of microbiome as a precision medicine modality will likely to continue being recognized and valued. We anticipate that this panel will be interesting to attendees working in the translational informatics realm, specifically those focusing on precision medicine.

**Discussion questions:**
1. What resources to support translational microbiome science currently exist?
2. What key expertise is lacking or inadequately present on the microbiome arena?
3. What translational value does the microbiome hold and in what specific areas of medicine and health?
4. What are the national and regional centers for microbiome research and what opportunities for involvement exist?
5. How is microbiome science incorporated in biomedical informatics training?

**Learning objectives:**
1. Recognize the existing technologies to survey the human microbiome and the biological aspects they capture;
2. Understand the history and development of microbiome research;
3. Recognize the potential and limitations of host-microbiome data for precision medicine;
4. Identify key issues with standardization of measurement and data sharing of microbiome data;
5. Identify translational areas which can be positively affected by better characterization of human microbiome.

The organizer confirms that all participants have agreed to take part on the panel.
Redesigning the “Choice Architecture” of the EHR to Reduce Clinical Decision Support Burden
Jessica S Ancker, MPH, PhD,1 Sameer Malhotra, MD, MA,1,2 Yiye Zhang, PhD, MS,1
Adam Cheriff, MD1,2
1. Department of Healthcare Policy & Research, Division of Health Informatics, Weill Cornell Medicine, New York, NY; 2. Department of Medicine, Weill Cornell Medicine, New York, NY

Abstract
In this panel, experts will illustrate how the design of health information technology can be optimized for better decisions, with a specific focus on ways to minimize the need for alert-based clinical decision support. These design interventions exploit the technology’s “choice architecture,” a term from cognitive psychology which describes the way a system presents choices to decision-makers. (For example, the US “choice architecture” presents organ donation as an opt-in decision; other countries present donation as an opt-out decision in which anyone who does not sign an opt-out registry is eligible to be a donor.) Through three case studies, panelists will demonstrate how decisions can be improved by choice architecture in the context of electronic prescribing, order sets, and patient portals. After the panel, participants will be able to: define “choice architecture”; describe how “choice architecture” interventions can provide alternatives to clinical decision support alerts; analyze ethical implications of intervening in the “choice architecture”; describe several successful EHR “choice architecture” interventions.

Introduction
One of the central goals of medical informatics is to facilitate better medical decision making. A wealth of research and practice has examined ways to improve decisions through electronic clinical decision support (CDSS).1-3 Despite the successes of CDSS, its effectiveness may currently be limited by overuse. Contemporary research shows that clinicians now override or ignore more than 90% of alerts they receive.4,5 Alert effectiveness may be impaired by high rates of false alarms, delivery approaches that interrupt workflow, a reliance upon relatively effortful cognitive processes such as reading, and other design flaws.1,6-8 Combined with the sheer volume of alerts in contemporary systems, these failings lead to “alert fatigue”9-11 and habitual override behavior.12 Less intrusive approaches are needed if EHRs are to live up to their potential to support effective decision-making.

A powerful and relatively poorly exploited influence on decisions is the “choice architecture” of the system. “Choice architecture,” a term originating in cognitive psychology and behavioral economics, describes the ways in which a social or technological system presents choices to decision-makers.13 The same choice may be presented in many ways, and these ways strongly influence decisions.13-18

One feature that affects how people perceive decisions is the default option. In countries where all individuals are organ donors by default, those who do not wish to donate may sign an opt-out registry. Organ donation rates are far higher in these countries than in others (such as the US) in which the choice is presented to citizens as an opt in.15 Similarly, savings rates in 401(k) plans are highest when new employees are defaulted to participate but allowed to opt out.19 A meta-analysis shows that the effect of changing a default is large and robust, with an effect size that averages 0.36 (i.e., a 36% difference between two proportions).20

When defaults and other components of the choice architecture are thoughtfully examined and altered, they can provide unobtrusive “nudges” to promote good decisions without constraining choice. This didactic panel will present several viewpoints on the choice architecture of health information technology; demonstrate ways in which defaults can be exploited to improve medical decisions; and, importantly, illustrate situations in which defaults can reduce the workload burden of EHRs by replacing or minimizing the role of CDSS.

Moderator: Adam Cheriff, MD, MS, chief medical information officer, Weill Cornell Medicine

1. Redesigning Choice Architecture: How does it work, and is it ethical? (Jessica S. Ancker, MPH, PhD)

A customer downloads a new map app on his phone. The app has three audible narration settings: a woman’s voice, a man’s voice, and no narration. The first time the customer opens the app, the setting is defaulted to the woman’s voice, and the customer proceeds to use the female narration. In fact, the majority – and sometimes the large majority – of users will do the same. In this talk, we will discuss the three mechanisms through which the default setting exerts its effects on decisions.21 The first is effort: by definition, the default option is the easiest, and anyone who does not make a choice will end up with the default. The second is endorsement: the default is perceived as an
implicit endorsement by authority. The third is endowment: often, the default option is perceived as something owned by the decision-maker, and overriding a default may feel like giving up something that the he/she is entitled to.21

Any intervention to alter medical decisions must be ethical, and that requirement is particularly important with highly effective interventions such as “nudges” that do not fully depend upon conscious analysis by the decision-maker. In this talk, we will analyze whether and under what circumstances “nudges” might be unethical, i.e., threatening autonomy, dignity, or welfare. We will also discuss recent research showing that “nudges” do not require deception, suggesting that they may and in fact should be freely disclosed to decision-makers. Finally, we will examine the ethical applications of failing to thoughtfully design a system’s choice architecture.

2. Increasing the Generic Prescribing Rate through Defaults (Sameer Malhotra, MD, MS)

Generic medications are not only cost-effective but also improve patient adherence because of their lower co-pays. In determining how to encourage clinicians at our medical center to use more generics, we realized that education and persuasion have limited effectiveness in changing prescriber behavior, and CDSS alerts are already probably oversized. Instead, we redesigned our e-prescribing interface to automatically substitute generic equivalents for brand-name medications during order entry (while allowing a one-click override for those who wished to order the brand-name medication).22 In this talk, we will describe how we accomplished this intervention and report the retrospective cohort study conducted to assess its effect. We found that among drugs with generic equivalents, the proportion of generic drugs prescribed more than doubled, rising dramatically from 39.7% to 95.9% (a 56.2% increase; P < .001). The higher rate was sustained after the intervention, with virtually no clinician complaints. Over time, prescribers became more likely to search by the generic name rather than the brand name, suggesting a learning effect over time. We will place this project in context of similar work23 demonstrating the universality of this type of intervention.

3. Paving the Cowpath: Exploiting defaults to optimize efficiency in order sets (Yiye Zhang, PhD)

When clinicians use computerized provider order entry (CPOE), they generally have a choice between placing orders a la carte or with the help of an order set. Order sets are a powerful component of the EHR choice architecture. For example, they can provide unobtrusive “nudges” in favor of evidence-based practice if they are constructed to promote recommended practices. They can also reduce workload by making it easier to select multiple relevant orders with a single click.24 We will present recent work25,26 on setting the default options within order sets on the basis of analysis of EHR data to identify previous physician choices about order placement. Our research finds that number of mouse clicks and cognitive workload associated with current ordering practices can be reduced substantially across medical conditions and severity levels through the heuristic optimization of order set content and defaults, thereby contributing to more efficient care. Current work further shows that combining experts’ knowledge with data-driven evidence in setting default options may further improve both the efficiency and scientific validity of order sets.

4. Reducing Disparities Through Defaults (Jessica Ancker, MPH, PhD)

It is widely accepted that giving patients access to their medical records via portals may help them manage their healthcare. However, it is also well-known that patients from socioeconomically disadvantaged groups are less likely to take advantage of portals. This raises the possibility that health information technology may disproportionately benefit patients who are already educated or affluent enough to have computer access. In this presentation, we will describe a “choice architecture” intervention that massively reduced socioeconomic disparities in portal access at a federally qualified health center (FQHC).27 In 2011, a study demonstrated that white and insured patients were much more likely to use the portal. It was initially believed that these patients might be more literate or computer-literate. However, further analysis showed that the institution was disproportionately likely to offer portal access to these patients, and most patients who received an offer accepted it. The organization responded by abandoning this “opt in” model in which portal access had to be requested. Instead, the default was changed to an “opt out,” in which portal accounts were set up for all patients except those who requested not to have one. A subsequent 2015-16 study showed the steady elimination of disparities between black and white patients, and between Hispanic and non-Hispanic patients. Changing the “choice architecture” improved equity in access to information technology.27

Learning objectives: By the end of this discussion, participants will be able to: (1) Define “choice architecture”; (2) Describe how “choice architecture” interventions within the EHR may provide less intrusive alternatives to alert-based clinical decision support; (3) Analyze ethical implications of intervening in the “choice architecture” of health information technology; (4) Describe several successful EHR-based “choice architecture” interventions.

Statement: All participants have agreed to take part on the panel.
Discussion points:

1. Are “choice architecture” interventions deceptive? Are they manipulative?
2. When is it appropriate to promote optimal decisions through “choice architecture,” and when would it be more appropriate to promote decisions through more cognitively effortful interventions such as education?
3. Who is responsible for the “choice architecture” of the electronic health record?

References

15. Nease RF, Frazee SG, Zarin L, Miller SB. Choice architecture is a better strategy than engaging patients to spur behavior change. Health Affairs. 2013;32(2):242-249.
The Good, The Bad, and The Ugly of Deploying and Adopting Machine Learning Based Models in Clinical Practice

Yin Aphinyanaphongs, MD/PhD\textsuperscript{1}, Parsa Mirhaji, MD/PhD\textsuperscript{2}, Berkman Sahiner, PhD\textsuperscript{3}, David Holmes III, PhD\textsuperscript{4}, Michael Draugelis\textsuperscript{5}

\textsuperscript{1}NYU Langone, New York, NY; \textsuperscript{2}Albert Einstein College of Medicine, New York, NY; \textsuperscript{3}US Food and Drug Administration, \textsuperscript{4}Mayo Clinic, Rochester, MN; \textsuperscript{5}Penn Medicine, Philadelphia, PA

Abstract

Over the past three decades, the growth of publications involving machine learning has grown in the healthcare literature [1]. More recently, to leverage the potential benefits of these machine learning based models, institutions have begun making investments in infrastructure to translate this academic research into clinical practice. Deployment and adoption of models in clinical practice is not trivial and many challenges and issues emerge during implementation. This panel will provide a broad view of the machine learning model pipeline through model inception, evaluation, deployment, adoption, and FDA regulation. Through a series of case studies, the panel will demonstrate the challenges and issues. The goal of this session is to form a body of work for best practices, establish a like-minded community, and demonstrate potential scientific research opportunities toward successful model deployment and adoption.

Learning Objectives

After participating in this session, attendants should be able to:

- describe challenges and issues in successful deployment and adoption of machine learning based models.
- describe operational principles supporting the production deployment of machine learning based models.
- describe the machine learning pipeline.
- define a strategy for promoting machine learning model adoption and engaging the clinical teams.
- describe methodology for evaluating the impact of these machine based models.
- describe the design of predictive analytics systems in anticipation of meeting FDA regulations.

General Description

Precision medicine [2] and precision delivery [3] are two emerging approaches to healthcare delivery. Both efforts propose to target care delivery toward those who can most benefit from an intervention. At its core, this targeting task is what every healthcare provider does when a patient walks into our offices. We gather patient data and allocate our resources (medicines, diagnostics, etc) to uncover and resolve underlying pathology. Precision medicine and precision delivery propose to expand this paradigm with omics and clinical data respectively.

Over many decades, researchers published papers and models with excellent discriminatory performance on multiple clinical problems. In the clinical domain, these models have typically been described as calculators. For example, PORT [4] and Reynolds [5] are two clinical calculators. Despite decades of literature, widespread adoption has been limited and a systematic deconstruction of the barriers and approaches to overcoming them have been limited. Multiple barriers may account for the limited adoption. One particular barrier was the difficulty in entering data into these calculators. With the advent of electronic health records and digital recording devices, this barrier is lower and it is now possible to build models more agilely from this data. Concurrently with the digitization of health data, massive gains in pattern recognition and machine learning have allowed massive analysis of this data. We are once again at a crossroads with many models being published and fewer being deployed.

Our premise is that barriers for pragmatic model implementation still exist between the academic research that demonstrates excellent discriminatory performance for multiple clinical problems and the translation of these models into routine practice. Naively it would seem easy to believe that translating a published model into practice is as easy as deploying a model on a server and feeding it variables and reporting the outcome. In practice, the challenges to
deployment and adoption of models are not trivial. This panel will highlight the lifecycle of a machine learning based model through the experiences of panelists deploying and implementing models at their respective institution. Through case studies, we will each highlight issues, challenges, and scientific research opportunities associated with deployment and adoption of clinical models.

**Description**

The panel will be organized as follows:

<table>
<thead>
<tr>
<th>Time</th>
<th>Speaker</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>14 minutes</td>
<td>Aphinyanaphongs</td>
<td>Predictive analytics in medicine and the predictive model pipeline.</td>
</tr>
<tr>
<td>14 + 4 minutes</td>
<td>Draugelis</td>
<td>Knowledge management to enable consistent model performance and real-time predictive system from a vast and diverse data stream</td>
</tr>
<tr>
<td>14 + 4 minutes</td>
<td>Holmes</td>
<td>Data input validation and post-implementation evaluation.</td>
</tr>
<tr>
<td>14 + 4 minutes</td>
<td>Mirhaji</td>
<td>Drivers of model adoption: human factors, workflow integration, education, interaction design.</td>
</tr>
<tr>
<td>6 + 2 minutes</td>
<td>Sahiner</td>
<td>Deploying analytic systems in anticipation of FDA regulation.</td>
</tr>
<tr>
<td>14 minutes</td>
<td>Aphinyanaphongs</td>
<td>Panel Discussion with Audience.</td>
</tr>
</tbody>
</table>

Each panelist will anchor their content with a case study of a predictive model that they have built, deployed, or evaluated at their institution.

**Aphinyanaphongs:** Dr. Aphinyanaphongs is the Director of Clinical Predictive Analytics at NYU Langone Medical Center. His job is to build, deploy, monitor, and evaluate machine learning based models for optimizing clinical operations and care. His team includes data scientists and software engineers that build sustainable infrastructure to translate clinical needs and research findings into hospital solutions. Dr. Aphinyanaphongs will moderate the session and discuss the overall lifecycle of a predictive model. He will discuss NYU’s experience in producing model diagnostics to understand the internals of a congestive heart failure classification model.

**Draugelis:** Mr. Draugelis is Penn Medicine's Chief Data Scientist where he leads the exploration and development of predictive healthcare applications. Michael's previous position was Lockheed Martin's Chief Data Scientist where he led application development in Missile Defense, National Intelligence, Healthcare and Cyber Defense. He will discuss the importance and methods of knowledge event management to enable consistent performance predictive applications.

**Holmes:** Dr. Holmes is a Collaborative Scientist in the Department of Physiology and Biomedical Engineering. His research focuses on the fusion of heterogeneous datasets and advanced analytic techniques to improve diagnostics and therapeutics. He works with industry, scientists, and practitioners to develop optimized processing pipelines to generate findings in “useful” time. Dr. Holmes will discuss data provenance and quality of inputs to the predictive models in the context of Mayo’s Unified Data Platform. Additionally, he will discuss approaches to model evaluation with a case study of in-hospital patient monitoring.

**Mirhaji:** Dr. Mirhaji is Associate Professor of Systems and Computational Biology at Albert Einstein College of Medicine, and the Director of Clinical Research Informatics at Montefiore Medical Center. He will discuss challenges regarding integration of analytically informed workflow processes into delivery of care process in healthcare ecosystem, implementation science and human computer interaction aspects of introducing data driven decision support without distraction, disruption, or overloading providers, and multi-system integration and coordinations necessary to present a streamlined process. He will highlight the role of workflow integration,
education, human factors, interaction design that drive model adoption and ultimately the success of a predictive model through the case study of “Accurate Prediction of PROlonged VEntilation-APPROVE” project that predicts inpatient mortality and respiratory failure 48 hours in advance, and providing personalized best practice advisories just-in-time and through institutional EMR.

**Sahiner:** Dr. Sahiner is an Electrical Engineer by training and a senior scientist with the Office of Science and Engineering Laboratories at the Center for devices and Radiological Health at the FDA. He is the leader of the image analysis laboratory at the Division of Imaging, Diagnostics and Software Reliability, with a focus on the appropriate evaluation methodologies for computational algorithms used to aid medical practitioners in the interpretation of diagnostic device results. Prior to joining the FDA, he was an Associate Professor of Radiology at the University of Michigan. His research interests include computer-aided diagnosis, machine learning, image perception, clinical study design, and performance assessment methodologies. He will discuss the regulatory environment faced by machine learning based clinical models and how to factor in regulatory considerations while the models are being developed.

**Significance of panel topic and intended audience**

Data science and predictive analytics are taking increasingly large roles in health care. A plethora of research has demonstrated that machine learning models can identify patients for care prioritization and potentially reduce costs [6]. Translating these findings into actual practice is essential to realize their potential gains. Healthcare organizations have recognized this potential and are making investments in building infrastructure and teams. Thus, this panel topic is timely and needed. Our intended audience include machine learning model researchers who want to translate their papers into practice, data scientists within healthcare organizations, healthcare executives who want to assess infrastructure and conceptual concerns in making investments in predictive analytics, change managers and behavioral economists who can address the challenges of model adoption, user interface designers who build the end users experiences, and medical educators who train future health care providers.

**Discussion Questions**

- How do organizations manage the risk of deploying an unproven predictive model at a clinical institution?
- How do we incorporate predictive analytic education for our healthcare providers?
- Predictive models have existed for many years as calculators. These calculators are not widely used in practice even though they provide potentially significant prognostic information. What potential barriers prevented the adoption of calculators and are predictive models in the era of electronic health records any different?
- What mechanisms are possible to pool our collective knowledge and experiences in deployments to identify best practices?

**Participation Statement**

All proposed panelists have agreed to participate in the panel if the proposal is accepted.

**References**

Problem List 2.0

Joel R. Buchanan, MD\textsuperscript{1}, William L. Galanter, MS, MD, PhD\textsuperscript{2},
DuWayne L. Willett, MD, MS\textsuperscript{3}, Adam Wright, PhD\textsuperscript{4}
\textsuperscript{1}University of Wisconsin School of Medicine and Public Health, Madison, WI; \textsuperscript{2}University of Illinois Hospital \& Health Sciences System, Chicago, IL; \textsuperscript{3}University of Texas Southwestern Medical Center, Dallas, TX; \textsuperscript{4}Brigham \& Women’s Hospital, Boston, MA

Abstract

The value of a comprehensive, concise list of all a patient's active medical conditions to provide accurate "situational awareness" during clinical decision-making has long been recognized. EHRs offer the potential of centralizing the patient problem list, streamlining its maintenance, enhancing its accuracy, and providing flexible viewing to fit clinical information needs. Yet many practices and healthcare organizations today struggle with realizing these benefits. Why is that? What can be done to unlock this potential?

In this panel, we first briefly review the evolution of the problem list in its transition from paper to the electronic health record and summarize current challenges. Then we explore the use of clinical decision support and automated approaches for keeping up an accurate problem list. The recognized benefits of providing context-appropriate, problem-oriented views of relevant medical record information require both EHR technical capabilities and clinically-valid concept maps: we review both, and describe a project to develop the latter in the public domain. We then explore the wide range of potential applications for valid problem list representation of a patient’s health conditions, including clinical decision-making, clinical communication, quality measurement, registries, and clinical and translational research.

Description

The overall outline of the talk is given in the abstract. The panel will be organized as follows:

\begin{table}[h]
\centering
\begin{tabular}{|c|c|p{10cm}|}
\hline
\textbf{Time} & \textbf{Speaker} & \textbf{Topic} \\
\hline
10 min & Willett & Introduction, overview of current state of the Problem List in the EHR era \\
5 min & & Discussion and questions from audience \\
15 min & Galanter & Toward automated maintenance of the Problem List \\
5 min & & Discussion and questions from audience \\
15 min & Buchanan & Problem-Oriented View (POV) \\
5 min & & Discussion and questions from audience \\
15 min & Wright & Additional uses of the Problem List \\
5 min & & Discussion and questions from audience \\
15 min & Panel & General discussion and questions from audience \\
\hline
\end{tabular}
\caption{Organization of panel by segment length, speaker and topic.}
\end{table}

The specific topics covered by each presenter will be:

1. Willett (introduction): Introduce the topic and panelists, including a brief description of what each panelist will be talking about and their affiliations. Provide an overview of the history of using a problem list in medical records, its evolution in the EHR era including context within the EHR, and common current issues\textsuperscript{1}.

2. Galanter: Describe methods to help automate problem list documentation. Demonstrate the use of clinical decision support (CDS) to prompt clinicians to provide problems while ordering medications and labs, as well as other methods which use CDS to help or directly place problems\textsuperscript{2}. The design of this type of decision support will be discussed briefly with an emphasis on the types of orders which lend themselves to this method, contrasted to those that work less well. The use of CDS to clean problem lists will also be discussed.

3. Buchanan: Context-dependent views of EHR data are considered essential for success of the Problem Oriented Medical Record, and implementation of these views will require programming by vendors and
creation of content maps by customers. We have created a methodology for obtaining expert consensus on
these maps, and we plan to create 140 vendor-neutral, open source maps in the next phase of the project.

4. Wright: Give an overview of the range of uses of problem lists, including the core use of clinical care, as well
as clinical decision support, quality measurement, clinical and translational research, registries, supporting
teamwork and documentation. Dr. Wright will also discuss the impacts of inaccurate or incomplete problem
lists on each of these uses.

5. Panel: After the presentations, the panel will briefly discuss additional issues and then open for questions
from and discussion with the audience. Audience members will be invited to address questions to specific
panelists, or to the panel as a whole

Importance

The need for accurate digital knowledge of a patient's health conditions continues to expand rapidly. In addition to
its core benefit in providing complete clinical context for individual medical decision-making, knowing the list of a
patient's currently active health issues—their Active Problem List—now has become valuable in several other contexts
as well. The patient's list of active health conditions is now shared (a) with other organizations via Health Information
Exchanges, reflecting on the sender's clinical assessment, and (b) with patients via patient portals. When accurate,
this same list can also help automatically define sub-populations for clinical registries and electronic quality
measurements. Advanced analytics as well as clinical and translational research likewise benefit from an
unambiguous record of the patient's active health conditions.

Yet at the practical day-to-day level of medical practice, challenges remain with populating and updating the Problem
List, and with optimizing the display of problem-related medical information stored in the EHR in visualizations that
enhance clinical decision-making.

In this panel we summarize current challenges with the EHR problem list, and then offer perspectives from three
organizations on approaches to a) enhancing the population and grooming of the problem list, b) optimizing a problem-
oriented display of information to maximize the clinical signal-to-noise ratio for decision making, and c) harnessing
the value of knowing our patients' health conditions to drive improvements in patient and population health.

We believe this topic, along with an opportunity for open discussion and audience questions, will prove valuable for
medical informaticists, practicing clinicians, and anyone interested in secondary uses of EHR-derived data for quality
improvement and analytics.

Discussion Questions

1. How well is the problem list used currently at your institution, and does it vary between primary care and
specialists?
2. What incentives do providers at your organization have to maintain accurate problem lists? (e.g. P4P, part of
being a good team player, makes it easier to document, etc.)
3. What can leaders do to create a culture of problem list usage in their organizations?
4. What are the limits on the use of CDS for automated and semi-automated problem list placement?
5. What do clinicians and administrators think about trying to automate the problem list?
6. If indications are used for ordering labs and medications, is there still a need for problem concept maps to
drive an aggregate data display?
7. If we create problem concept maps for common problems, how can we encourage vendors to create the
necessary software?

Participation Statement

All proposed panelists are aware of this panel submission, and have agreed to participate in the panel if the proposal
is accepted.
References


Social Determinants of Health: Applied Informatics Approaches
Incorporating Context into Care

Hossein Estiri, PhD¹, Rachel Gold, PhD, MPh², Michael N. Cantor, MD³, Theresa Cullen, MD, MS⁴,⁵

¹Harvard Medical School, Boston, MA; ²Kaiser Permanente Northwest Center for Health Research, Portland, OR, and OCHIN, Inc.; ³NYU School of Medicine, NY, NY; ⁴Global Health Informatics, Regenstrief Institute, Indianapolis, IN; ⁵Indiana University School of Medicine, Indianapolis, IN

Abstract
There is substantial empirical evidence that many of the factors that contribute to health outcomes originate from outside of the conventional confines of the U.S. healthcare system. Social Determinants of Health (SDH) constitute the majority of these factors. The uptake of Electronic Health Record (EHR) systems offers great potential for addressing such socio-contextual determinants of health. Integrating SDH data into EHRs would provide invaluable information for healthcare organizations, helping to improve precision in diagnosis, treatment, and monitoring of disease and care management programs. Yet, the current healthcare system generally lacks the infrastructure to assimilate SDH data within clinical practice and research. This panel will discuss theoretical aspects and practical steps for developing informatics infrastructures to incorporate data on socio-contextual determinants into healthcare practice and research.

Keywords:
Social and Contextual Determinants of Health; Biomedical Informatics; Electronic Health Records; GIS

Introduction
Most determinants of health originate from the socio-contextual settings wherein we live (i.e., the social, cultural, historical, and economic settings, and the built, natural, and institutional environments), which are often outside the confines of disease-based healthcare.[1–6] The WHO defines social determinants of health (SDH) as “the conditions in which people are born, grow, work, live, and age, and the wider set of forces and systems shaping the conditions of daily life. These forces and systems include economic policies and systems, development agendas, social norms, social policies and political systems.” Since the early 2000s, several national advisory committees have highlighted the need for addressing SDH in clinical settings and made explicit recommendations on incorporating SDH into health care. However, in most cases, the U.S. healthcare system still lacks the resources and motivation needed to assimilate SDH into healthcare practice and related research.[7–9] As a result, many healthcare providers operate without fully understanding the challenges to optimal health that their patients may face, as those challenges fall outside of their historical purview.

Accountable Care Organizations and the Precision Medicine Initiative promote population health management both as a cost containment strategy and as a means of detecting disease earlier, understanding it better, and intervening more precisely. Electronic Health Record (EHR) systems are key to transforming health care from an episodic, reactive model to a more preventive, proactive system. To implement more efficient population health management, healthcare providers need point of care access to a full range of information about their patients. Yet while EHRs provide valuable information about clinical determinants of health and treatment effectiveness,[4] it is rare that they are configured to systematically capture SDH information.[3,10]

A few studies have recently applied geospatial technology to integrate SDH data with EHRs, arguing that integrating EHR data with SDH data would help healthcare organizations improve precision in diagnosis, treatment, and monitoring of disease and care management programs.[3,5,11] Attaining these potentials, however, is contingent on availability of health information technology that integrates EHR data with SDH. Also, little is yet known about best methods for synthesizing, interpreting, and intervening on SDH. We lack empirical evidence on the clinical utility of SDH data – i.e., evidence that SDH can be directly applied in clinical interventions / decision-making. In addition,
data sharing, governance, and privacy concerns introduce new challenges that discourage healthcare organizations from integrating data from outside into their clinical repositories.

This panel is sponsored by the AMIA Social Determinants of Health (SDH) discussion group.

**Discussion Questions:**

- Who are the social and institutional actors/stakeholders, and their roles and relationships in a context-informed healthcare system (i.e., one in which patients’ social and cultural context is considered in clinical care decisions)?
- What are available/potential sources of social and contextual information, and related data integration technologies?
- Can distributed clinical data networks facilitate integration of EHRs with SDH data?
- What is needed to incorporate the integrated SDH data into clinical processes?
- What is known about how patient-reported or contextual SDH data can be used in clinical practice?

**The Panel**

This panel seeks to discuss practical steps and opportunities for developing informatics infrastructures to incorporate socio-contextual determinants data into healthcare practice and research. For this panel, participants will describe their experience in integrating EHR data with social and contextual data and will discuss lessons learned, informatics requirements, and decision support needs that can be applied in real-world clinical settings.

**Dr. Estiri** will use the social and ecological theory as an organizing framework to identify and discuss social and institutional actors/stakeholders, elements, and their interactions in a context-informed healthcare system. Based on this framework, he will discuss informatics questions that need to be addressed to operationalize such a system. He will describe the experience, as well as socio-technical challenges and opportunities, of architecting an informatics infrastructure for incorporating social and contextual data into Electronic Health Records (EHR) from a large-scale distributed clinical data network at HMS.

**Dr. Gold** will discuss her team’s pilot study of the development and implementation of EHR-based tools for collecting and acting on patient-reported SDH data. This is the first federally-funded study of strategies for SDH data collection among adult patients in the primary care community health center setting – a setting with highly vulnerable patients, where SDH have profound impacts. This talk will describe implementation barriers and lessons learned in this pilot study; these lessons may be informative to others seeking to implement patient-reported SDH data collection. Dr. Gold will also discuss her team’s theoretical model for using SDH to augment clinical data, and their plans for future research on how to use such data to connect primary care teams to community services that might address SDH.

**Dr. Cantor** will discuss the experience of NYU Langone Medical Center’s efforts around capturing and acting on individual and community-level SDH, and efforts on integrating the two. For individual-level determinants, he will discuss NYULMC’s work with New York State’s DSRIP (Delivery System Reform Incentive Payment) program which aims to improve care for Medicaid patients. NYULMC has taken several different approaches to both capture and integrate individual-level SDH into treatment and care management approaches, including partnering with and referring patients to community-based organizations. NYULMC is also using community-level SDH both within its network and as part of the NYC-CDRN, our local instance of PCORnet. We are using publicly-available, census tract level data, captured in the FACETS (https://github.com/mcantor2/FACETS) database, to evaluate the impact of community-level determinants on various medical outcomes.

**Dr. Theresa Cullen** will discuss her experience with developing screening tools for SDH and associated analytics in Indian Health Service. Starting in 2004, the Indian Health Service, at the behest of tribal communities, responded to community needs to capture data on domestic violence as well as sexual assault. Dr. Cullen will describe how SDH screening tools, developed in conjunction with endabuse.org (formerly the Family Violence Prevention Fund), were integrated into RPMS, the Indian Health Service electronic health record. This process, a joint effort between tribal communities and the Indian Health Service, is a model for shared decision-making with communities. Screening tools included attention to safety, including ‘opt-out’ options, non-inclusion of screening data on health summaries as well as extra security requirements for aggregating and exporting the data. Screening rates for
domestic violence are tracked; outcome reports are generated by tribal communities but are not available for research or publication without tribal IRB approval.

Following the panelists’ presentations, the moderator will start the discussion by asking a few questions to help the audience understand the available data and technology resources, and socio-technical challenges for integrating social and contextual data with EHRs. We will then open the floor for a discussion of the practical informatics steps and possibilities for addressing social and contextual data in healthcare research and practice.

**Learning Objectives**

After participating in the session, the attendant should be able to:

- Identify elements (i.e., actors/stakeholders, technology, and interactions) of a context-informed healthcare system.
- Describe available data and technology resources for integrating social and contextual data into EHRs.
- Discuss possibilities, challenges, and concerns related to integration of social and contextual data with EHR data.

**Panel Organizer Statement**

All participants have agreed to take part in the panel.

**Conflict of interest**

The participants have no conflict of interest to declare.

**References**

The Data and Research Center of the All of Us Research Program: Framework for a National Cohort Program and Research Opportunities

Robert J Carroll, PhD¹, Joshua Mandel, MD², Karthik Natarajan, PhD³, Scott Sutherland⁴, Joshua C Denny, MD MS¹

¹Vanderbilt University Medical Center, Nashville, TN; ²Verily, Cambridge, MA; ³Columbia University, New York, NY; ⁴Broad Institute of MIT & Harvard, Cambridge, MA

Abstract

The Precision Medicine Initiative All of Us Research Program is a national effort to recruit and engage at least one million participants who consent to provide health information (including data from health surveys, electronic health records, and baseline physical measures), biospecimens, and to be recontactable in the future. The Data and Research Center (DRC) sits at the confluence of all of these varied data and, in collaboration with the many All of Us partners, has been creating a framework to shape and support biomedical research in the future. This panel will discuss a number of primary topics including operations support, the back end data structure, EHR data ingest, processing, and curation, and the researcher tools and public facing sites with a number of interwoven secondary topics including principles driving the creation of the All of Us Research Program, the ethical, legal, and social implications of the effort, and the privacy and security framework for protecting participants and their data. We are interested in engaging the AMIA community in discussions around the plans to engage and protect participants and empower this community of researchers.

Overview

In 2015, President Barack Obama first introduced his vision for the Precision Medicine Initiative (PMI) during his State of the Union Address. With a goal to recruit and engage one million participants representing the diversity of the United States, the PMI All of Us Research Program has received bipartisan support and has been an important focus of the National Institutes of Health. After months of work by individuals across a broad set of fields, a set of working recommendations were presented. The All of Us Research Program was launched in 2016 with a number of grants addressing key areas of the project including recruitment sites, a biobank, a participant technologies center, and the data and research support center (DRC). All of these groups are collaborating on solving the many barriers to creating such a large, nationwide cohort. As members of the DRC, the panelists have been involved in the many aspects of this project, including operations support, the back end data structure, EHR data ingest, processing, and curation, and the researcher tools and public facing sites. Josh Denny from Vanderbilt will be the moderator for this panel.

Operational Support

Robert Carroll will present on the topic of the DRC’s role in operational support for the All of Us Research Program starting with an overview of the participant enrollment process. The design, feedback, and implementation approach of the research associate facing HealthPro Portal used for in person enrollment will be the primary focus for this segment. Concerns around data sharing and visibility on a national program will be discussed in addition.

Raw Data Import and APIs

Josh Mandel will introduce the data storage architecture, domain models, and APIs that support the DRC. The discussion will incorporate a set of motivating use cases outlining communication requirements between the DRC and other All of Us stakeholders (Participant Technology Center, Biobank, and healthcare provider organizations). This topic will also include an overview of the use of the Fast Healthcare Interoperability Resources (FHIR) standard within the DRC, as well as the security model that protects participant data.

¹Precision Medicine Initiative, PMI, All of Us, the All of Us logo, and The Future of Health Begins With You are service marks of the U.S. Department of Health and Human Services.
Aggregation of EHR Data

Karthik Natarajan will present on the ingestion, processing, and curation of electronic health record data for the All of Us program. Topics of interest include the Observational Medical Outcomes Partnership Common Data Model (OMOP CDM), the Observational Health Data Sciences and Informatics (OHDSI) collaborative and its open-source toolkit, and solutions to integrating multiple implementations of the OMOP CDM into a single, usable dataset. Additionally, this topic will cover the quality control processes employed on such data in order to deliver high-quality EHR data for researchers.

Data Use and Research Support

Scott Sutherland from the Broad Institute of MIT & Harvard will present on public facing and researcher tools. This segment will include public dashboards, researcher support, tools available in the researcher portal, and data access mechanisms. Special focus will be given to the privacy and security policies and how researchers can access and use the data from the All of Us Research Program.

Relevance

Simply by the merit of being one of the largest programs initiated by the United States National Institutes of Health, the All of Us Research Program is likely to be of interest to a diverse set of attendees. In addition, each topic appeals to a broad, potentially different, set of informatics researchers. Implementation and evaluation of an operational electronic support system, the design of interoperable systems communicating health data using a number of major current standards, the management and quality control of multi-site EHR data, and the design and availability of a researcher portal are all key areas of interest to AMIA members. It is currently anticipated that roll out of research tools using this data may be available, providing some introduction to the pipeline and data available (or soon to be available) for use generally.

Discussion Questions

1) Since the All of Us Research Program will be delivering many petabytes of data into a cloud-based scalable computing environment, what kinds of tools should we provide to minimize the need to download data to local systems.
2) What are the opportunities and concerns with integrating health data from a number of systems, including differing EHR sources, participant self report data, and other national systems?
3) Are there sources of participant level data that would be useful to integrate within the DRC for use by researchers?
4) Are there sources of reference data that would be helpful to provide in a research environment?
5) What level of quality control for the various types of data would be helpful? How can the DRC flexibly provide accurate data without excluding or preventing certain types of research, eg on noisier data?
6) What types of studies would be easily enabled in this framework?
7) Are there studies which would be difficult to perform in the current scheme? How could those be facilitated?
8) In what ways can data be provided back to participants that would encourage engagement in the program?

Participation Statement

All participants have agreed to take part in this panel presentation.
“My work will surely speak for itself:”
Visibility, Networking, and Self Promotion in Informatics
Sponsored by the Women in AMIA Taskforce

Wendy W. Chapman, PhD 1, Murielle Beene, DNP, MS, MPH, MBA, RN-BC, PMP, FAAN 2, Omolola Ogunyemi, PhD3, Genevieve B. Melton, MD, PhD 4, Laura K. Wiley, PhD 5
1 University of Utah, Salt Lake City, UT, 2 US Dept of Veteran Affairs, Atlanta, GA, 3 Charles R. Drew University of Medicine and Science, Los Angeles, CA 4 University of Minnesota; University of Minnesota Health and Fairview Health Services, 5 University of Colorado Anschutz Medical Campus.

Abstract. Women and men interested in better understanding how to obtain leadership positions and to enhance leadership opportunities for women within our field should attend this interactive panel sponsored by the Women in AMIA Taskforce. Women continue to be under-represented in leadership positions in medicine and within informatics. Contributing factors include a person’s confidence level and degree of networking. Our objective is to discuss the confidence and networking differentials between men and women and to share strategies for reconciling the fear many women (and some men) have of appearing over-confident with the need for self promotion and visibility needed to achieve positions of leadership and increase impact. The panel will focus on four areas:
(1) Selling yourself: the confidence spectrum; (2) Climbing the ladder: Academic career progression; (3) Women in the C-Suite; (4) Strategies for Students.

Intended Audience: The intended audience for the panel discussion includes all biomedical informaticists (both men and women) who are interested in obtaining leadership positions themselves and in enhancing leadership opportunities for women within our field. This topic is timely as we seek to increase the proportion of female nominations for awards and of leadership positions within our field.

Introduction of the Topic.
Women continue to be under-represented in leadership positions. In fact, men with moustaches outnumber women in medical school leadership. An informal survey of women in informatics leadership showed that 43% of NLM training programs have female PIs, 21% of CTSAs have a female BMI Core director, and only 10% of BMI departments have a female chair. Confidence level is potentially a contributing factor to this gap. Research shows that across the world men overestimate their intelligence, whereas women underestimate theirs. This confidence differentiation influences whether individual seek for or accept leadership opportunities.

259
Moreover, differences in displays of confidence affects reception by peers and other leaders. Male qualities are often equated with leadership qualities. Because people “commonly misinterpret displays of confidence as a sign of competence, we are fooled into believing that men are better leaders than women.”

**Aim of Discussion.** Our objective is to discuss the confidence differential between men and women and to share strategies for reconciling the fear many women (and some men) have of appearing over-confident with the need for self promotion, networking, and visibility needed to achieve positions of leadership and increase impact.

Panel presenters will provide background, share personal experiences, and lead interactive discussions with the audience on the following four topics:

1. **Selling yourself: the confidence spectrum (Wendy Chapman).** The motto of many women is “My work will surely speak for itself.” As much as we hope that may be true, the path to leadership requires promoting one’s work and one’s personal abilities, and women on the whole are less successful at this.
2. **Climbing the ladder: Academic career progression (Omolola Ogunyemi).** Is there a best approach to working towards promotion? How do we balance leadership opportunities with building our research labs? And what role do sponsors play?
3. **Women in the C-Suite (Genevieve Melton-Meaux).** There are a variety of leadership opportunities in health care settings. How many women are in the C-Suite? What are the particular challenges in this setting, and how do women make themselves more visible for these types of positions?
4. **Strategies for Students (Laura Wiley).** Leadership qualities can be learned and practiced from the very beginning of informatics studies. How can students effectively network and self promote to improve their visibility to broader informatics community? What opportunities are there for female students in AMIA leadership?

**Contributions of Each Speaker:** All participants have agreed to take part on this panel.

**Murielle Beene, DNP, MS, MPH, MBA,** is the CNIO and acting director, Office of Informatics and Analytics, Department of Veterans Affairs. Dr. Beene is the first CNIO for the Veterans Health Administration and has been a CNIO since 2009 and a nurse since 1996. Dr. Beene will moderate the interactive panel.

**Wendy Chapman, PhD, FACMI,** is the chair of the Department of Biomedical Informatics at the University of Utah. She discovered biomedical informatics serendipitously and fell in love with the field. Since receiving her PhD in 2000, Dr. Chapman was a faculty member at the University of Pittsburgh then UC, San Diego. She returned to her alma mater as the department
chair. Dr. Chapman serves as an elected member of the AMIA Board of Directors, as the chair of the AMIA Student Paper Awards Committee, and as the Chair of the Women in AMIA Taskforce. Dr. Chapman will discuss the role of confidence in career advancement. She will discuss research on gender differences in confidence, will describe some personal experiences, and will suggest an approach to confidence for women.

**Omolola Ogunyemi, PhD**, is Director of the Center for Biomedical Informatics at Charles R. Drew University of Medicine and Science (CDU) and an Associate Professor in the Department of Preventive and Social Medicine. She was introduced to the field of Biomedical Informatics while a computer science graduate student at the University of Pennsylvania. She served as an informatics faculty member at Brigham and Women’s Hospital for over seven years before joining the faculty at CDU. She is a member of the Women in AMIA Taskforce Steering Committee and AMIA’s Membership and Outreach Committee. Drawing on her background as a reviewer for NLM, NIBIB and AHRQ research grants and her experience serving on appointments and promotions committees, she will talk about career advancement strategies and potential pitfalls.

**Genevieve Melton-Meaux, MD, PhD,** is Chief Health Information Officer for Fairview Health Services and University of Minnesota Physicians. At University of Minnesota, she is an Associate Professor of Surgery, Core Faculty in the Institute for Health Informatics, and co-directs the Natural Language Processing-Information Extraction Program. After studying computer engineering and mathematics as an undergraduate, she went to medical school and while there was introduced to informatics and surgery. With experience as a successful academic surgeon, informatics researcher, and healthcare administrator, Genevieve will discuss administrative leadership in health care settings, research around women in the boardroom, and factors for success.

**Laura Wiley, PhD,** is an Assistant Professor in the Division of Biomedical Informatics and Personalized Medicine at the University of Colorado School of Medicine. She recently completed a PhD in human genetics (2016) and Master’s in biomedical informatics (2014) at Vanderbilt University. She is currently the student representative to the AMIA Board of Directors, Chair of the Working Group Steering Committee, and a member of the Women in AMIA Taskforce. Previously she has served as Chair of the Student Working Group and as a member of the AMIA Public Policy Committee. Dr. Wiley will discuss how students can effectively network and self promote to improve their visibility to broader informatics community. She will also describe ways that students can become involved in AMIA leadership (all of her listed AMIA roles began while she was still a student).
The Hashemite Kingdom of Jordan is located in the Middle East at the crossroads of Asia, Africa, and Europe with a population of over 9 million. In 2009 under the patronage of His Majesty King Abdullah II Bin Al Hussein, Electronic Health Solutions (EHS) was founded to advance the healthcare sector in the country through its flagship program Hakeem. The Hakeem program facilitates efficient, high-quality healthcare in the Kingdom through the nationwide implementation of an Electronic Health Record (EHR). Jordan selected an open source version of VistA. The Jordanian version of VistA, a collaboratively supported free and open platform adopted from World VistA, is currently in use in over 104 facilities (primary healthcare clinics, comprehensive healthcare clinic, and hospitals) across various public healthcare authorities in the Jordan Ministry of Health, Royal Medical Services, and King Hussein Cancer Center. The benefits of open, collaboratively developed health information technologies (HIT) have been well described, and strongly advocated for over the past decade. Many have found it difficult to successfully initiate, grow, and maintain these initiatives. The Jordan experience with WorldVistA EHR provides a potent reminder of the possibilities and promise of open source HIT as well as the possibilities and impact of full spectrum HIT in low and middle income countries (LMIC).

Keywords

Global Health, Clinical Information Systems, Clinical Systems Implementation

Introduction and Background

This panel is designed to share the experience of the last eight years of health information technology in Jordan from the vantage point of a non profit government supported health information technology company (EHS) that has helped guide the development, deployment, implementation, and assessment of a nationwide roll-out of an EHR within Jordan. This panel will discuss the history of the decision and the benefits that have accrued due to the use of an open source version of VistA (WorldVistA EHR) which has been customized and modified to meet the need of the various Jordanian public healthcare sectors.

The HIT solution, Hakeem, is based upon the Veterans Health Information Systems and Technology Architecture (VistA), an enterprise-wide information system used throughout the United States Department of the Veteran Affairs (VA) medical system. The Hakeem VistA-based structure selected based upon the needs of the country. VistA was customized and modified to meet Jordanian regulations, language, culture and standards. Members of EHS, the nonprofit organization supporting the HIT initiative, coupled with healthcare team members and other stakeholders, identified the local health sector needs and requirements for a customized implementation of VistA. EHS successfully developed needed functionality, implemented quality tests to ensure the capacity and efficiency of the system (i.e. drug information, localization, etc.) and developed robust deployment and implementation capacity. Ultimately, EHS has become a pioneer, locally, regionally and globally, in developing a technology solution for the national public health sector in low and middle income countries.

In 2017, the Hakeem initiative is engaged in the deployment of the Hakeem (VistA) open EHR in more than 75 new sites across the Jordanian public healthcare sectors. This scale of a project, leading to over 70% of the population...
receiving care at a facility that has deployed Hakeem by the end of 2017, represents successful collaboration between multiple stakeholders who shared the vision of transforming and sustaining a continuously improving healthcare system in Jordan by leveraging technology. EHS stakeholders include the Ministry of Health, Ministry of Information and Communications Technology, Royal Medical Services, King Hussein Cancer Center, Private Hospital Association, The Royal Health Awareness Society, and the King Abdullah II Fund for Development. Over the past eight years, EHS and its various stakeholders have worked side-by-side to elevate the healthcare process and achieve health equity by providing easy access to a complete electronic health record for patients of healthcare facilities within the Hakeem implementation network. This work reflects shared responsibilities and processes such as vision and mission, governance, technologies, industry standards, principles, and the gathering and utilization of resources. This shared commitment has helped ensure efficient access to vital patient information, resulting in improved patient care quality. Jordan is an exemplary example of a middle-income nation that has actualized improving a country’s health status through a commitment to health information technology. Jordan stands at the precipice of the next stage of data use - embracing the continuum of data maturity with a commitment to a complete analytics and business intelligence framework. The lessons learned from Jordan’s experience are important to global as well as US national health informatics work.

Panel Description

Dr. Terry Cullen will facilitate a panel discussion on the history and current status of VistA within Jordan. This panel will help the audience better understand the importance of open source software, the many collateral benefits that emerge from them, and the potential for using the Jordanian experience to help inform other health information technology journeys. The panel will include a review of how a decision was made to select VistA, and the various implementation activities and the factors that contributed to the current success. The panel will also discuss the historical and future management strategies employed by project leadership, which will provide details on the nuanced decision-making that has proven fundamental to the initiative’s success. These experiences and lessons learned from this global health informatics initiative are applicable to not only HIT development and implementation initiatives working under fiscal and human constraints, but to other countries who are embracing a commitment to health system strengthening and resiliency with limited resources. Anticipated audience includes those interested in enterprise wide deployments of HIT, those interested in the use of open source software, and AMIA members interested in seeing a successful mature HIT deployment in a LMIC.

Panel

Health Information Technology (HIT) in Jordan: Why VistA? (Feras Kamal)

Mr. Feras Kamal will present on the history of Jordan’s decision to use HIT to improve care at a national level, as well as the choice of VistA as the solution for this need. Mr. Kamal will share the last nine years of experience, culminating in the current situation where the Hakeem program supports care provision to over five million patients from all over Jordan. The Hakeem system enables healthcare professionals in clinics as well as hospitals to care for patients through electronically accessing the patient’s medical records from within participating health facilities by simply entering the patient’s national ID number (a unique identifier for all patients receiving treatment within the public healthcare sector). The creation of a HIT system responsive to the clinical needs of Jordan, as well as an integrated approach to deployment and implementation of a cloud based HIT solution, has proven to be a successful model for the Jordanians health care system. Mr. Kamal will also discuss the challenges of deployment in a middle income country, as well as Jordans approach to security, privacy and confidentiality within the VistA deployment.

The Benefits of VistA (Nancy Anthracite)

A technical overview of the various VistA packages which have been implemented, as well as the packages currently in development and packages that have received extensive customization to enhance usability (Pharmacy GUI + Dental Package), will be presented. The Hakeem Team has developed a package enhancement process that is low cost, supports the needs of Jordan and can be a model for other countries. The panel presentation will emphasize lessons learned by EHS, including predictive factors for the success of an open source electronic health records deployment on a national level as well as a successful open source health information technology community.

Looking Forward – Analytics and Business Intelligence (Theresa Cullen)
EHS is a case study in how to use open source technology to improve health status of a country. As a result of the successful deployment of the Hakeem program, EHS has moved to the next phase of data use with a commitment to quality and business intelligence. Following a rigorous alternative review, EHS elected to use an open source dB, Maria, for their ABI platform. Currently in its infancy, this platform is designed to support the production of key performance measures as well as multiple data marts. Ideally, EHS will soon produce national level indicators and measures that reflect the needs and priorities of the Kingdom. Plans for the future, as well as the use of analytics and business intelligence to drive future implementation projects and improve health quality and safety, illustrate the data continuum and a comprehensive HIT model for LMIC countries.

Increasing attention to health system strengthening and resiliency, data collection, aggregation and data use and the constraints of focusing on monitoring and evaluation data in LMIC reinforces the need for deploying and integrating affordable HIT systems that can support big data and analytics/business intelligence. The Jordanian example is a case study in a country’s commitment to improving the health status of its citizens through the use of HIT.

Discussion Questions

1. What are the top three benefits and challenges of open source HIT systems for global health IT?
2. What differences exist between the US based HIT and ABI maturity models and HIT and ABI maturity models in LMIC?
3. How can AMIA leverage the lessons learned from Jordan to benefit the US HIT and health informatics community?
4. What can AMIA do to assist the global health informatics community?

Panel Organizer Statement: All participants have agreed to take part in the panel and discuss the topics as outlined above.

References

Opportunities and Challenges for Development, Implementation, and Investigation of Acute Care Patient Portals (ACPP): Recommendations from the ACPP Task Force

Anuj K Dalal, MD,1 Sarah Collins, RN, PhD,1 Victoria L Tiase,2 MSN, RN-BC, Kristin O’Reilly, RN, MPH,4 S. Ryan Greysen, MD, MHS, MA5

1Brigham and Women’s Hospital, Boston, MA, 2New York-Presbyterian Hospital, New York, NY, 4Beth Israel Deaconess Medical Center, Boston, MA, 5University of Pennsylvania Health System, Philadelphia, PA

Abstract

The adoption of patient portals has been predominantly in ambulatory settings—use for acute and post-acute care is a rapidly emerging field. An Acute Care Patient Portal (ACPP) task force consisting of stakeholders from early adopter institutions has identified opportunities and challenges for use of patient portals to improve the acute and post-acute care patient experience and is creating a road map to advance this evolving field. This interactive panel will discuss key findings and recommendations from the ACPP 2020 conference that engaged a national group of stakeholders. Session participants will discuss and vet recommendations, refine a road map for development, implementation, and investigation, and identify future directions for innovation.

General Description

The use of patient-facing technologies to engage patients and care partners is viewed as a strategy to achieve better outcomes, improve satisfaction, and optimize healthcare delivery.1 Healthcare systems have been adopting patient portals at an accelerated rate, driven in part by the Meaningful Use (MU) program. Forthcoming legislation—the Caregiver Advise Record and Enable (CARE) Act—will require hospitals to engage care partners more formally in the acute and post-acute care setting. However, the vast majority of use has been limited to the ambulatory setting—use in the hospital and care transitions settings is nascent at best.2 Nonetheless, hospitalized patients represent a vulnerable population with unmet information and communication needs—and there remains the task of promoting self-management after discharge. Although preliminary evidence is encouraging, there is little consensus regarding the development, implementation, and investigation of acute care patient portals.2,3

As part of the Libretto Consortium supported by the Gordon and Betty Moore Foundation, we created an Acute Care Patient Portal (ACPP) task force consisting of early adopter institutions. Early work identified stakeholder perspectives on current practices for acute care patient portals.4 These findings have generated broad interest and culminated in a national conference in June 2016 in Palo Alto, CA: Acute Care Patient Portals 2020: Opportunities and Challenges for Development, Implementation, and Innovation. Participants included CMIOs, CNIOs, clinicians, healthcare executives, policy officials, representatives from electronic health record vendors, patient advocates, and others. Meeting transcript and notes were synthesized, summarized, and reviewed by task force members. Using a group consensus approach, we identified three main themes from our panel of patient advocates—cognitive support, respect and boundaries, and patient and family empowerment—that should help drive future implementation and innovation efforts in a patient-centric manner. From the perspective of the multiple stakeholders engaged during the conference, the ACPP task force identified a set of priorities for development, implementation, and investigation.

This interactive panel will engage a wide variety of participants (clinicians, researchers, CMIOs, CNIOs, etc.) in a range of topics based on the collective work of the ACPP task force. Panelists will discuss recommendations from the ACPP 2020 national conference and share lessons learned from individual institutions. Participants will be encouraged to share experiences from their institutions to corroborate, refute, and refine recommendations, and identify new directions for innovation and investigation.

Addressing the discussion topics below, panel members will present a recommendation in context of their perspective, institutional experience, and involvement in the ACPP task force, and engage audience members in a
robust discussion to elicit participants’ opinions, experiences, and practices.

**Discussion Topics**

- What are sustainable approaches to engage patients and care partners in the acute and post-acute care setting?
- What strategies can be employed to sustain use of the patient portals across the continuum of care?
- How do we leverage patient portals to improve the patient experience and maximize value?
- What outcomes should be used to measure the impact of patient portals over the acute episode of care?
- What are the interoperability and standards necessary to facilitate adoption?
- How should institutions integrate patient portals across the enterprise? Is integration with the ambulatory patient portal the best approach for all institutions?
- What is the role of patient portals in promoting self-management and patient-provider communication during hospitalization and after discharge?
- How do we ensure equitable access securely? Is bring your own device (BYOD) practical, secure, and feasible?
- What are the advantages and disadvantages of hospital-issued versus a BYOD strategy?
- Should we leverage patients’ social networks and existing social media profiles to share with providers?
- What policies should be in place to drive development, innovation, adoption? How do we deal with “portal proliferation” from a policy perspective?
- What policies should be in place regarding sharing of data (test/study results, mental health data) with patients and care partners? What is the role of patient-generated health data?
- How will a new emphasis on APIs (application programming interfaces)–driven by Meaningful Use–impact patient engagement, and how should portals support or supplement this trend? How do we leverage the growth of wearable technologies?

**Moderator:** Anuj K Dalal, MD is an Associate Physician within the Hospital Medicine Unit, Division of General Medicine at Brigham and Women’s Hospital (BWH). He is an Assistant Professor at Harvard Medical School and Fellow of Hospital Medicine. He completed the Program in Clinical Effectiveness at Harvard School of Public Health, and holds a Graduate Degree in Medical Informatics from Oregon Health & Science University. He has expertise in designing, implementing, and evaluating innovative HIT strategies to improve the delivery of care in the inpatient and transitions setting. He has worked on multiple initiatives, including innovative projects to improve communication of clinically significant test results, optimize care team communication, and facilitate patient engagement with an overall goal of improving patient safety, quality, and healthcare value. He co-led the BWH PROSPECT (Promoting Respect and Ongoing Safety through Patient-centeredness, Engagement, Communication, and Technology) effort to design, develop, implement, and evaluate a novel acute care patient portal to engage patients and care partners. He has co-led the ACPP task force to disseminate findings from the Libretto Consortium and the ACPP2020 national conference. **He will provide an overview of the emerging field, review key opportunities and challenges to improve the acute and post-acute patient experience, highlight key recommendations from the ACPP2020 workshop, and introduce panel members.**

**Panel Members**

**Sarah Collins, RN, PhD** is a Senior Clinical and Nurse Informatician at Partners eCare, a researcher within the Division of Internal Medicine at BWH, and an Instructor in Medicine at Harvard Medical School. Dr. Collins received her Bachelor of Science in Nursing from the University of Pennsylvania and holds a PhD in Nursing Informatics from Columbia University. She is an experienced critical care nurse. She was a National Library of Medicine Post-Doctoral Research Fellow at Columbia University’s Department of Biomedical Informatics. Her research and applied clinical informatics work focuses on modeling, developing, and evaluating standards-based, collaborative informatics tools to further patient safety, knowledge development, clinical decision-support, and coordinated patient-centered care. Dr. Collins was first author of a 2011 JAMIA publication titled: Policies for Patient Access to Clinical Data via PHRs: Current State and Recommendations. Currently, she is leading the MySafeCare research project at BWH. MySafeCare is an application for patients and care partners to electronically submit safety concerns while in the hospital for real-time review via clinical dashboards. This AHRQ funded work includes integration of MySafeCare with the acute care patient portal at BWH. She has co-led the ACPP Task Force to disseminate findings from the Libretto Consortium and the ACPP2020 national conference. **She will discuss potential new directions for informatics research based on recommendations from the workshop.**
Victoria L Tiase, MSN, RN-BC is Director, Informatics Strategy at New York-Presbyterian Hospital. She received a BS in Nursing from the University of Virginia and the MS in Nursing Informatics from Columbia University. She serves on the AMIA Nursing Informatics Working Group Board and recently complete a fellowship in the Alliance for Nursing Informatics Emerging Leaders Program. She is currently responsible for planning, organizing, and implementing a range of clinical information technology projects related to electronic health record adoption and meaningful use, clinical workflows, patient engagement, and care coordination. She will discuss institutional strategies to sustain adoption and use of patient portals in acute care settings, including EHR integration.

Kristin O’Reilly, RN, MPH, is currently Manager of Network Medicine at athenahealth. Previously she held the position of Manager of Critical Care Quality at the Beth Israel Deaconess Medical Center (BIDMC). She holds a Master of Public Health in Health Services, Management, and Policy from Tufts University. Prior to working in healthcare quality, she served as a front-line clinical nurse at the Massachusetts General Hospital for five years. This unique combination of formal training and real-world healthcare experience has made her exceptionally effective as a project leader in Critical Care Quality and Safety. At BIDMC she oversaw a large portfolio of grant funded initiatives intended to reduce preventable harm in critical care incorporating informatics, patient engagement and systems science. This portfolio includes the development, design and rollout of an acute care inpatient portal for use in the intensive care unit. She will address recommendations for future investigation to address knowledge gaps, challenges, and emerging opportunities drawing upon her experience designing and implementing an acute care patient portal.

S. Ryan Greysen, MD, MHS, MA is a practicing hospitalist, NIH-funded investigator, and Chief of Hospital Medicine at the University of Pennsylvania. His research focus is on improving engagement of care for hospitalized older adults and interventions to improve post-discharge continuity of care including novel uses of patient-centered technology such as tablets and wearable mobility sensors. Dr. Greysen’s mixed-methods evaluation studies have leveraged advanced quantitative and qualitative techniques to inform health policy at the intersection of mobile health, hospital medicine, and geriatrics. His ongoing work expands this bedside coaching with portals and integrates mobility sensors to engage older adults/caregivers in goal-setting to avoid complications such as functional decline during hospitalization. He will discuss opportunities and challenges for effectively engaging vulnerable populations (such as older adults) in use of acute care portals and related applications (such as mobility sensors connected to the EHR via APIs) to improve patient experience and outcomes.

Acknowledgements: The work of the Acute Care Patient Portal Task Force was supported by the Gordon and Betty Moore Foundation. Additionally, we thank the members of the planning and steering committee for their participation and contributions to this effort.

References:
From Large-Scale Network Analytics to Clinical Solutions in OHDSI

Jon D. Duke, MD MS¹, George Hripcsak, MD MS², Patrick Ryan PhD³,³尼gam Shah MBBS, PhD⁴
¹Georgia Institute of Technology, Atlanta, GA; ²Columbia University, New York, NY; ³Janssen Research and Development, Titusville, NJ; ⁴Stanford University, Palo Alto, CA

Abstract

The Observational Health Data Sciences and Informatics collaborative (OHDSI, pronounced ‘Odyssey’) aims to create reliable scientific evidence through large-scale analysis of observational health data from around the world. OHDSI promotes the translation of this work to improve clinical research and medical decision-making. In the proposed panel, members of the community will discuss how OHDSI is working to support clinical research efficiency as well as support evidence sharing for medical decision-making by adopting the FHIR integration. To ensure transparency and reproducibility in our research, we have created shared knowledge artifacts such as phenotype libraries. In the proposed panel, we will discuss our approach to building shareable knowledge assets, conducting research using these assets, and supporting dissemination beyond the OHDSI community.

Organizer: Jon D. Duke, MD, MS

Background and Panel Description

The Observational Health Data Sciences and Informatics collaborative (OHDSI, pronounced ‘Odyssey’) was formed in 2013 with the goal of creating reliable scientific evidence through large-scale analysis of observational health data from around the world.¹ To advance this goal, OHDSI has had to rapidly expand its scientific, technical, and community infrastructure. OHDSI has now grown to over 200 participants from 17 countries. Across the collaborative, there are 55 databases covering over 650 million patient lives that have been transformed into a common data model. OHDSI has conducted multiple international network-based observational research studies using this infrastructure. In this panel we will focus on efficiency of evidence generation, and on translating large scale analytics into clinical solutions. We will report back on the results of a two day face to face meeting and hackathon from March 2017 as well as describe the development of the OHDSI on FHIR framework designed to facilitate the integration of OHDSI predictive models into clinical decision support. By the end of this panel, attendees will gain an understanding of the issues and approaches to sharing of large-scale observational research with clinicians.

Panel Importance and Target Audience

The target audience includes health informaticians, clinical and health services researchers, health IT leaders (e.g., CMIOs), implementation scientists, epidemiologists, biostatisticians, and data scientists. This panel will provide attendees with insight into the opportunities and challenges of conducting reproducible research across diverse datasets as well as a clear path for sharing research results outside the OHDSI network. The proposed panel illustrates how international collaborative research can be conducted in a transparent and reproducible fashion; and in the process generate large scale shared resources.²

Presentations

An Informatics Framework for Systematic Clinical Research across Observational Data Network (P. Ryan)
Observational healthcare data from electronic health records and administrative claims can be used to explore and test clinical hypotheses and generate aggregate evidence that can meaningfully inform medical decision-making.
While the community has seen tremendous advances in technical infrastructure, data standards, and statistical computing over the past decade, the paradigm for conducting clinical research has largely remained stagnant: a researcher poses a question about a particular clinical phenomenon (such as an exposure-outcome relationship), seeks to find a dataset that can be used to test the hypothesis, generates results that confirm or refute the hypothesis, drafts a publication that interprets the findings, rinses and repeats. Each clinical research project is a one-time journey, often inefficient but rarely reproducible. Other forms of replication - applying similar methods to different data, or different methods to the same data - to assess the robustness of clinical findings are difficult to coordinate and execute. The OHDSI community has been actively developing and testing informatics frameworks to facilitate the clinical research process - replacing the bespoke approach with a systematic pipeline to consistently guide researchers from research conception through protocol development through analysis implementation through network execution through results coordination. We have applied this process across questions of clinical characterization, population-level estimation, and patient-level prediction with the aim to the efficiency, accuracy, and reproducibility of observational analyses. In this talk, we will share the experiences and lessons learned from executing clinical research questions across the OHDSI network using this framework.

**Large-Scale Dissemination of OHDSI Prediction Models via FHIR Integration (J. Duke)**

One of the strengths of the OHDSI analytics environment is its predictive modeling engine known as PatientLevelPrediction (PLP, www.github.com/OHDSI/PatientLevelPrediction). PLP is an R-based package for generating predictive models using a large set of covariates such as all drugs, diagnoses, procedures, demographics, and comorbidity indices. PLP incorporates a range of machine learning techniques for feature selection and model generation. Current outcome models include logistic regression, random forest, gradient boosting machine (GBM), naive bayes, k-nearest neighbors (KNN), and multi-layer perceptron (MLP). While generating high-performing models is a key step in advancing the value of observational data, model generation alone is insufficient. Indeed, these models must be applied to individual patients and delivered at the right time and place (e.g., point of care) to improve medical decision-making. Dr. Duke will discuss the development of the OHDSI on FHIR framework designed to facilitate the integration of OHDSI predictive models into clinical decision support. This framework wraps PLP-based predictive models into RESTful APIs that ingest standard FHIR resources (e.g., medications, conditions), transiently converts the data to OMOP CDM format, and generates an individual patient prediction. This framework allows any FHIR-enabled EHR to incorporate OHDSI-based predictive models without the need to maintain an OMOP database. We will discuss how the OHDSI on FHIR framework can be used to expand evidence dissemination in clinical and research settings.

**Large-Scale Electronic Phenotyping (N. Shah)**

One of the first steps in using EHR data for research is to reliably identify a cohort of patients that have a condition of interest or a phenotype. Typically, methods for identifying patients with a given phenotype have relied on rule-based definitions. Recently, statistical learning approaches have also been employed for electronic phenotyping. Here the rate-limiting step is manual creation of training sets for statistical learning approaches. Given the heterogeneity of the data models in use, missing data values, and differences in standardization, in commercial EHR systems phenotype definitions are difficult to port across different EHR systems and institutions. Developing such definitions against common data models is one way to create reusable phenotype definitions. In the OHDSI community, we have demonstrated that by using semi-automatically assigned, and possibly noisy labels in training data, it is possible to build phenotype models that are comparable to rule-based phenotype definitions. The key intuition is that the large volume of training data which can be collected using an automated labeling process, can compensate for the inaccuracy in the labels. This presentation will discuss our effort to scale electronic phenotyping methods to build a library of 100 phenotype definitions for the OHDSI community for shared use. We will describe our experience in creating such phenotype definitions at scale. We will discuss automated mechanisms to evaluate the validity of phenotype definitions; and review the progress made at the OHDSI face to face meeting and hackathon held March 17-18 2017.

**Large-Scale Incidence Rates (G. Hripcsak)**

The large-scale calculation of incidence rates serves as a first step toward a comprehensive biomedical knowledge base. For every given drug, condition, procedure, etc., we can calculate the incidence of all other drugs, conditions, procedures, etc. Such a knowledge base is easily understandable, requires no assumptions, is efficient to calculate,
and is straightforward to visualize. The challenge is that incidence does not assign causality and is subject to many forms of confounding. We can see the probability of something happening but not the probability of improving it. For example, a condition that begins frequently when a drug is taken may be a side effect of the drug or may be part of the indication for the drug. Stopping the drug may or may not remove the condition. Nevertheless, the information is useful. Incidence rates can serve as upper bounds to side effect rates, and they can be interpreted in the context of deeper medical knowledge, such as whether a condition is likely to be part of a drug’s indication. As we move to large-scale causal assessments, incidence rates will remain an important foundation to interpret the causes. For example, something that is highly causal but very low incidence may not be of concern. OHDSI is calculating all-by-all incidence in a set of large databases and making it available to the public.

Discussion Questions

- What are the main challenges associated with disseminating evidence?
- What is the value of reusable knowledge assets such as phenotype libraries?
- How can phenotype performance be consistently evaluated across diverse institutions?
- What are the barriers to disseminating predictive models?
- How can FHIR be used to connect evidence to providers in diverse clinical settings?
- How can large-scale incidence rate enable better care or better research?
- What are the limitations of large-scale incidence rates.
- What motivates large collaborative studies?

Assurance
The organizer Jon D. Duke vouches that all listed participants have agreed to take part on the panel.

References

An Overview of Emerging Real-World OpenHIE Use Cases: Successes, Challenges, and Future Opportunities

Shaun Grannis, MD, MS, FACMI1, Eric-Jan Manders, PhD2, Carl Leitner, PhD3, Annah Ngaruuro, MS4, Jack Bowie, ScD5

1Indiana University and the Regenstrief Institute, Indianapolis, IN; 2Centers for Disease Control and Prevention, Atlanta, GA, USA; 3PATH, Chapel Hill, NC, USA; 4ICF International Inc., Rockville, MD, USA; 5Apelon Inc., Hartford, CT, USA.

Abstract

A growing international collaborative of diverse organizations and people have formed a community to develop open processes and technologies to support essential clinical data sharing needs of low and middle income countries throughout the world. OpenHIE (Health Information Exchange) is a global initiative convening communities of practice, implementers, and reference software components that together support national-level health information sharing activities. In this panel we will describe multiple real-world use cases supported by OpenHIE’s operational infrastructure, and will discuss the evolving community process, opportunities for broader collaboration, and future development plans. Learning objectives for the panel include: understanding the basic elements of a component-based eHealth architecture for low and middle-income countries, as exemplified by OpenHIE; describing specific use cases supported by OpenHIE; and identifying potential barriers and solutions for implementing specific health information exchange use cases for low and middle-income countries.

Introduction

Healthcare systems and public health practice rely on information and communication methodologies and tools for enabling the optimal use of data and information for action. Recent pandemics have further emphasized the need for information derived from flexible, standards-based, and interoperable electronic health information systems. Experience has shown that application of information and communication technologies (ICTs) for health, i.e., eHealth, especially in a global, interconnected world, requires strong leadership at the strategic and implementation levels to develop a sustainable model. To address this need, OpenHIE has emerged over the last two years, and numerous countries in Africa and Asia have engaged in the initiative. OpenHIE is a community of individuals and organizations who are committed to serving the underserved by bringing responsible technical architectural solutions to support large-scale electronic health information sharing. The OpenHIE community welcomes all and is represented by government agencies, private sector companies, academia, and non-governmental organizations contributing to the development of an open source, globally relevant health information exchange (HIE). Information sharing has become increasingly central to improving healthcare systems and health outcomes, The OpenHIE community seeks to share our experience and lessons learned with those interested in data sharing regardless of setting. This session will leverage the perspectives shared by attendees to inform future trajectory for health information architectures and data sharing use cases.

The 66th World Health Assembly (WHA) proposed a progressive new agenda in support of the health information architectures for it’s member states. WHA resolutions implored global health and informatics experts to support member states as they developed their national eHealth strategies through a multistakeholder and multisectoral approach including national authorities, relevant ministries, relevant private sector parties, and academic institutions. In parallel, a broad consortium of donors, stakeholders and development partners developed the nine Principles for Digital Development to serve as a set of living guidelines meant to inform the design of technology-enabled development programs. Both OpenHIE and this proposed panel respond to these calls to action, which furthers the dialogue around the effective, real-world implementation of health information architectures via community of practice models.

OpenHIE Introduction (Grannis)

OpenHIE’s mission is to improve the health of the underserved through open, collaborative development and support of country driven, large scale health information sharing architectures. The initiative is multifaceted and includes a flexible component-based architecture to support a diverse set of health information exchange clinical workflows. As
chief architect of OpenHIE, Dr. Grannis will serve as moderator and will provide an overview of the OpenHIE architectural elements, the reference software components and document artifacts, and the inclusive community model. The panelists will highlight use cases that illustrate the broad use of the OpenHIE construct to improve care and outcomes.

**Integrating Standard Terminologies: OpenHIE Terminology Services (Bowie)**

As developing countries contemplate comprehensive eHealth strengthening strategies, a primary function of a fully functional health information exchange is to aggregate, normalize, and communicate diverse healthcare information among disparate sources for various use cases. As data from different systems is exchanged and used outside its originating system, it is challenging to maintain the original meaning of the data. The use of standard definitions, terminologies and ontologies provides a technical language that can enable the data to have meaning outside its originating system.

This presentation will describe the role of one of the critical components of OpenHIE: the Terminology Services (TS). The TS component of the OpenHIE architecture provides a centralized source for the HIE’s standards and definitions, including terminologies, ontologies, dictionaries, code systems, and value sets. We will discuss the value of standardized data representations and the difficulties inherent in their adoption, particularly in the context of low and middle-income countries. We will describe examples of TS integration with other OpenHIE architecture components. The presentation will conclude with a discussion of TS integration in a variety of current OpenHIE implementations, touching on issues of required clinical and demographic domains, selection of national and international standards, and ongoing governance of the overall standards environment.

**The DATIM experience supporting indicator reporting (Ngaruro)**

The President's Emergency Plan For AIDS Relief (PEPFAR) is a U.S. governmental initiative to address the global HIV/AIDS epidemic, strengthen health systems, and help save the lives of those suffering from HIV. Its reach is global and includes partnerships with many countries. To manage information generated in participating countries, PEPFAR’s Data for Accountability, Transparency and Impact Monitoring (DATIM) system seeks to build a health information exchange infrastructure to enable data exchange among PEPFAR in-country teams, implementing partners and Ministries of Health. To accomplish this, the DATIM initiative has begun implementing various components of the OpenHIE architecture. We will discuss the DATIM use case, the experience and challenges, and value of implementation of the OpenHIE architecture from a programmatic and system perspective as well as the value proposition for stakeholders. Further, we will describe the minimum set of technologies necessary to accomplish the DATIM use case as well as how a discussion around the use of other HIE component to enrich and enhance the data exchange infrastructure.

**mHero: Supporting health workers for routine care and Ebola crisis care (Leitner)**

We will describe how the OpenHIE framework has been used to support sustainable health information systems by exploring a two-way SMS communication platform called mHero. mHero has been deployed in West Africa as part of the USAID Ebola Grand Challenge to link existing routine health worker registry and facility information systems (iHRIS and DHIS2) to provide integrated scalable tools to reach health workers in support of both routine health services crisis response situations. We will discuss the enabling environment needed to move a digital health intervention from a pilot to a national scaling implementation.

We will describe how factors including a participatory design processes, the installation of comprehensive governance structures, the inclusion of mHero in written eHealth policy, and investments in Ministry human resources and ICT capacity have produced a strong sense of local ownership of mHero. Participants will learn how an approach using OpenHIE simplifies the deployment of complex integrated health information systems and the technological tools that can be managed by ministries of health even in severely resource-constrained settings.

**HIV case-based surveillance (Manders)**

In many countries, the national response to the global HIV epidemic has contributed to the development of health information systems. With an initial focus on reporting program performance metrics, stakeholders recognize that system architecture and design considerations are critical to sustainability and successful implementation at scale. One such consideration is health information exchange, and the preference to achieve it through standards-based interoperability.
For the context of a country that has a substantial and growing footprint of facility-based EMR systems, we focus on leveraging the OpenHIE framework to implement EMR-based HIV case surveillance. We use the OpenHIE architecture to create a reporting channel to transmit selected clinical counter data of public health interest to a centralized data repository. Data transmission is based on clinical trigger events, and uses a standards-based message format. An OpenHIE client registry component can be used to support record linkage for patients who seek health services from different facilities, resulting in more accurate client counts for reporting. Security controls support the implementation of data access policies. Our aim is to leverage OpenHIE to facilitate program planning and public health decision-making from patient level data without the requirement to establish a fully developed shared health record for care coordination. The reusable architecture can accommodate interoperable systems for laboratory reporting, vital registration, and data sources that originate from paper registers for more complete data analysis.

**Discussion Questions**

1. What are the key elements of a component-based eHealth architecture supporting low and middle-income countries?
2. How can elements in component-based eHealth architecture be orchestrated to support specific data exchange workflows?
3. What are potential barriers to implementing specific data exchange use cases?
4. What key non-technical factors can enable successful implementation of data exchange workflows?
5. How can terminology services integrate into an overall eHealth architecture?
6. What lessons learned from implementing Health Information Exchange in the US could translate to LMIC settings?

**Panel Organizer Statement:** All participants have agreed to take part in the panel and discuss the topics as outlined above.

**Disclaimer:** The findings and statements in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention.

**References**

Clinical Decision Support in the Era of Precision Medicine

Panelists
Barry Blumenfeld, MD, MS1, Joni L. Rutter, PhD2,
Eliot Siegel, MD3, Linda Wedemeyer, MD, RN, MS4

Moderator
Michael A. Grasso MD, PhD3

1RTI International, Raleigh, NC, 2National Institutes of Health, Bethesda, MD, 3University of Maryland, Baltimore, MD, 4Veterans Health Administration, Los Angeles, CA

Abstract
The All of Us Research Program from the Precision Medicine Initiative will create new opportunities and challenges for clinical decision support (CDS). An important goal of the program is to create a research cohort of at least one million U.S. volunteers, which will be used to develop customized approaches to treat diseases and improve health. This panel will discuss new opportunities and challenges for CDS using the large research cohort. This panel, which was organized by the AMIA Clinical Decision Support Working Group, brings together leaders from government, academia, and industry with diverse backgrounds in Biomedical Informatics.

Background
During his 2015 State of the Union address, President Obama announced a new Precision Medicine Initiative (PMI). Renamed the All of Us Research Program, its goal is to develop new approaches to treat diseases and improve health customized by an individual’s genetic, environmental and lifestyle characteristics1,2,3. The program received initial funding of $215 million, which was allocated to NIH to build a large national cohort and to support cancer genomics. The Precision Medicine Initiative, PMI, All of Us, the All of Us logo, and The Future of Health Begins With You are service marks of the U.S. Department of Health and Human Services.

One long-term goal of the All of Us Research Program is to create a research cohort of at least one million U.S. volunteers4, who will contribute genetic data, lifestyle and environmental information, biological samples, and links to their electronic health records (EHRs). This new All of Us cohort will be unique in its balance of participants, reflecting the diversity of the U.S. population based on age, gender, ethnicity, geography, social characteristics, and health status. The program is a partnership between the cohort participants, health care providers, and the research community, to work together to develop prevention and treatment strategies with greater precision.

Similar large cohorts have been created, for example, IMS Health nation cohort of three million primary care patients in United Kingdom’s National Health Service5. In addition, the U.S. Veterans Health Administration manages research data on roughly thirty million veterans through the VA Informatics and Computing Infrastructure (VINCI)6, and is currently developing the Million Veteran Program to study genetic influences on health and disease7.

Introduction
This panel will discuss new opportunities and challenges for CDS in light of the All of Us Research Program. A wealth of new knowledge is expected to flow from the initiative. However, its effective use in clinical care depends heavily on its incorporation into CDS applications and tools. The barriers to such use are many. The provision of CDS will hinge on the development and deployment of new knowledge environments based upon flexible data models and extensive use of standard-based ontologies and terminologies8. Access will depend on wide scale adoption of standardized application programming interfaces. Current rule-based systems used in EHRs for most CDS systems might not be adaptable to the types of data and conditions represented by genomic data, nor is it known if they can scale for the very large numbers of data points represented by mutations and deletions in large genome datasets9. Other challenges include a lack of data independence, as well as complexity, quality, and integration10. Additionally, at the
point of care, such knowledge must be displayed in user-friendly formats so that physicians, laboratory personnel, nurses, researchers, and other end-users can enter data, access information, and understand the output.

**Topics for Discussion**

Clinicians are already dealing with a knowledge explosion that exceeds their capacity for absorption. The addition of genomic knowledge will only strain their capacity further. CDS has the potential to bridge the gap between the promise and realization of precision medicine, but it depends on progress in multiple arenas. The goal of this panel is to discuss these challenges and barriers, and to relate possible solutions. Panelists will address the following topics:

- Research goals and the timeline of the All of Us Research Program.
- The role of CDS and participants as partners in achieving the All of Us Research Program goals.
- Enabling CDS in the All of Us cohort: Knowledge engineering, implementation, and measurement.
- Using CDS in the All of Us cohort to drive a learning system.
- Challenges and barriers to the effective use of CDS within large cohorts.
- Real-world CDS experiences and lessons learned from other efforts.

**Panel Organizers**

This panel is sponsored by AMIA Clinical Decision Support Working Group.

- Yang Gong, MD, PhD, Chair.
- Barry Blumenfeld, MD, MS, Vice Chair.
- Michael A. Grasso, MD, PhD, Immediate Past Chair.
- Eric Pan, MD, MSc, Past Chair.
- Laura S. Crawford, MIS, Member-At-Large.

**Intended Audience**

The intended audience includes clinicians, researchers, and other informatics professionals who have an interest in healthy living, health policy, disease management, clinical decision support, and electronic health record integration.

**Speaker Contributions**

**Barry Blumenfeld, MD, MS - Senior Physician Informaticist, RTI International**

Barry Blumenfeld is a Senior Physician Informaticist at RTI International. He currently serves as the principal investigator for the AHRQ sponsored Patient Centered Outcomes Research Clinical Decision Support Learning Network. He is an authority on CDS and clinical documentation, and has led the design of systems in both academic and commercial settings. Other areas of expertise include health care analytics, health care information standards, interoperability, and knowledge management. He is also the current Vice Chair of the AMIA Clinical Decision Support Working Group. Dr. Blumenfeld will discuss the value of knowledge sharing and CDS in clinical practice, and the role of the All of Us cohort in knowledge engineering and management.

**Joni L. Rutter, PhD - Director of Scientific Programs, All of Us Research Program, The Precision Medicine Initiative, NIH**

Joni Rutter is the Director of Scientific Programs within the All of Us Research Program of the Precision Medicine Initiative at the National Institutes of Health. She leads the scientific, programmatic development, and implementation efforts to build a national research cohort of one million or more U.S. participants to advance precision medicine. She is internationally recognized for her work in basic and clinical research in human genetics and in the study of genetic and environmental risk factors focusing on the fields of cancer and addiction. She received her Ph.D. from Dartmouth Medical School and completed a fellowship at the National Cancer Institute. Her primary scientific objective is to integrate genetic principles with environmental influences to increase our understanding of how individual and societal influences impact health and disease. Dr. Rutter will discuss the All of Us cohort efforts and its application to CDS.

**Eliot Siegel, MD - Professor, University of Maryland School of Medicine**

Eliot Siegel is a Professor and Vice Chairman in the Department of Diagnostic Radiology and Nuclear Medicine, and Chief of Imaging Services for the VA Maryland Healthcare System. He has been recognized internationally for his
work in imaging informatics, served as the National Cancer Institute Cancer Biomedical Informatics Grid (caBIG) Imaging Workspace lead, and is a designated informatics lead within the national VA network. He is a fellow of the American College of Radiology and of the Society of Imaging Informatics in Medicine. His areas of interest and responsibility at both the local and national levels include digital imaging and PACS, telemedicine, the electronic medical record, informatics and artificial intelligence systems. Dr. Siegel will discuss related CDS efforts with large cohorts, including the IBM Watson program and caBIG.

Linda Wedemeyer, MD, RN, MS - Physician Informatician, VA Greater Los Angeles Healthcare System

Linda Wedemeyer is a Physician Informatician within the Veterans Health Administration Office of Informatics and Information Governance. She is part of the Clinical Decision Support team in Knowledge Based Systems. Her work includes CDS and health IT evaluation at the VA, and health IT standards development for the American Academy of Ophthalmology. She is working on a model for CDS evaluation and on standardized, shareable CDS knowledge artifacts. She has also been responsible for writing metrics to determine the impact of interoperability and health IT modernization on healthcare processes and outcomes. Dr. Wedemeyer will discuss real-world CDS experiences within the VA system, and the availability of large clinical repositories within the VA system.

Michael A. Grasso, MD, PhD – Assistant Professor, University of Maryland School of Medicine

Michael Grasso is an assistant professor of Emergency Medicine, Internal Medicine, and Computer Science at the University of Maryland School of Medicine. He practices Emergency Medicine at the Baltimore VA Medical Center, is board certified in Clinical Informatics, and is Director of the University of Maryland Clinical Informatics Group. He is currently working with the national clinical repository from the Veterans Health Administration. He is developing new approaches to knowledge representation and reasoning, which can be optimized for very large clinical repositories. The clinical focus for this work includes several chronic diseases, prescription opioid abuse, and quality improvement in Emergency Medicine. Dr. Grasso will be the panel moderator.

Participation Statement
All of the panelists have agreed to participate if the panel is selected for the AMIA 2017 Fall Symposium.

References
Successes and Challenges in Developing and Implementing Electronic Informed Consent Tools for Research

Panelists
Christopher A. Harle, PhD1, Hyeoneui Kim, RN, MPH, PhD2, David R. Nelson, MD3

Moderator
Kenneth W. Goodman, PhD, FACMI4

1Indiana University and Regenstrief Institute, Indianapolis, IN; 2University of California, San Diego, CA; 3University of Florida, Gainesville, FL; 4University of Miami, Miami, FL

Abstract
The U.S. health care system’s increased adoption and use of electronic health records (EHRs) over the past decade has increased the utility of health records for research. With these technological changes, new approaches and systems are needed to efficiently and appropriately inform patients about and consent them to participate in research, especially research that uses their electronic health information. In this panel, three researchers from different disciplines will describe three novel efforts to develop, implement, and evaluate point-of-care electronic systems that administer broad research consent. These systems span multiple institutions, are integrated with institutional EHRs, and involve consent for research re-contact and patient sharing of health records and bio-samples for research. At the conclusion of this panel, audience members will be able to: (1) discuss technological, process, and ethical challenges and keys to success in designing electronic consent tools for point-of-care use, (2) describe operational and technical challenges and keys to success in integrating research consent tools with EHRs and other enterprise information systems, and (3) critique qualitative and quantitative approaches to evaluating the effectiveness of electronic informed consent tools.

Introduction
The U.S. health care system’s increased adoption and use of electronic health records (EHRs) over the past decade has increased the utility of electronic health information for clinical research. EHR systems serve as direct and indirect, via integrated warehouses and registries, sources for researchers to analyze phenotypic, treatment, and outcomes data for patients over time, across care settings, and across organizations. EHRs can be used to identify patients for clinical trials and linked to biosample data for genomics research.1, 2 In addition, where racial and ethnic minorities are often underrepresented in clinical research,3, 4 EHRs may contain more representative data. Indeed, it has been suggested that failure to make efficient use of EHRs for research is itself an ethical failure.5

Given the expanded content and uses of EHRs, a key question is whether new approaches are needed to ensure ethical obligations are met when informing and consenting patients about research that uses their health information. However, while ethically appropriate approaches to informing and consenting patients are critical, research institutions also demand efficient approaches for administering broad research consents, and integrating and sharing consent and research data with partner organizations. Therefore, electronic applications for consenting patients at the point of care may provide an ethically appropriate, patient-centered, and efficient approach to involving patients in research. Compared to traditional paper-based consents, well-designed interactive electronic consent tools may be better at allowing patients to explore personally-relevant information and communicate granular consent preferences, and allowing research institutions to integrate consent data with other clinical and research data.

Why this Panel is Timely and Urgently Needed
This panel involves an interdisciplinary group of clinicians and researchers who will be discussing a timely topic with important policy, administrative, and research informatics implications. In recent years, the Department of Health and Human Services (HHS) prioritized improved knowledge and policies related to the increased scale and scope of electronic health information use in research. In 2014 and 2015, this led the National Institutes of Health (NIH) to request grant applications on Ethical Issues Related to Consent for Research Using Clinical Records and Data. This panel will discuss two of the projects funded by this request. Furthermore, in early 2017, HHS finalized updates to the Common Rule,6 the primary regulation that protects research participants. Several of the Common Rule updates aim to overcome the concern that informed consent processes are administratively and legally
burdensome, and not designed to help people make informed decisions about participating in research, including sharing their health information for research. Therefore, this panel will discuss the potential for electronic informed consent tools to integrate with clinical research information systems and support consent processes. We expect the panel will attract clinicians, information technology administrators, and informatics researchers interested in developing and leveraging electronic tools for efficient and patient-centered research.

General Description of the Panel
In this panel, three researchers from different disciplines will describe three novel efforts to develop, implement, and evaluate point-of-care electronic systems that administer research consent. These systems span multiple institutions, are integrated with institutional EHRs, and involve patient consent for research re-contact and sharing of health records and bio-samples for research. The panelists will describe technological, process, human, and ethical keys to successes and challenges in designing, implementing, and rigorously evaluating electronic consent tools.

Panelists
Dr. Harle is an Associate Professor of Health Policy and Management at Indiana University and Affiliated Scientist at the Regenstrief Institute’s Center for Biomedical Informatics. He is also currently a Co-Principal Investigator on the NIH-funded study “An Interactive Patient-Centered Consent for Research Using Medical Records.” In this panel, Dr. Harle will describe findings from this study, which aims to design and evaluate an interactive electronic informed consent application that asks patients for consent to use their identifiable health record data in research. The application is designed to be efficient for use in clinical practice settings while also enhancing patient trust and flexibly meeting their information needs, as compared to traditional paper-based consents. First, Dr. Harle will describe patient preferences and information needs related to electronic research consent. These were identified during a series of 42 think-aloud interviews conducted during a mock electronic consenting process. Second, Dr. Harle will describe the successes, challenges, and results of a randomized experiment examining how the interactive and trust-enhancing features of an electronic consent application affect peoples’ satisfaction with an electronic consent process, understanding of the consent, perceptions of coerciveness, and trust in medical researchers.

Dr. Kim is an Associate Professor of Biomedical Informatics at the University of California, San Diego. She is also currently a Co-Principal Investigator on the NIH-funded study “iCONCUR: informed CONsent for Clinical data and Biosample Use for Research.” In this panel, Dr. Kim will describe findings from this study, which aims to guide the development and implementation of informed consent management systems that elicit patient preferences for clinical data sharing based on what data are shared and who the recipients of shared data are. First, Dr. Kim will describe the successes and challenges involved in developing, pilot testing, and establishing feasibility of a web-based informed consent tool that allows patients to express granular preferences for the use of their EHR data in research. Second, Dr. Kim will describe the second phase of iCONCUR, where consenting options are refined and survey functions are expanded to incorporate the care dependents of the main survey respondent. The current status of the iCONCUR tool will be demonstrated and preliminary results of the phase two survey will be shared.

Dr. Nelson is a Professor of Medicine, Assistant Vice President for Research, and Director of the NIH-funded Clinical and Translational Science Institute (CTSI) at the University of Florida (UF). In this panel, Dr. Nelson will describe the UF’s Consent2Share program. The Consent2share program encompasses processes and electronic consent systems that ask patients at the point-of-care for their consent to be re-contacted about research studies in which they may be eligible to participate. The Consent2Share program was pilot tested in an internal medicine and medical specialties practice and subsequently rolled out at an enterprise level for both pediatric and adult patients. First, Dr. Nelson will describe successes and challenges in Consent2Share’s rollout and integration with UF’s Epic EHR system, integrated data repository, and i2b2 system. Second, Dr. Nelson will describe Consent2Share’s value in supporting clinical trial recruitment, including relative yield of different recruitment strategies.

Dr. Goodman is a Professor of Medicine, Director of the University of Miami School of Medicine’s Institute for Bioethics and Health Policy, and co-founder of the North American Center for the Study of Ethics and Health Information Technology. He is also past chair of the Ethics Committee of AMIA and co-founder of the Ethical, Legal and Social Issues Working Group. Dr. Goodman will moderate this panel, discuss overlapping themes in the panelist presentations, and lead the audience discussion.

All participants have agreed to take part on the panel.
Learning Objectives
At the conclusion of this panel, audience members will be able to:

1. Discuss technological, process, and ethical challenges and keys to success in designing electronic consent tools for point-of-care use.

2. Describe operational and technical challenges and keys to success in integrating research consent tools with EHRs and other enterprise information systems.

3. Critique qualitative and quantitative approaches to evaluating the effectiveness of electronic informed consent tools.

Discussion Questions
The panelists will encourage in-depth discussion from the audience. Several discussion questions are listed below:

1. What ethical obligations do researchers have to inform and consent patients regarding the use of their health record data for research? How has that obligation evolved with widespread EHR adoption, increasing health information exchange, and new forms of personal health data that are stored in EHRs?

2. For what types of research and patients can well-designed electronic informed consent tools serve as an effective substitute for in-person consent processes? What are the potential legal, ethical, and operational strengths and weaknesses of replacing in-person informed consent with electronic tools?

3. What are the appropriate metrics for assessing the value and impact of electronic informed consent tools on research activity?

4. What challenges have audience members’ institutions faced when considering, developing, implementing, and using electronic informed consent tools?

References


A Structured Approach to Measuring Individual Nurse’s Contribution in Patient Outcomes

Ellen M. Harper, DNP, RN, MBA, FAAN¹, Catherine H. Ivory, PhD, RNC-OB, RN, BC ², Lisa A. Moon, BSN, CCMC, LNC, PhD², Amy Garcia, DNP, RN, CENP ⁴
¹University of Kansas, Kansas City, KS, ²Vanderbilt University, Nashville, TN, ³ University of Minnesota, Minneapolis, MN, ⁴ Cerner Corporation

Abstract

Nurses are key drivers of patient outcomes as well as a significant cost center in any healthcare setting. How do we measure nursing care value – or more importantly, the unique value each nurse contributes to patient outcomes? At the most basic level of definition, value is the relationship between outcomes of care and the costs to produce that care. These two basic elements of cost and outcome are measured using the individual patient as unit of analysis. Traditional hospital outcome metrics are measured in terms of length of stay, the presence or absence of adverse events such as infection, fall, or pressure ulcer, or inpatient mortality. However, the volume of data generated by clinical and operational electronic health records (EHRs), permits the examination of short term changes in patient outcomes attributable to nursing care. EHR data allows measurement of both the patient response to treatments or nursing interventions, at the individual nurse level and correlate both with nurse characteristics such as education and experience levels. This didactic panel will describe the methodological approach taken by the Nursing Value Expert Workgroup to structure, define, abstract and evaluate discrete data from disparate electronic systems to measure nurse’s contribution to patient outcomes.

Introduction

Nurses are key drivers of patient outcomes as well as a significant cost center within any healthcare setting¹². How do we measure nursing care value – or more importantly, the unique value each nurse contributes to patient’s outcomes? At the most basic level of definition, value is the relationship between outcomes of care and the costs to produce that care³⁴. These two basic elements of cost and outcome are measured at the individual patient unit of analysis, for example, what is the total nursing care time and costs to care for a patient with acute pneumonia and what are the patient’s outcomes of the care? Conversely, we do not measure care cost and outcomes at the individual nurse level. Traditional metrics of hospital outcomes are measured in terms of length of stay, potential adverse events such as infection, falls, pressure ulcer or inpatient mortality⁵. However, the volume of data generated by clinical and operational electronic health records (EHRs), permits the examination of short term changes in patient outcomes attributable to nursing care⁶. Discrete electronic data allows measurement of both the patient response to treatments or nursing interventions, at the individual nurse level and correlate both with nurse characteristics such as education and experience levels. This didactic panel will describe the methodological approach taken by the Nursing Value Expert Workgroup to structure, define, abstract and evaluate discrete data from disparate electronic systems to measure nurse’s contribution to patient outcomes.

Nursing care is a practice discipline and occurs between one nurse and a person, family, or community. Nurses can be linked to individual patients within the EHR in several ways (e.g., the nurse-patient assignment or other data that capture touch points such as bar coding for medication administration or recording an assessment in a patient’s home or clinic documentation⁷). These data can provide new and unique opportunities to measure nursing care in several different dimensions (e.g., the 1:1 relationship of a nurse caring for a patient, the impact of many nurses on the outcomes of care for a single patient, or the potential to investigate similar outcomes across many nurses and many patients across many different settings⁸).

This didactic panel will describe the structured methodology and work products of a Nursing Value Expert Workgroup⁹¹⁰, organized through the Nursing Big Data to Knowledge initiative¹¹. The expert workgroup is a component of ongoing efforts of a National Action Plan to ensure that nursing data is captured in EHRs and other information systems – and that the data is available in shareable and comparable formats so that clinicians, nursing administrators, researchers, policy makers and others can use it to gain useful, actionable insights. The aim is to ensure that nursing data is used to inform changes leading to better outcomes, lower cost and an improved patient and staff experience.
Objectives

1. Develop a national consensus model to measure patient-level nursing intensity and costs per patient in multiple care settings to support the continuum of care and to produce objective measures of nursing care value.

2. Develop new nursing business intelligence and analytic tools that will utilize the rich clinical, operational, financial, and quality/safety outcome data currently available to measure and compare nursing care value.

3. Develop and test new nursing financial models to bring transparency to support risk sharing within Accountable Care Organizations, value-based purchasing, and pay-for-performance models.

The panel will include presentations on the following 4 topic areas:

1. Presentation of the recently published Value Based Nursing Data Model which demonstrates the relationships across disparate electronic data and primary key (relationships). The Value Based Nursing Data Model is EHR agnostic and venue setting neutral to support the continuum of care.

2. Processes used to develop a data dictionary that structurally and operationally defines individual nursing value data elements mapped to standardized terminologies, such as SNOMED CT and LOINC. We will describe efforts to use existing definitions from established dictionaries such as the Centers for Medicare/Medicaid Services (CMS), the Nursing Management Minimum Data Set (NMMDS), the National Database of Nursing Quality Indicators (NDNQI) and others, to eliminate redundancies. We also propose a framework to harmonize recognized terminologies that describe nursing interventions, processes and outcomes, that recognizes facility preferences but permits data sharing and benchmarking. Finally, we propose value sets for standards organizations, to approve and use for the measurement of nursing care across inpatient, outpatient, home care and public health domains.

3. Development of User Stories used to deconstruct complex nursing scenarios into a standardized set of features, item responses and data elements. The User Stories use conventional agile information technology processes. The tool intentionally links the identified user story features to the Value Based Nursing Data Model. This approach provides a standardized process that can be used to describe nursing value for data analytics and business intelligence purposes.

4. Presentation of the results of the pilot study site used to test the work products. The research questions- Do the characteristics (certification, education, nursing experience, and experience on the unit or service line) of registered nurses assigned to care for a patient significantly impact patient care and outcomes, including cost? How does the workload of the individual nurse impact the outcomes of the patient? Additionally, presentation on the preparation of a multi-site study which includes the Nursing Value Data Warehouse (NVDW) and business intelligence to facilitate research on the value of nursing across the continuum of care.

The presentation will demonstrate how we can use discrete data to improve efficiency, effectiveness, productivity, performance, and quality, to achieve exceptional outcomes of nursing care.

Discussion questions:
What strategies can be used to define and categorize data elements used to measure nursing value?

What discrete data elements can be used to measure both the patient’s response to treatments or nursing interventions, at the individual nurse level and correlate with nurse characteristics such as education and experience levels?

How can user stories be used to deconstruct complex nursing interactions so that they might be tested as part of an approved data model to measure nursing value?

Can we leverage patient and nurse level data to identify new micro-costing methods, to predict patient care trajectories that improve outcomes of care, and link nurses and patient to identify best nursing outcomes?
Conclusion
The imperative is to better understand how nursing care is expended for each person/patient and measure the specific costs of that care, associated with defined outcomes. Prior research measures nursing care in the aggregate, such as average number of nurses on a unit or nursing skill-mix on a unit. This presentation will describe the ample data in emerging EHRs and other electronic systems and how it can be used to quantify nursing care at the individual nurse level. The structured methodology presented provides a model for usable and actionable information that can be the basis to improve the quality and efficiency of nursing care delivery systems. Such actionable information provides understanding of how to optimize nurse staffing and assignment patterns that will improve outcomes by aligning individual nurse’s characteristics (experience, education, certification) to individual patient’s needs. The nursing value model, data dictionary, user stories and their corresponding analytic and business intelligence tools will provide a mechanism for the use of data sensitive to the practice of nurses in big data and value-based measurement, providing a metric by which the unique value of nursing contribution in patient outcomes can be measured.

Working Group Affiliation: Endorsement of this panel presentation is provided by the Nursing Informatics Working Group (NIWG).

References

Clinical Informatics in Medical Education: Innovations from the AMA Accelerating Change in Medical Education Initiative

William R. Hersh, MD1, Susan Skochelak, MD2, Anderson Spickard, MD3, Blaine Takesue, MD4, Paul Gorman, MD1

1Oregon Health & Science University, Portland, OR; 2American Medical Association, Chicago, IL; 3Vanderbilt University, Nashville, TN; 4Indiana University, Indianapolis, IN

Competence in clinical informatics is required of all 21st century health professionals. This panel will describe efforts to bring clinical informatics competencies, curricula, and integration into medical school curricula in the context of the American Medical Association initiative, Accelerating Change in Medical Education. After an overview of the initiative by one panelist, the other three will describe specific projects in curriculum development in clinical informatics for medical students, data-driven monitoring of medical student achievement and progress, and the development and use of a teaching electronic health record.

Modern physicians, nurses, and other health care providers must have knowledge and skills in applying modern informatics tools, from the electronic health record to information and knowledge retrieval systems [1]. They must have a practical understanding of key informatics issues that impact clinical practice, including privacy and security, clinical decision support, patient safety, and more.

In 2013, the American Medical Association (AMA) launched an initiative called Accelerating Change in Medical Education (ACE), with the aim of creating “medical schools of the future” [2]. The AMA awarded five-year grants to 11 medical schools to undertake innovations in medical education. An additional 21 schools were brought into the consortium in 2016, although funded at a lower level. The foci of the awardees was diverse and included projects to introduce competency-based education and lifelong learning, shorten the duration of medical school, and expand curricula beyond traditional basic and clinical sciences. One focus across the consortium has been to expand curricula under the rubric of health systems science (also known as healthcare delivery science or the “third science”) [3], including the publication of a new textbook on the topic [4].

Four of the original 11 grantees had components of clinical informatics in their project plans, and a subgroup of interest focused around aspects of clinical informatics in the curriculum. Each of the four medical schools had distinct projects, yet they collectively addressed a spectrum of key issues related to implementing competencies in clinical informatics medical training. The four institutions included Indiana University, New York University, Oregon Health & Science University (OHSU), and Vanderbilt University.

Although the development of competencies and curricula in clinical informatics for clinicians is not limited to physicians, the AMA ACE initiative has provided unique resources to enable aspects of clinical informatics in the training of health professionals that can be adapted to other health professionals besides physicians. The overall goal of the panel is to share the projects and lessons learned from the initiative to inform other efforts and stimulate potential collaboration. The panel will begin with an overview of the AMA ACE initiative, which sets the context for three of the grantees to describe their work. Each panelist will speak for 12-14 minutes and leave at least 30 minutes for audience interaction to provide feedback on work done and engage in dialogue for collaboration going forward.

This panel is timely and relevant given the increasing role that informatics plays in the lives of clinicians, especially the near-universal adoption of electronic health records in the United States and increasing engagement and use by patients, researchers, and other of information technologies.

Dr. Skochelak will be the initial speaker and will present a big-picture view of the AMA ACE consortium. She will then drill down to describe the relevance of clinical informatics to the consortium and describe AMA efforts to lead innovation in medical education across the country.

Dr. Hersh will next describe efforts at Oregon Health & Science University (OHSU) to thread clinical informatics throughout the four years of a newly revised MD curriculum. He will discuss initial efforts to establish competencies in clinical informatics for medical education and then detail how the curriculum has been implemented. He will also
describe barriers and challenges, which include the critical need for student assessments, managing competition for curricular time, and engaging students who are technology-savvy but possess very little understanding of its optimal and appropriate use in healthcare.

Dr. Spickard will then discuss how Vanderbilt has implemented a restructured curriculum in which medical student learning is competency-based, individualized, and situated in the workplace. He will describe innovative means of gathering student performance information through the electronic health care record and mobile assessment applications. He will share lessons learned after 4 years of implementation of a robust electronic portfolio system that collects and displays student data for analysis to inform individual and programmatic success.[5].

Finally, the Regenstrief Institute at Indiana University developed an electronic health record (EHR) designed specifically to support education. Dr. Takesue will describe the implementation and use cases of the teaching EHR (tEHR) at Indiana University and several other undergraduate and graduate health professions schools. He will discuss challenges in using real, misidentified data for education. He will also discuss the challenges in developing and executing a clinical informatics curriculum in schools without a robust clinical informatics tradition.

Dr. Gorman will serve as the panel moderator and also share his experiences when appropriate concerning implementation of clinical informatics curriculum at OHSU.

Although discussion with the audience will be attendee-driven, we propose some questions to facilitate participation:

- What strategies have we used to successfully address the challenges of incorporating informatics concepts and competencies?
- How are we using the EHR educationally?
- How can clinical information systems and specially designed systems be used to capture student performance data?
- How can these methods be transported to other disciplines?
- What strategies have we used to successfully address the challenges of incorporating informatics concepts and competencies?
- How can clinical information systems and specially designed systems be used to capture student performance data?

This topic is timely and urgent because competence in informatics is not limited to informatics professionals. All health care professionals must have skills and knowledge in informatics to practice optimally in the 21st century. Although focused on medical school curricula, this panel will provide insights and methods applicable to all health care professionals. The panel will likely draw an audience not only of educators, but all conference participants who are concerned with the informatics competence of health care professionals generally.

The panel organizer (WRH) states that all participants have agreed to take part on the panel. A CV for each presenter (“NIH-style” bio-sketch is preferred) is provided.

References

The Best of Imaging Informatics Research 2017

Panelists: William Hsu, PhD1, Charles E. Kahn, Jr., MD, MS2
1Department of Radiological Sciences, University of California, Los Angeles, CA;
2Department of Radiology and Institute for Biomedical Informatics, University of Pennsylvania, Philadelphia, PA

Abstract
The field of imaging informatics is rapidly advancing in its ability to address challenges related to clinical big data and harnessing this information for precision medicine. In the past year, the field has experienced growth in a variety of areas including machine learning (learning from imaging data for diagnostic and prognostic predictions), radiomics (the generation of high-dimensional features from images), patient-oriented sharing and communication of images and image-derived findings, and the growth of the use of imaging beyond radiology. Novel approaches go beyond pixel data to integrate imaging with other biomedical data, standardize imaging workflows, and improve the quality and utility of image-derived information in clinical practice. In this session, we will review key advances in imaging informatics research published this past year.

Overview
Biomedical informatics spans a spectrum from basic to applied research, and from molecular to population scale. Imaging informatics is a discipline that focuses on improving patient outcomes through the effective use of images and imaging-derived information in research and clinical care. As the storage and display of multidimensional image data have become commonplace, imaging informatics has evolved from primarily characterizing information from pixels and voxels in images to integrating evidence from images and other biomedical data sources to provide a systems-level understanding of disease etiology and progression. These developments have underscored the need to engage individuals from different disciplines to address shared challenges in managing heterogeneous datasets, developing common standards, and interpreting multiscale evidence for precision medicine. This session provides attendees with a topical review of new and interesting developments in imaging informatics in areas of quantitative image analysis, image annotation and markup standards, data curation and integration with clinical, pathology, and genomic datasets, deep learning and machine learning for clinical decision support, and patient-oriented image sharing. We expect this session to appeal to AMIA attendees given its emphasis on shared challenges and techniques such as data standards, information extraction, annotations, and predictive modeling that are broadly studied and applicable across the biomedical informatics field.

Methods
We will perform a broad search of relevant papers published in the past twelve months using targeted PubMed, Web of Science, and Google Scholar queries. To ensure coverage, we will also review articles published in Radiology, Radiographics, Journal of the American Medical Informatics Association, Journal of Biomedical Informatics, and Journal of Digital Imaging, Journal of Pathology Informatics, Neuroimage, Medical Image Analysis, and Academic Radiology. Abstracts will be reviewed and ranked by the two panelists with disagreements discussed. Identified articles will also be supplemented by a call for nominations from luminary imaging informatics groups, including the AMIA Biomedical Imaging Informatics Working Group, the Radiology Informatics Committee of the Radiological Society of North America (RSNA), the Commission on Informatics of the American College of Radiology, and the Board of Directors of the Society for Imaging Informatics in Medicine (SIIM).

Drs. Hsu and Kahn will present the best of research in ontologies, structured data capture, decision support, radiomics, data integration and predictive modeling, novel healthcare applications, and other science focusing on data other than the images themselves. The content will be divided between the two panelists based on the final set of papers that are identified as a result of the selection process.

Panel Organizer Statement
The panelists listed above have agreed to present the material and lead a discussion. A similar session will be presented at the RSNA annual meeting in December 2017. This session is supported by RSNA, SIIM, and the AMIA Biomedical Imaging Informatics Working Group.
Can the Clinical Information Modeling Initiative (CIMI) Enable the Semantic Interoperability Promise of Fast Healthcare Interoperability Resources (FHIR®©)?

Stanley M. Huff, MD⁠¹, Julia Skapik, MD, MPH⁠², Claude Nanjo, MPH, MAAS⁠³
¹ Intermountain Healthcare, Murray, UT, ² Department of Biomedical Informatics, University of Utah, Salt Lake City, UT, ³ Office of the National Coordinator for Health IT (ONC), Washington, D.C., ⁴ Cognitive Medical Systems, San Diego, CA

Abstract As implied directly in its name, the Health Level 7 (HL7) Fast Healthcare Interoperability Resources (FHIR) standard is intended to enable semantic interoperability. However, the base FHIR-defined classes (known as FHIR “resources”) do not have sufficient specificity to enable such interoperability. Constrained versions of those classes (known as FHIR “profiles”) are intended to provide sufficient specificity to support the semantically interoperable exchange of clinical data. However, multiple, semantically inconsistent FHIR profiles on the same topic can and have been developed; there is no governance around profile development; many profiles are still insufficiently specified to enable semantic interoperability; there is no mechanism to ensure consistency between FHIR profiles; and vendor support for FHIR has focused primarily on resources and not profiles, and are therefore potentially (and likely) inconsistent with one another. HL7 CIMI is working with the FHIR team to address these problems and to enable FHIR to be truly semantically interoperable. This panel will be an interactive discussion to address two questions: first, is FHIR capable of supporting true semantic interoperability without an effort such as CIMI? And second, is CIMI the answer to FHIR’s challenges in this area?

Description Dr. Huff will serve as the moderator and introduce each of the panel members. He will then provide a brief overview of the topic (5 min). Dr. Skapik will provide a position statement that FHIR is a highly promising platform for interoperability but that it alone is insufficient for achieving true semantic interoperability (10 min). Dr. Huff will provide a position statement that CIMI and its model-driven approaches are essential for achieving true semantic interoperability (10 min). Mr. Nanjo will provide a position statement that CIMI can be leveraged by the FHIR community to enable true semantic interoperability (15 min). These position statements will be followed by an interactive panel discussion (50 min) on the proposed position statements and will be moderated by Dr. Huff.

Intended Audience The intended audience for this interactive panel are developers and implementers of the FHIR standard, including EHR vendors, clinical informaticists, standards developers, and healthcare leaders and policy makers.

Introduction of the Topic As noted explicitly in its name, The HL7 FHIR standard is intended to enable semantic interoperability. The primary focus of the FHIR standard to date has been on its definition of classes such as Condition, Encounter, Patient, and Observation, which are known as FHIR “resources”. However, FHIR resources alone do not have sufficient specificity to enable semantic interoperability. In particular:
- **FHIR resources may not represent important concepts.** For example, the Patient resource lacks a race attribute, which is commonly used in the U.S. This situation is due to FHIR resources not including attributes unless they are commonly collected globally.
- **FHIR resources make most attributes optional.** Because FHIR resources can be used in a wide variety of contexts, most attributes are optional. For example, the value of an Observation is optional in FHIR.
- **FHIR resources have minimal terminology constraints.** Because FHIR resources are intended to be used in a wide variety of contexts, terminology constraints are minimal in nature. For example, codes to identify Observations, Conditions, Procedures, and Medications can use any terminology, including local or vendor-specific codes.
- **The FHIR “Basic” resource, combined with FHIR’s limitless “extension” capability, allows any data structure to be represented in any number of ways while remaining “FHIR compliant.”** For example, a Basic resource with tires, doors, and an engine can be defined to represent a car and would be considered completely FHIR compliant, as would a Basic resource with a windshield, drivetrain, and hood ornament.
As a consequence, for example, an EHR vendor returning laboratory results as an Observation can be fully compliant by using a vendor-specific, or even implementation-specific, code to represent a lab and by omitting the value completely. It should be clear, then, that on its own, saying that one supports FHIR resources is practically meaningless in terms of semantic interoperability.

In order to address this interoperability challenge, FHIR allows the definition of more specific versions of the resources as FHIR Structured Definitions, more commonly known as FHIR “profiles”. These profiles can specify, for example, that a U.S. laboratory result observation must use LOINC for its code and use UCUM as its units of measure.

Despite the potential promise of profiles, in practice they are insufficient for achieving semantic interoperability. In particular, multiple, semantically inconsistent FHIR profiles on the same topic can and have been developed; there is no governance around profile development; many profiles are still insufficiently specified to enable semantic interoperability; there is no mechanism to ensure consistency between FHIR profiles; and vendor support for FHIR has focused primarily on resources and not profiles, and are therefore potentially (and likely) inconsistent with one another.

HL7 CIMI was established to rigorously and consistently define the semantics of clinical information models. The HL7 CIMI Work Group has been working with the HL7 FHIR team to address challenges to semantic interoperability and to enable FHIR to be truly semantically interoperable. In particular, CIMI is developing models from which FHIR profiles can be auto-generated, and it is establishing governance on the selection of a preferred model and associated FHIR profile among multiple semantically equivalent representations of the same information.

**Aim and Expectations of the Discussion**

The aims of the discussion are to (i) inform the AMIA audience on the interoperability limitations of FHIR and work that is currently underway to address it and (ii) obtaining constructive criticism and feedback on the proposed approach from the audience.

**Contribution of Each Speaker**

Dr. Huff is Professor (Clinical) of Biomedical Informatics at the University of Utah, and the Chief Medical Informatics Officer at Intermountain Healthcare. He is currently a co-chair of the LOINC Committee, and a co-chair of the HL7 CIMI Work Group. He will moderate the panel and make the case for CIMI as the basis of semantic interoperability in health care.

Dr. Skapik is an internist and Medical Officer at ONC. She has been working with ONC for the past several years on clinical quality standards, terminology, policy and programs. The culmination of the standards work is the Clinical Quality Framework, which has taken disparate clinical quality measurement and clinical decision support standards and moved them into a unified modular stack within FHIR. She hopes that this work, when fully integrated with a robust data element specification and detailed clinical models like CIMI, could form the basis for a unified clinical data model that will enable seamless but consistently meaningful data capture by end users and a basis for future health IT certification. She will make the case that FHIR alone is insufficient to achieve true semantic interoperability.

Mr. Nanjo is Chief Scientist at Cognitive Medical Systems, and he has been a leader in the development of FHIR profiles for the purposes of clinical decision support and electronic clinical quality measurement. He is also developing tooling to allow CIMI models to auto-generate FHIR profiles. He will make the case that CIMI can enable FHIR to achieve true semantic interoperability.

**Why this Topic is Timely and Controversial**

FHIR is rapidly being advanced as a panacea to semantic interoperability in health care. However, early and deep adopters of the FHIR standard, including the members of this interactive panel, have identified significant issues with achieving semantic interoperability using FHIR alone. This issue is particularly timely because important stakeholders including healthcare leaders, vendors, and policy makers may be advancing FHIR without a full understanding of its current limitations. Addressing this issue will be critical for achieving the promise of interoperability envisioned by the FHIR community. This topic is controversial because many stakeholders are under the impression that FHIR alone can achieve semantic interoperability.
Participation Statement
All panelists have agreed to participate in the panel if it is accepted.
Big Data in the Intensive Care Unit*

Andrew James, MBChB¹, Mohammad Adibuzzaman, PhD², John Zaleski, PhD³, Peter Haug, MD⁴
¹Department of Paediatrics, University of Toronto, CA, ²Regenstrief Center for Healthcare Engineering, Purdue University, Indiana, USA, ³Bernoulli, USA, ⁴Homer Warner Center, Intermountain Healthcare and Department of Biomedical Informatics, University of Utah, USA

Moderator
Soojin Park, MD⁵
⁵NY Presbyterian Hospital and Columbia University College of Physicians and Surgeons

Abstract
Big data has brought much promise for discovery of treatment and therapies, drug safety, and the care delivery processes by identifying which treatment would work best for which patients. NIH has recently taken the initiative, ‘All of Us’, to collect one million or more patients’ data (electronic health records (EHR), imaging, genomics, environmental data, etc.) over the next few years. The intensive care unit represents a unique data source in this context, with carefully captured detailed high volume data from different systems such as EHR, electrocardiogram, blood pressure, infusion pumps, and photoplethysmogram, among other data. However, the mere availability of data does not translate into knowledge or improved outcome. Questions remain on what data is needed, how to integrate these high volume data with high throughput infrastructure for near real-time decision making by the clinicians. In this panel, we present our work in integrating this heterogeneous high volume data with state of the art technologies for retrospective analysis and near real-time decision making with different systems such as Medical Information Mart from Intensive Care Unit (MIMIC III) and Artemis: both from technological and clinical perspectives.

Panel Description
The panel addresses the technological challenges to overcome the gap between large volume of data collection and useful applications with the data. The discussions would focus on high throughput analysis, both real time and retrospective, and specific clinical applications of such systems in the domain of prediction of sepsis, multivariate alarm generation, and diagnosis and treatment of pneumonia.

Brief description of each panelist’s discussion
1. The Artemis Project: Real-time intelligent data analysis in the NICU. Andrew James, MBChB
   Enormous data is collected in the ICU setting from different sources such as nursing notes, infusion pumps, and physiological data from as many as sixteen different streams. Clinicians often are overwhelmed with the volume of data, and cannot make data driven decisions because of the gap between data collection and necessary tools for real-time data analysis and visualization. For example, it is very common for critically ill babies to have significantly abnormal variation in the measured parameters minute by minute that are not recorded in the medical record. Motivated by this need, the Artemis project was initiated to develop the necessary software and hardware tools for real-time continuous ingestion of high volume data for robust temporal analysis to identify variations that are temporally associated with conditions of interest. As a use case, the early diagnosis algorithm for onset of neonatal sepsis, which is a leading cause of morbidity, is illustrated.

2. An integrated high throughput open access analysis platform for the MIMIC database for retrospective analysis. Mohammad Adibuzzaman, PhD
   Medical Information Mart for Intensive Care (MIMIC) database is a well-known database in the scientific community for detailed data captured from an intensive care unit and freely available for research and collaboration. Although this database is a great resource for scientific analysis, it is not straightforward to do a holistic analysis of the clinical

* Sponsored by the Intensive Care Informatics Working Group
In this talk, we present a new cloud based model for collaborative access, exploration, and analyses of the MIMIC III database for translational clinical research (1). The proposed model addresses the significant disconnect between data collection at the point of care and translational clinical research. It addresses problems of data integration, pre-processing, normalization, analyses (along with associated compute back-end), and visualization. The proposed platform is general, and can be easily adapted to other databases. The pre-packaged analyses toolkit is easily extensible, and allows for multi-language support. The platform can be easily federated, mirrored at other locations, and supports a RESTful API for service composition and scaling.

3. Multivariate real-time data can provide insight into cascades towards adverse events. John Zaleski, PhD

In recent years focus has been on alarm signal reduction in critical care and, more recently, general care floors regarding mitigation of false alarm signals and other non-actionable alerts to clinical staff, to both mitigate alarm fatigue and improve patient safety. Alarm signals, in general, provide reactive mechanisms for indicating a potential problem with patients. Yet, oftentimes the trends leading towards the resulting alarm signal can be detected earlier through an understanding of the behavior of measured signals that, individually, may not be interesting. Yet, when taken together, paint a picture of deterioration that can be used to suggest earlier intervention to head off an adverse event. One area in particular is that of opioid-induced respiratory depression, and the interaction between patient respiratory drive, the onset of CO2 narcosis, and the detection of trending events. This talk illustrates the case for using multi-variate data through real-time measurement to indicate trending behavior that can lead towards adverse events.

4. Impact of electronic clinical decision support tool with high volume data for diagnosis and treatment of pneumonia. Peter J Haug, MD

The talk illustrates the usefulness of big data systems in clinical settings for the diagnoses and treatment of specific clinical problems, such as pneumonia and sepsis, from the experience of Intermountain Healthcare. Despite evidence that guideline adherence improves clinical outcomes, management of pneumonia patients varies in emergency departments (EDs) and ICUs. Motivated by this problem, the effect of a real-time, electronic clinical decision support tool that provides clinicians with guideline-recommended decision support for diagnosis, severity assessment, disposition (including recommending ICU admission), and antibiotic selection is presented(2, 3).

Learning Objective

After participating in this panel, the attendant should be able to:

1. Identify the technological challenges to integrate and analyze high volume data in ICU setting, both for real time and retrospective analysis.
2. Identify the need for such systems in care delivery process.
3. Learn about specific clinical applications of such systems, in the domain of prediction of sepsis, multivariate alarm generation, and prediction of pneumonia.

Biography of the panelists

1. Andrew James, BSc, MBChB, MBI, FRACP, FRCPC

Dr. Andrew James is the Interim Clinical Director, Neonatal Intensive Care Unit at The Hospital for Sick Children, and an Associate Professor, Department of Paediatrics, Faculty of Medicine, University of Toronto. He obtained his medical degree from the University of Auckland, Auckland, New Zealand and trained in neonatology at National Women’s Hospital, Auckland and The Hospital for Sick Children, Toronto, Canada. Dr James has a graduate degree in biomedical informatics from the Department of Medical Informatics and Clinical Epidemiology, Oregon Health & Science University, Portland, Oregon. He has also completed the Certificate in Continuing Studies in E-Learning, School of Continuing Studies at the University of Toronto. Dr. James’ current activities include knowledge representation, the visual display of information, predictive analytics, and advanced clinical decision support. He is co-lead for the Artemis Project: a collaborative, multidisciplinary, international team that is developing advanced clinical decision support applications for the early identification of late onset neonatal sepsis, the recognition and classification of neonatal spells, the recognition of pain, the recognition of sleep/wake cycling in newborn infants and the quantification of retinal exposure to oxygen. Other active projects include the structured representation of the
clinical care of the newborn infant by SNOMED CT; the development of SNOMED RefSets for disorders of the newborn infant; and the ontological representation of the clinical care and disorders of the newborn infant.

2. Mohammad Adibuzzaman, PhD

Mohammad Adibuzzaman is a research scientist at the Regenstrief Center for Healthcare Engineering located in Purdue University, Indiana. His research interest includes mathematical model development and big data system for understanding of healthcare data for better patient outcomes. He has a PhD in Computational Sciences from Marquette University. He received his Master’s degree in Computational Sciences in 2012 from the same university. Before going to Marquette, he worked as a Junior Research Assistant at the National University of Singapore and as a Software Engineer in Bangladesh. He also worked as an Oak Ridge Institute of Science and Engineering (ORISE) Fellow at the U.S. Food and Drug Administration in 2013 and 2014.

3. John Zaleski, PhD

John Zaleski, PhD, CAP, CPHIMS, is Chief Analytics Officer of Bernoulli, a leader in real-time connected healthcare. Dr. Zaleski brings 21 years of experience in researching and ushering to market devices and products to improve healthcare. He received his PhD from the University of Pennsylvania, with a dissertation that describes a novel approach for modeling and prediction of post-operative respiratory behavior in post-surgical cardiac patients. Dr. Zaleski has a particular expertise in designing, developing, and implementing clinical and non-clinical point-of-care applications for hospital enterprises. Dr. Zaleski is the named inventor or co-inventor on seven issued patents related to medical device interoperability. He is the author of numerous peer-reviewed articles on clinical use of medical device data, information technology and medical devices and wrote three seminal books on integrating medical device data into electronic health records and the use of medical device data for clinical decision making, including the #1 best seller of HIMSS 2015 on connected medical devices.

4. Peter J. Haug, MD

Dr. Peter J. Haug earned his MD from the University of Wisconsin, Madison. Currently he is a Medical Informatics Director and a member of the Homer Warner Center for Informatics Research at Intermountain Healthcare and is a professor in the University of Utah’s Department of Biomedical Informatics, where he is involved in research and postgraduate education. His expertise includes the following: the study of tools for natural language processing in medicine; the evaluation of probabilistic decision support systems for use in the health-care process; the development of computer-based tools to deliver detailed medical protocols; and various applications of data mining to assist in the development of innovative medical software. Previously, he developed components of two different medical information systems. Currently, he is focusing on the construction of environments designed to support applied informatics research and to support the testing of innovative tools for implementing decision support in active clinical settings. His research explores natural language processing to extract clinical information from medical documents, sharable approaches to representing decision support logic, models of disease capable of informing clinical care, and secondary uses of data collected in the course of care. Joining the department in 1983, Professor Haug, MD, is a Fellow of the American College of Medical Informatics, and co-chairs the Arden Syntax Working Group, Health Level VII. He is certified by the American Board of Internal Medicine and the National Board of Medical Examiners.

References


Big Data to Knowledge (BD2K) and the Application of Metadata

Guoqian Jiang, MD, PhD1, Walter Campbell, PhD, MBA2, Timothy Clark, PhD3, Cui Tao, PhD4, Mark Musen, MD, PhD5

1Mayo Clinic, Rochester, MN; 2University of Nebraska Medical Center, Omaha, NE; 3Massachusetts General Hospital, Boston, MA; 4University of Texas Health Science Center – Houston, Houston, TX; 5Stanford University, Stanford, CA

Abstract
Clinical and translational research studies increasingly involve the manipulation of large datasets. There is a critical need to develop metadata-driven informatics technology to both lower the effort required for, as well as incentivize the process of applying metadata in clinical and translational research. The NIH Big Data to Knowledge (BD2K) initiative seeks to develop an interactive data ecosystem through adopting the FAIR principles (i.e., data must be findable, accessible, interoperable and reusable), and is supporting various technology platforms in these contexts to make it easier for researchers to share their data with the community so that it can broadly benefit the community. This thematic panel will highlight five projects funded through the BD2K initiative that are developing technology infrastructure and tools for metadata management, standardization and applications. Attendees will learn about novel open source metadata informatics platform, and how data/metadata standards are used in data sharing platforms. Attendees can engage with panelists on metadata challenges and potential solutions in support of clinical and translational research applications.

Introduction
Clinical and translational research studies increasingly involve the manipulation of large datasets. There is a critical need to develop metadata-driven informatics technology to both lower the effort required for, as well as incentivize the process of applying metadata in clinical and translational research. The NIH Big Data to Knowledge (BD2K) initiative seeks to develop an interactive data ecosystem through adopting the FAIR principles (i.e., data must be findable, accessible, interoperable and reusable), and is supporting various technology platforms in these contexts to make it easier for researchers to share their data with the community so that it can broadly benefit the community. This thematic panel will highlight five projects funded through the BD2K initiative that are developing technology infrastructure and tools for metadata management, standardization and applications.

The use of data standards plays a key role in these applications of metadata as existing and emerging data standards can enhance comparability and consistency of heterogeneous data sources. The data standards used in these projects include clinical data standards such as HL7 FHIR and ONC national standards, the data citation standards such as JDDCP, and vocabulary standards through biomedical ontology services and developments. The use cases and application domains included in some of these applications are generic (e.g., focusing on broad clinical and translational research studies) whereas some applications are very specific (e.g., focusing on data regulation and citation).

Metadata Topics
In this panel, we highlight five metadata technology development projects that describe various platforms to demonstrate the informatics platform and framework applying metadata that would drive data discovery in the era of rapidly accumulating big data. These initiatives are described briefly below:

1. **Tools for indexing clinical research metadata using HL7 FHIR (Guoqian Jiang)**

   Fast Healthcare Interoperability Resources (FHIR) is an emerging HL7 standard; it leverages existing logical and theoretical models to provide a consistent, easy to implement, and rigorous mechanism for exchanging data between healthcare applications. However, currently the toolbox that enables HL7 FHIR as a global data model to standardize clinical research metadata is very limited. We aim to provide a scalable standards-based framework that enables effective and efficient big data integration and semantic sharing leveraging emerging Semantic Web technologies. In this panel, we will present the methods and tools we are developing at Mayo Clinic for indexing clinical research metadata, including 1) a semantic data element repository for managing metadata in HL7 FHIR; 2) a schema.org extension for enabling
2. Deploying ONC National Standards in Support of Metadata for Big Data Research Warehouse Management of Repurposed Laboratory, Pathology and Patient Findings Data from the EHR (Walter Campbell)

The Patient Centered Outcome Research Network (PCORnet) seeks to support large scale, multi-site outcomes research through collection and aggregation of data across a broad array of medical centers. Central to the success of PCORnet is the ability to perform federated data queries across numerous heterogenous clinical data warehouse (CDW) instances. Researchers at UNMC propose the use of ONC standards as metadata as a mechanism for normalizing data across multiple, diverse CDW’s. In this panel, we will present how UNMC deployed ONC standards metadata into its i2b2-based CDW including the complex polyhierarchies of SNOMED CT and RxNorm to support federated queries across PCORnet sites. Included in this discussion is the development and deployment of an observables ontology for pathology data, as well as, the current status of software tooling developed and available for ONC metadata deployment in an i2b2 environment.

3. JDDCP-Compliant identifiers and Metadata for Biomedical Publishers and Data Repositories (Timothy Clark)

The BD2K-sponsored Data Citation Implementation Pilot (DCIP) was run from late 2015 through 2017 in conjunction with bioCADDIE, a project to develop a data discovery index for biomedicine. DCIP’s goal was to develop, enable and encourage adoption of a common model for data to be archived in robust repositories, cited in publications, identified by globally unique persistent identifiers, and discoverable by search engines and indexers. DCIP’s methodology was to convene expert groups of early adopters across several communities to develop common specifications and initiate adoption, while coordinating with metadata harvesters and standards bodies. This talk will briefly report on the metadata and identifier approaches that developed out of the DCIP program and how they can support more robust, reusable and translatable biomedical research.

4. Metadata Applications On Informed Content To Facilitate Biorepository Data Regulation And Sharing (Cui Tao)

The application of suitable metadata in support of the complex set of regulatory, legal, privacy and security requirement processes and information flows involved in regulated research is a field in an early phases of development. The complicated legal and technical requirements involved in the processes challenge our ability to effectively build information systems that support sharing of research data, specimens and other research artifacts at scale. Our project is trying to develop suitable machine-based metadata representation and automatic annotation of regulatory processes focusing on informed consent forms and the associated documents. Our approach will presumably provide a formal and computable basis for data sharing and information release policies.

5. Intelligent Authoring of Metadata for Translational Research (Mark Musen)

The Center for Expanded Data Annotation and Retrieval (CEDAR) is developing methods to ease the authoring of the metadata needed to annotate data sets from experiments in translational research. Using community-derived metadata standards and controlled terms from biomedical ontologies, CEDAR offers a Web-based platform that simplifies the development of complete, comprehensive metadata. CEDAR has been deployed in several real-world settings, and we will discuss user experiences with the technology. We also will discuss how CEDAR-based software might assist with the authoring of metadata for more clinical applications.

This panel is timely and relevant given the urgent need to share clinical research data to facilitate clinical and translational data discovery and analytics, and advance our understanding of human health and disease. Initiatives such as the precision medicine and cancer moonshot will spur the generation of large datasets. We need appropriate tools and technologies applying metadata to manage and analyze these to inform clinical research and patient care.

Discussion questions

1. What would you like to see in a research-data sharing platform?
2. What types of metadata/data should be shared?
3. What metadata tools will be most useful to help your translational research?
4. What are your most critical metadata/data sharing challenges?

Learning Objectives

1. Attendees will learn about novel open source metadata informatics platform
2. Attendees will learn about how data/metadata standards are used in data sharing platforms
3. Attendees can engage with panelists on metadata challenges and potential solutions in support of clinical and translational research applications

Organizer statement: All speakers have agreed to attend AMIA 2017 Annual Symposium and participate on this panel.

Acknowledgement

The five metadata projects highlighted in this panel proposal are supported by the following NIH Awards: U01 HG009450 (PI: Jiang); U01 HG009455 (PI: Campbell); 3U24 AI117966-02S1 (PI: Ohno-Machado), U01HG009452 (PI: Clark); U01 HG009454 (PI: Tao); and U54 AI117925 (PI: Musen).
Enabling Knowledge-Driven Care at Scale through CDS Hooks and the FHIR Clinical Reasoning Module

Kensaku Kawamoto, MD, PhD, MHS¹, Kevin Shekleton,² James Doyle³, Bryn Rhodes⁴, Howard R. Strasberg, MD, MS⁵

¹Department of Biomedical Informatics, University of Utah, Salt Lake City, UT,
²Cerner Corporation, Kansas City, MO,
³Epic Systems Corporation, Verona, WI,
⁴HarmonIQ Health Systems Corporation, Orem, UT,
⁵Wolters Kluwer Health, San Diego, CA

Abstract

Interoperable clinical decision support (CDS) holds great promise for ensuring that patients receive the best care possible based on the latest medical knowledge. Building on prior efforts at standards-based CDS interoperability, as well as rapid adoption of the HL7 FHIR standard by the EHR vendor community, various stakeholders including EHR vendors, knowledge vendors, and healthcare systems are working together to enable interoperable knowledge-driven care through the CDS Hooks and FHIR Clinical Reasoning specifications. These specifications are being harmonized into a single unified HL7 specification, and the goal of this unified specification is to enable EHR systems to both (i) invoke external CDS Web services and (ii) consume external knowledge artifacts such as order sets, rule specifications, and documentation templates in a standard manner. In this panel, the latest developments will be discussed by leaders of this effort from the EHR vendor community, the CDS knowledge vendor community, the standards development community, and the healthcare provider community. The panel will demonstrate interoperable CDS implemented on multiple EHR platforms using this approach. The panelists will also describe key challenges and lessons learned, solicit feedback from the audience, and describe planned future work to continue advancing health and health care through interoperable CDS.

Description

The panel will be organized as follows:

<table>
<thead>
<tr>
<th>Time</th>
<th>Speaker</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>10 min</td>
<td>Strasberg</td>
<td>Need for CDS interoperability, history, and challenges</td>
</tr>
<tr>
<td>15 min</td>
<td>Rhodes</td>
<td>Overview of unified CDS specification (CDS Hooks + FHIR Clinical Reasoning)</td>
</tr>
<tr>
<td>15 min</td>
<td>Shekleton</td>
<td>Cerner perspective</td>
</tr>
<tr>
<td>15 min</td>
<td>Doyle</td>
<td>Epic perspective</td>
</tr>
<tr>
<td>15 min</td>
<td>Kawamoto</td>
<td>Health care system perspective</td>
</tr>
<tr>
<td>20 min</td>
<td>All</td>
<td>Panel discussion with audience</td>
</tr>
</tbody>
</table>

Dr. Strasberg will serve as the moderator and introduce each of the panel members and their organizations. Dr. Strasberg will then describe why CDS interoperability is needed, briefly summarize prior work, and outline challenges to achieving interoperable knowledge-based care. This overview will be followed by presentations by the panelists, who will describe the unified FHIR-based standards for CDS interoperability (Rhodes); provide EHR vendor perspectives for supporting these CDS interoperability standards (Shekleton and Doyle); and describe how interoperable CDS can and is improving patient care at the healthcare delivery system level (Kawamoto). Panelists will also showcase interoperable CDS using this standard approach. These presentations will be followed by a panel discussion with the audience moderated by Dr. Strasberg.
Strasberg: Dr. Strasberg is VP of Medical Informatics at Wolters Kluwer Health, a leading CDS knowledge vendor with CDS solutions such as UpToDate® and Medi-Span®. He is co-chair of the HL7 CDS Work Group. Dr. Strasberg leads Wolters Kluwer’s efforts in interoperable CDS. Dr. Strasberg will provide an overview of why CDS interoperability is needed to enable knowledge-driven care at scale. He will also provide a brief history of prior standardization efforts upon which the present initiative builds, and he will discuss key challenges to CDS interoperability.

Rhodes: Mr. Rhodes is EVP of Advanced Consulting at HarmonIQ Health Systems Corporation, a health IT company focused on interoperable solutions to improve patient health. Mr. Rhodes is also the principal subject matter expert for the Clinical Quality Framework (CQF) initiative sponsored by the Office of the National Coordinator for Health IT (ONC) and the Centers for Medicare and Medicaid Services (CMS) to enable interoperable CDS and electronic clinical quality measurement at scale. Mr. Rhodes will describe the unified FHIR-based CDS interoperability specification, which represents a unification of the CDS Hooks and FHIR Clinical Reasoning specifications. This unified specification is anticipated to be balloted as an HL7 standard in September 2017. This unified specification supports both (i) the sharing of CDS capabilities through Web services and, equally importantly, (ii) the sharing of CDS knowledge artifacts (e.g., event-condition-action rules, order sets, and documentation templates) as FHIR resources. The sharing of CDS knowledge artifacts involves the use of the HL7 Clinical Quality Language (CQL) expression language, and the sharing of CDS capabilities through Web services can leverage CQL or any other expression language. Both CDS interoperability approaches utilize the FHIR data model and profiles, which are expected to continue to evolve through the work of various HL7 Work Groups including the HL7 Clinical Information Modeling Initiative (CIMI) Work Group.

Shekleton: Mr. Shekleton is a Vice President and Distinguished Engineer at Cerner Corporation. Mr. Shekleton is the Project Lead of the multi-stakeholder CDS Hooks initiative, which has brought together a diverse group of EHR vendors, CDS knowledge vendors, and healthcare systems around the vision of interoperable CDS. Mr. Shekleton will describe Cerner’s strategy and roadmap for supporting interoperable CDS, which includes providing production-level support for CDS Hooks in its product lines. He will also show examples of interoperable CDS solutions operating in the Cerner environment using these interoperability specifications, and he will provide his thoughts on the future of interoperable CDS at Cerner and across the industry, including his thoughts on how standards-based CDS knowledge artifacts could potentially be supported by EHR vendors such as Cerner.

Doyle: Mr. Doyle is a senior Software Developer and a leader of CDS development and initiatives at Epic, including Epic’s efforts in the area of standards-based CDS interoperability. He will describe why Epic sees value in CDS interoperability, and how Epic is seeking to support the best care possible for every patient through the sharing of best clinical practice and the latest medical knowledge. Mr. Doyle will describe Epic’s strategy and roadmap for interoperable CDS and will provide examples of interoperable CDS solutions operating in the Epic environment. Mr. Doyle will also provide his thoughts on the future of interoperable CDS at Epic and across the industry, including how CDS interoperability could be supported not just through services but also through the sharing of interoperable CDS knowledge artifacts such as rule definitions and order sets.

Kawamoto: Dr. Kawamoto is Associate Chief Medical Information Officer, Assistant Professor of Biomedical Informatics, Director of Knowledge Management and Mobilization, and Chair of the CDS Committee at the University of Utah. Dr. Kawamoto is also co-chair of the HL7 CDS Work Group and co-Initiative Coordinator for the ONC/CMS-supported CQF initiative. Dr. Kawamoto also co-leads the University of Utah’s Interoperable Apps and Services (IAPPS) initiative, which is a multi-stakeholder effort to enable standards-based, interoperable applications and software services to improve health and health care. Dr. Kawamoto will describe why interoperable CDS is urgently needed to enable the ubiquitous availability of advanced CDS across healthcare delivery systems. He will then describe how the University of Utah is supporting the standards described in the panel through an open-source implementation of a CDS Hooks adapter for the Epic EHR platform, as well as through OpenCDS (www.opencds.org), an open-source CDS framework that provides support for standards-based CDS delivery. Dr. Kawamoto will also show how the University of Utah is leveraging these standards to improve patient care and the physician experience, including through the use of these CDS interoperability approaches within SMART on FHIR applications interfaced with the Epic EHR. Dr. Kawamoto will also discuss key lessons learned and the work that remains in achieving optimized patient care supported by effective, comprehensive, and interoperable CDS.
**Strasberg:** Dr. Strasberg will lead a moderated discussion with the audience. The objectives of this discussion will be to answer questions from the audience, engage in a stimulating exchange of ideas, and gain further insights on the topic from audience members’ collective experiences. The questions listed below will be used to stimulate this discussion.

**Significance of panel topic and anticipated audience**

Effective CDS is critical for ensuring that patients receive optimal patient care and for healthcare systems to survive and thrive in a value-based payment environment. While history has shown that it is not feasible for individual healthcare organizations to develop the breadth and depth of effective CDS needed for realizing the full potential for knowledge-driven care, the unified CDS Hooks/FHIR Clinical Reasoning specification holds the potential to finally realize this vision given the strong support it is receiving from major EHR vendors such as those represented in this panel. Thus, it is imperative for biomedical informaticists seeking to improve the quality and value of health care to gain knowledge and insights into this important and rapidly progressing area of clinical informatics and health care delivery.

**Discussion questions**

- What areas of clinical medicine do you think can benefit most from interoperable CDS?
- What recommendations do you have for standards development and implementation in this area?
- What gaps do you see in the current interoperability frameworks offered by EHR systems?
- What can be done to facilitate order mapping from external CDS systems to local order catalogs?

**Participation statement**

All proposed panelists have agreed to participate in the panel if the proposal is accepted.
Are EHRs “Overloading” Health Professionals? Issues, Advances and New Directions from Cognitive Science and Usability Engineering

Andre W. Kushniruk, Ph.D., FACMI 1, Elizabeth Borycki, Ph.D.1, Yalini Senathirajah2, Darren Hudson3

1 School of Health Information Science, University of Victoria, Canada
2 SUNY Downstate Medical Center, Brooklyn, New York, USA
3 Department of Critical Care Medicine, University of Alberta, Canada

Abstract

Issues about the usability and safety of electronic health records have come to the fore. Reports of systems that user find difficult to use, interfere with healthcare workflows and that make them feel “cognitively overloaded” continue to be reported. The concept of “cognitive load” emerged from the cognitive literature several decades ago and refers to the amount of mental effort required to carry out tasks. In areas such educational design the concept has proven useful, as well as in other areas such as aviation and cockpit design. However, far less work in cognitive load has been applied in the area of human factors in healthcare. In this panel presentation, experts from the United States and Canada will explore the concept of cognitive load, including its definition, its measurement and its implications, for improving the usability and safety. Implications for the design and testing of a wide range of health information systems and technologies will be explored. The learning objectives include increasing awareness of cognitive load, understanding the concept, understanding how it can be measured, and understanding what can be done to decrease “cognitive overload”.

Introduction

Usability testing and clinical simulations can provide data that can be used to assess not only the usability of health information technology (HIT), but also the impact of systems on the cognitive processing of end users of these systems 1-3. Despite increased consideration nationally and internationally about the usability of HIT, reports of unusable, difficult to work with and unsafe HIT continue to be reported worldwide. One major complaint is that these systems may be difficult to use under real contexts of use where there may be time constraints, missing data, and stressful conditions. Working with HIT under these conditions might be difficult and users still complain about inability to process information presented by HIT and potential negative impact on workflow, decision making and reasoning 4.

A concept that has emerged from the cognitive literature is that of cognitive load. Cognitive load refers to the amount of mental effort being used in working memory to carry out tasks. The theory emerged from research in problem solving and the work of John Sweller in the 1980s 5. It has since been applied in design and evaluation in areas ranging from education to aerospace. For example, the NASA Task Load Index has been used to assess the cognitive load of carrying out tasks using technology in a range of areas from combat aviation, commercial aviation, office workers and electrical power groups. Related measures of workload have begun to be applied in medical specialties such as critical care medicine, anesthesia, emergency medicine, pharmacy and nursing health records 6-9. Applying such measures hold promise to help in understanding critical human factors issues with systems such as electronic health records and to help in decreasing the potential for error due to cognitive overload 10. A number of authors have begun to argue that cognitive overload may be responsible for many of the usability, workflow and safety problems being reported in the literature 10. With the increased triggering of system generated medical alerts, complex user interfaces and less than optimal screen designs in HIT, the potential for cognitive overload increases and is beginning to be recognized as being a major problem (and as a major underlying source of many usability and safety issues in healthcare today). As a result, there is a need to develop effective measures of cognitive load that can be used to compare HIT, select systems and to understand when HIT user interfaces and displays need to be reconfigure to reduce cognitive load and resulting potential for technology-induced error in healthcare.
In this panel, experts from Canada and the United States will describe their work in applying assessment of cognitive load to leading to improved user interface design, improved user interaction and safer systems in healthcare. A particular focus of the panel will be on the impact of cognitive load on the inadvertent introduction of technology-induced error, and how methods from human factors and usability engineering can be applied to assess the impact of cognitive load on medical error in order to mitigate such negative impact. This is a new perspective from which to view cognitive load as the literature on technology-induced error in HIT is barely a decade old and needs to be integrated with work on assessing cognitive load. New approaches are needed in the context of making HIT safer from the impact of cognitive load on technology-induced error, which will be discussed and will be a focus. The panelists will also engage the audience in a discussion of the impact and importance of this essential human factors issue (and growing concern) in the design, use and selection of healthcare information systems. The intent will be to make the audience aware of the critical importance of considering cognitive load in understanding problems with system usability and safety and to provide practical ways of measuring cognitive load.

Panelist Presentations:

Andre Kushniruk, Ph.D., FACMI (moderator and panelist) will describe the concept of cognitive load and relate it to the design and usability of a range of healthcare information systems. He will provide a background to the concept from the cognitive literature and will review areas in healthcare information systems where concepts and methods from cognitive science have been applied in an attempt to improve healthcare user interfaces by understanding, measuring and decreasing the cognitive load of health professionals using these systems. He will draw on examples from the area of usability and human factors in healthcare, including the application of a method known as rapid low-cost in-situ usability testing as a way of assessing the impact of systems on users’ cognitive processes and assessing the impact of systems on cognitive load in order to mitigate error and make HIT safer.

Elizabeth Borycki, Ph.D. will describe the integration of “clinical simulations”, i.e. simulations involving in-depth qualitative study of humans (e.g. nurses, doctors) conducted “in-situ” (i.e. in real hospital settings) with simulations that are “computer-based”. The approach involves several phases. First complex human interactions with systems such as medication order entry are analyzed by giving representative users scenarios which are recorded in-situ (i.e. in actual healthcare settings, including recording of user interactions with interfacing technologies – e.g. bar coding). The data from these studies of humans interacting with systems in highly realistic conditions are then analyzed to assess cognitive load and the likelihood of medication error given the presence of specific user interface and design features. Finally, the results from such analysis can be fed into computer-based simulations to form the basis for further modeling and forecasting of impact of systems on a larger scale and over time. The approach can also be used to assess the impact of specific design features that may be causing cognitive load and that may be responsible for causing technology-induced error. It is argued that a range of usability testing and simulation methods can be practically used in conjunction in order to identify user issues and dramatically decrease cognitive load and the risk of technology-induced error, and increase the potential benefits of healthcare information systems.

Yalini Senathirajah, Ph.D. will describe her work in developing composable user interfaces that allow for more flexibility in designing and deploying EHR user interfaces, in an effort to decrease cognitive load of these systems and to arrive at new ways of measuring cognitive load. Almost all current EHR systems have information location fixed by the vendor, with menu-based navigation and information access. This has the result that the user must usually traverse multiple screens and scroll, zoom, pan, and take other actions to view all relevant parts of the record. This ‘display fragmentation’ has consequences for cognitive load as the user must then remember information in mind between screens, loading working memory and also imposing visual and hand coordination requirements. A different ‘composable’ approach in which the relevant data are assembled on the same screen via flexible user-controlled UI can be used to decrease the cognitive load as well as discover better information aggregation and interaction design meeting user needs. Methods of studying cognitive load (such as with eye tracking) and assessing and redesigning EHR UI will be discussed.

Darren Hudson, M.D., M.Sc. will describe his work in integrating the NASA task load index to assess workload in electronic health records. Few studies have specifically studied workload in electronic medical records even though there increased workload is cited as a frequent cause of dissatisfaction. The NASA task load index can be used to understand the factors that impact on the perceived workload associated with the use of electronic medical records in order to make them safer. It is easy to administer and produces consistent results. The NASA task load index is also able to distinguish between different groups of subjects within a study and demonstrates face validity.

Topic Rationale: Issues related to poor usability and potential for introduction of technology-induced error continue to be reported worldwide as HIT such as EHRs are deployed and become ubiquitous. Theory and methods emerging from human factors in the area of measuring cognitive load will become increasingly important to
understand and apply in health informatics to increase the potential benefits of systems and decrease the potential for cognitive overload and associated negative effects on usability and safety of the systems we deploy. Health informatics professionals require better methods for identifying issues related to cognitive load and to formulate approaches to mitigating impact of cognitive load.

**Learning Objectives:** After participating in this panel, attendees will be able to understand issues related to cognitive load in the design, use and testing of HIT such as EHRs. They will also have learned about how cognitive load can be measured in relation to using HIT, and how cognitive overload can be identified and mitigated.

**Discussion Questions:**
- Are currently available vendor based systems causing health professionals to feel they are overloaded or overwhelmed?
- What is the impact of cognitive overload on health IT safety and ultimately patient safety?
- Are negative impacts of systems on workflow related to the health professionals’ ability to process increasing complexity and amounts of information being made available electronically?
- Does the concept of “cognitive load” help to explain the type of end user issues complaints related to lack of fit with workflow, too much information to process in a reasonable time, poor usability etc.?
- What methods can be used to assess cognitive load involved in using HIT?
- How could methods for assessing cognitive load be used to improve system design and customization and provide input when deciding amongst HIT products during procurement?

**Participant Agreement Statement:** All participants have agreed to participate in this panel.

**Keywords:** user-computer interface, patient safety, usability, cognitive load, system evaluation

**Panel Participant Details:**
- **Andre W. Kushniruk**, MSc., Ph.D., FACMI, Professor and Director, School Health Information Science, University of Victoria, Victoria, BC, Canada Email: [andrek@uvic.CA](mailto:andrek@uvic.CA)
- **Elizabeth Borycki**, Ph.D. Professor, School Health Information Science and School of Nursing, University of Victoria, Victoria, BC, Canada Email: [emb@uvic.CA](mailto:emb@uvic.CA)
- **Yalini Senathirajah**, Ph.D. Department of Medical Informatics, SUNY Downstate Medical Center, 450 Clarkson Ave, Room 7-048 Brooklyn New York, 11203-2012 Email: [Yalini.Senathirajah@downstate.edu](mailto:Yalini.Senathirajah@downstate.edu)
- **Darren Hudson**, M.D., M.Sc. Department of Critical Care Medicine, University of Alberta, 2-124 Clinical Sciences Building, 8440-112 Street, Edmonton, Alberta, Canada Email: [darrenandamyhudson@gmail.com](mailto:darrenandamyhudson@gmail.com)

**References**

Accelerating Evidence Into Practice: AHRQ’s Clinical Decision Support Initiative

Edwin A. Lomotan, M.D.1, Dave deBronkort, S.B.2, Barry Blumenfeld, M.D.3, Robert McCready, M.S.4
1Agency for Healthcare Research and Quality, Rockville, MD; 2 Society for Participatory Medicine, Nashua, NH; 3 RTI International, Research Triangle Park, NC; 4 The MITRE Corporation, Bedford, MA

Abstract
The Agency for Healthcare Research and Quality (AHRQ) has launched an ambitious multi-component initiative with two primary aims: to accelerate the movement of evidence into practice through clinical decision support (CDS) and to bring CDS closer to becoming more shareable, standards-based, and publicly-available. This panel will discuss progress thus far and visions for the future based on lessons learned. The initiative includes: 1) The Patient-Centered CDS Learning Network, which is building a community for engaging patients, clinicians, CDS developers and implementers, professional societies, and other stakeholders; 2) CDS Connect, which is building prototype infrastructure for sharing CDS, including a CDS authoring tool and a national CDS repository; 3) research grants to develop new CDS and to scale existing CDS to new sites and systems; and 4) evaluation. Audience members will learn how the Patient-Centered CDS Learning Network is engaging stakeholders to explore and advance the concept of “patient-centered CDS,” how to contribute and use shareable CDS artifacts through CDS Connect and its repository, about current grants and funding opportunities to move evidence into practice on a wider scale, and how to participate in the overall initiative.

Introduction
The time between evidence generation and incorporation into clinical practice continues to be problematic. Clinical decision support (CDS), as a component of quality improvement and when properly implemented, has the potential to accelerate the movement of evidence into practice. This process may be further expedited by tools and infrastructure that facilitate sharing of CDS among developers, implementers, and target audiences such as patients, clinicians, health care organizations, and others stakeholders dedicated to improving health care quality. In 2016, the Agency for Healthcare Research and Quality (AHRQ) launched an ambitious multi-component initiative, with two primary aims: 1) to accelerate the movement of evidence into practice through CDS and 2) to make CDS more shareable, health IT standards-based, and publicly-available.

Panel Description
This didactic panel will feature the leads of the initiative and provide a patient advocate’s perspective. Panelists will focus on progress thus far and visions for the future based on lessons learned. Additional information can be found at http://cds.ahrq.gov.

Edwin Lomotan, M.D., serves as Chief of Clinical Informatics for the Division of Health IT in the Center for Evidence and Practice Improvement at AHRQ. He will provide background for the CDS initiative, including its legislative mandate and requirements, its overall purpose and aims, and will describe the collaborative effort the initiative represents across several Federal agencies. Dr. Lomotan will also moderate the panel.

Barry Blumenfeld, M.D., is a Senior Physician Informaticist at RTI International and serves as the Principal Investigator for the Patient-Centered Clinical Decision Support (PCCDS) Learning Network. He will describe the Learning Network’s efforts thus far, including its Analytic Framework for Action developed through multi-stakeholder input, which places CDS into the context of a learning cycle addressing patient engagement, development, implementation, measurement, legal, governance, marketplace, and policy factors (see Figure 1). Dr. Blumenfeld will also describe the Learning Network’s efforts to advance the concept of “patient-centered CDS,” which extends CDS beyond its traditional focus on clinicians to include CDS-enabled, shared decision-making with patients.
Robert McCready is a Senior Principal Healthcare Systems Engineer at the MITRE Corporation and serves as the Project Lead for CDS Connect, the AHRQ-funded contractual effort to build prototype infrastructure for sharing CDS. Mr. McCready will describe a new national CDS repository, CDS Connect. CDS Connect allows CDS developers and stewards to contribute to a curated, online library of CDS artifacts (see Figure 2). CDS artifacts range from computer-interpretable, guideline recommendation statements encoded in the HL7 draft standard, Clinical Quality Language, to resources to guide CDS implementation, results of CDS testing, and links to primary research evidence. CDS Connect also allows CDS implementers to view, download, sort, and provide feedback on the library of CDS artifacts. Mr. McCready will describe a CDS authoring tool that aims to improve the efficiency and interoperability of CDS between systems and technologies. Finally, he will describe the development of CDS in the area of cholesterol management, chosen as the clinical domain for the first year of the contractual effort, as a proof-of-concept for developing standards-based CDS, testing CDS in a live clinical environment, and contributing to CDS Connect for public use of the artifacts.
Dave deBronkart, S.B., known on the internet as “e-Patient” Dave, is a leading evangelist for patients having active partnerships with their clinicians, both through sociological advancement (empowerment and role change) and by recognizing and leveraging new competencies that become available when patients connect with information and with peers on the internet. He will provide a patient advocate’s perspective. Mr. deBronkart serves on the Advisory Council for the PCCDS Learning Network and on a working group for CDS Connect.

Relevance to Annual Symposium Theme and Importance of the Topic

Our proposed panel fits directly into the theme of the AMIA 2017 Annual Symposium, “Precision Informatics for Health: The Right Informatics for the Right Person at the Right Time.” The right information for the right person at the right time are three of the “Five Rights of CDS,” which are woven throughout the AHRQ initiative. CDS Connect helps developers and implementers find what is relevant for health care providers and their patients. A national repository of publicly-available CDS artifacts has never before existed. Further, the PCCDS Learning Network has identified “patient-centered CDS” as its central focus. It has engaged a wide range of stakeholders – patients, clinicians, CDS and other health IT developers, Federal agencies, payers, and more – to explore what it means for CDS to be relevant to and incorporate patient preferences and for CDS to reflect shared decision-making with patients and their families.

As part of CDS Connect, MITRE has developed CDS in the area of cholesterol management. The CDS provides an assessment of cardiovascular risk and applicable recommendations, which can be viewed by both patients and clinicians in real-time. Going beyond simple alerts and reminders, the tool facilitates shared decision-making by enabling informed discussions to guide medication management and other treatment decisions that are reflective of patient preferences.

The concept of “precision informatics” depends on knowing what is precisely relevant to patients and making those preferences actionable. The AHRQ CDS initiative is exploring and developing tools and infrastructure to advance CDS forward in this regard. The anticipated audience will include patients (or patient advocates), CDS developers and implementers, other health IT developers (e.g., EHR vendors, vendors of population health platforms that may support CDS), Federal agencies, and clinicians.

Discussion Questions for Audience Participation

We anticipate the panel will generate discussion in a wide range of areas, from the technical aspects and standards for CDS development to further exploration with the audience how CDS Connect may serve their organizations and their patients. Discussions questions may include:

- What does “patient-centered CDS” mean to you?
- Would contributing artifacts to CDS Connect by you (or your organization) be valuable? What about using artifacts made available through CDS Connect? Why or why not?
- How might the artifacts and the authoring tool enable you to develop CDS more efficiently?
- How can the PCCDS Learning Network continue to build the CDS community? What do you see as pressing CDS issues that should be addressed moving forward?
- How can we incorporate feedback from both patients and clinicians into patient-centered CDS so that the CDS artifacts and their development continuously improve in a learning health system?

Author statement

All authors have agreed to take part in the panel.

References

A Spoonful of Structure Helps the Workload Go Down: Modeling Clinical Cognition in Inpatient Documentation Tools

Panelists: Subha Airan-Javia MD\textsuperscript{1,2}, Evan Orenstein, MD\textsuperscript{3}, Eric Shelov, MD MBI\textsuperscript{2}, Kai Zheng, PhD\textsuperscript{4}

Moderator: Mark V. Mai, MD MHS\textsuperscript{3}

\textsuperscript{1}University of Pennsylvania Health System, Philadelphia, PA; \textsuperscript{2}Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA; \textsuperscript{3}Department of Biomedical and Health Informatics, Children’s Hospital of Philadelphia, Philadelphia, PA; \textsuperscript{4}Department of Informatics, Donald Bren School of Information and Computer Sciences, University of California Irvine, Irvine, CA

Abstract

As inpatient care has grown in complexity and front-line clinicians (FLCs) must create and update a larger volume of information, inefficient and fragmented documentation systems put hospitalized patients at risk of harm from communication errors. In addition to writing and updating admission notes, progress notes, and discharge summaries, FLCs must also update plans of care and problem lists in their handoffs—often independently. Further, the information therein is generally locked in free-text, which cannot be easily reused for clinical decision support. With this duplicative workflow, FLCs struggle to maintain high quality notes and handoffs, and discrepancies between information documented between the two are known to occur. Problem-oriented documentation tools from electronic health record vendors have begun to address these issues by mandating coded problem names and allowing users to specify a block of information as historical. However, these tools lack the granularity required to optimize task management and handoffs. In this didactic panel, we will review the successes and failures of structured and flexible documentation systems, describe strengths and weaknesses of a vendor approach to problem-oriented documentation, and discuss novel models that organize clinical reasoning documentation into discrete elements that can be reused to improve clinical workflow.

Description

The tension between structured and free-text data in clinical documentation has limited the promise of the electronic health record (EHR). Information quality for secondary re-use is more complete when clinical data are structured at the time of data entry, however few structured documentation systems have demonstrated ongoing adoption and dissemination due to limited expressivity, user interface complexity, poor integration, and the differing goals of busy clinicians and administrators. Inpatient documentation represents a missed opportunity to leverage structured data for workflow enhancement. In addition to admission notes, progress notes, and discharge summaries, inpatient front-line clinicians must create and update written handoffs and problem lists to optimize patient outcomes. These tools must be independently updated, and the information therein is generally locked in free-text, which cannot be easily reused for clinical decision support. FLCs struggle to maintain high quality notes and handoffs, and discrepancies between information documented in the handoff and recorded in the electronic health record (EHR) are known to occur. Resolving this tension requires a sufficiently flexible system where clinicians can divine downstream workflow benefits from structuring data at the point of entry. In this didactic panel, we will review successes and failures of structured and flexible documentation systems, describe vendor approaches to problem-oriented documentation, and discuss novel models that allow users to tag clinical reasoning elements to improve clinical workflow.

Some early efforts to promote structured data entry demonstrated increased data quality and provider satisfaction for specific use cases such as SISCOPE for reporting gastrointestinal endoscopy. However, more generalized efforts ranging from the Structure-Meta Knowledge formalism to OpenSDE subsequently struggled to sustain adoption due to decreased efficiency in data entry and difficulty balancing limited expressivity with an excessive number of data elements. More recent approaches have recognized the importance of tolerating flexibility through structure-enhanced narratives and tightly integrating documentation with clinical workflow. The chronological order of studies on effective documentation strategies suggest a paradigm shift from an expert focus to end user focus, from a one-size-fits-all solution to customization, from a strong preference on structured data to a balance between structure and expressivity, and from individuals to teamwork.
In this environment, EHR vendors have begun to develop inpatient documentation strategies that explicitly connect the problem list with note documentation while allowing users to create individual problem plans through free text or note templates. Simultaneously, wiki-style free-text handoff sections allow for team collaboration, although these sections are often disconnected from other documentation tools. New methods may allow for some degree of parsing information that can be reused for different user-tasks. While this solution does not yet fulfill all user goals in inpatient documentation, it encourages the use of reference terminologies for the problem list and bypasses many obstacles of a third-party solution including integration, authentication, and the long-term need for local maintenance and associated institutional resources. We will discuss strategies for using vendor-based functionality to reduce inpatient documentation redundancy. Nonetheless, custom systems based on a more detailed model of clinical cognition may offer new strategies for minimizing redundancy and facilitating inpatient workflow.

At the University of Pennsylvania, analyses of communication failures and tasks during clinical handoffs led to the development of a home-grown, interdisciplinary handoff system. Carelign© is a responsive web application with a mobile interface being used in production at four of five hospitals within the University of Pennsylvania Health System. It utilizes an overarching, constantly changing care plan which can then be used for handoffs, progress note assessment and plans and hospital course summaries for discharge transitions. It pairs objective data with the qualitative handoff utilizing HL-7 messages, direct database queries, and web services/APIs. Carelign’s© care plan functionality uses discrete, collapsible data elements organized in a way that facilitates rounding, critical thinking, and list management. Key features include the ability to archive information, tag items as anticipatory guidance, arrange the care plan by problem priority or by system, view tasks for all problems by time due, create progress note content, and filter views based on provider role and context. For example, overnight cross-covering providers are able to limit their view to a one-liner, anticipatory guidance, and overnight tasks for the patients they are covering, with one touch access to drill down into more detailed information about the patient’s history and the ability to enter information from the same view that writes back to the overall care plan. The system handles almost 5000 unique weekly users with ~50,000 weekly sessions, and internal medicine residents reported significantly higher satisfaction and safer written handoffs with Carelign©, as well as a near-significant improvement in workflow efficiency.

At the Children’s Hospital of Philadelphia, cognitive engineering methods including artifact analysis and directed stakeholder interviews have informed the development of a markdown-based handoff application. Snap Chart (Signout to Note Assessment and Plan) leverages existing written handoff culture to give structure to free text at the point of data entry. It uses common annotations to identify data element types as the user is entering them, then parses that information into a data model representing elements of clinical reasoning (e.g. differential diagnosis, current therapy, anticipatory guidance, etc.). This model can then populate a note assessment and plan as well as workflow tools that can read and write to the data model such as a to-do list, an IPASS-compliant handoff, or a graphical user interface. By taking advantage of frequent habits, such as beginning a line of text with a “#” that will contain the name of a problem, the application suggests standardized terminology or autocompletes to existing patient data. This approach provides structure to unstructured text with the primary goal of streamlining the workflow of the providers responsible for data entry and upkeep. It also yields concepts that can drive clinical decision support and billing.

Learning Objectives

- Understand how the inpatient care plan is distributed across multiple documents and how data can be re-used to improve workflow and avoid inconsistencies.
- Review the successes and failures of structured and flexible documentation systems.
- Describe the strengths and shortcomings of vendor-based tools to address these problems.
- Discuss new approaches to documentation that structure impressions and plans into a data model that can be filtered into workflow tools optimized for provider role and context.

Description of panelist contributions

Mark Mai is a Pediatric Resident Physician in his 3rd year at The Children’s Hospital of Philadelphia and incoming Clinical Informatics Fellow. He is also a co-founder of Snap Chart, a markdown-based web application that allows clinicians to tag elements of their plan to facilitate rapid data entry, internally consistent inpatient documentation, and context-specific displays of information. He will serve as moderator of this panel.

Kai Zheng is Associate Professor in the Department of Informatics at the Donald Bren School of Information and Computer Sciences, Associate Adjunct Professor in the Department of Emergency Medicine, and Co-Director of the Center for Biomedical Informatics at the Institute for Clinical and Translational Science of the University of
California, Irvine. He will present a systematic review of the literature on successes and failures in structured and flexible clinical documentation and describe lessons for future work.

**Eric Shelov** is Associate CMIO, Medical Director for the Inpatient Electronic Health Record, and Attending Physician in the Division of General Pediatrics at The Children’s Hospital of Philadelphia. He will present the problem created by independent workflows for different inpatient documents and describe the strengths and shortcomings of a vendor-based solution.

**Subha Airan-Javia** is Associate CMIO at the University of Pennsylvania Health System, Director of Handoff and IT Education for the Internal Medicine residency program and Assistant Professor of Clinical Medicine at the Perelman School of Medicine. She is also the founder of Carelign©, a responsive web application with an intuitive presentation of real-time clinical data and an interactive, editable, electronic handoff with full mobile functionality. She will discuss how Carelign© organizes the care plan to reduce redundancies and improve communication.

**Evan Orenstein** is a Clinical Informatics Fellow at The Children’s Hospital of Philadelphia. He is also a co-founder of Snap Chart, a markdown-based web application that allows clinicians to tag elements of their plan to facilitate rapid data entry, internally consistent inpatient documentation, and context-specific displays of information. He will discuss the cognitive engineering methods that led to the data model underlying Snap Chart and discuss approaches to structuring free text using autocompletion tools at the point of data entry.

**Discussion of anticipated audience**

We believe this panel will be of interest to a wide variety of attendees including clinicians and trainees interested in streamlining documentation workflow, administrators and CMIOs looking to maximize the utility of clinical documentation and improve quality of care, and informatics researchers interested in increasing the proportion of structured data in the electronic health record.

**Expected discussion**

1. What are the challenges in integrating new approaches to documentation into vendor-based electronic health records and how can we avoid the creation of a “shadow chart”?
2. How should documentation interfaces be designed to optimally capture clinical reasoning and display easily digestible summaries for clinical users?
3. What are barriers to implementation of new documentation strategies in clinical practice?
4. How can we maximize structured data in the electronic health record for secondary consumption?

**Statement of participation:** All panelists have agreed to take part in this panel.

**References**

The PCORnet Learning Cycle

Keith Marsolo, PhD1, Bradley Hammill, DrPH2, Laura Qualls, MS2, Jeffrey Brown, PhD3, Lesley Curtis, PhD2
1Cincinnati Children’s Hospital Medical Center, Cincinnati, OH; 2Duke Clinical Research Institute, Durham, NC; 3Harvard Pilgrim Health Care Institute, Boston, MA;

Abstract

The National Patient-Centered Clinical Research Network (PCORnet) is a distributed network-of-networks, where partners transform data from electronic health record (EHR) and claims data sources into a Common Data Model. In order to ensure that the underlying data are suitable for use in network queries, PCORnet employs a robust data characterization process. The tools and processes that support PCORnet are not static, however. They are constantly being refined in response to stakeholder needs as well as the findings from analytic queries and characterization results. This panel will describe the Learning Cycle that informs the activities of the PCORnet Distributed Research Network Operations Center (DRN OC) and will highlight the relationship between the PCORnet CDM, the Data Characterization process, the findings of the PCORnet Demonstration Projects, and downstream analytical tool development. In addition, the panel will discuss how all of these findings are communicated out to network participants and other stakeholders. Attendees will gain an understanding of why these components are all necessary to the successful operation of a distributed research network, and recognize how similar Learning Cycles can be used to improve the quality and reproducibility of any network’s research.

General description of panel

Distributed research networks (DRNs) are critical components of the strategic roadmaps for agencies like the National Institutes of Health and Food and Drug Administration as they seek to move towards large-scale systems of evidence generation (1). The promise and sustainability of these systems hinges on the ability to extract usable, high-quality data from sources like the electronic health record (EHR), claims and others that are fit to support translational, interventional, and observational research initiatives. There are several active DRNs in the United States, including the Sentinel Initiative (formerly Mini-Sentinel) (2, 3), the Health Care Systems Research Network (HCSRN) (formerly the Health Maintenance Organization Research Network (HMORN))(4-6) and the National Patient-Centered Clinical Research Network (PCORnet) (7, 8). While these DRNs all have different operational characteristics, they share some commonalities. The first is that each network utilizes a common data model (CDM) to ensure that all partners (network participants) represent their data in a standardized manner. The second is that the data stay local, and the queries or analyses are distributed. The analytical tools differ by network, but are generally tailored to leverage the available data and to support the major use cases or study designs of interest. What they do have in common is that typically only the results, in the form of aggregate counts or summary statistics, are sent back to the requestor (9, 10). Therefore, before conducting any analysis, it is important to first understand the quality of the data that are to be queried to ensure that the results returned by each partner are valid and trustworthy (11). To that end, many DRNs have developed data characterization routines which are used to summarize missing values, outliers, and frequency distributions, and they also ask partners to complete surveys about the provenance of their data (3, 11). The characterization results then help inform an investigator’s decision about whether a dataset is suited to answering a given research question.

These activities do not occur in a silo, however, but inform one another. Data characterization routines may uncover variability in how network partners are populating certain elements of the network’s CDM. Study-specific analyses may uncover quality issues that are masked by a data characterization package that is designed to provide a broad assessment on data quality. Such findings can have a number of downstream impacts. For instance, variability in the population of the CDM may result in specific guidance to partners on how they should structure their extract-transform-load procedures and to the analytic teams on how they should structure their queries. Quality issues uncovered through study-specific characterization routines may result in changes to the base characterization process, or modifications to the network’s CDM in order to better represent the data in question. All of this then has an impact on the design and developmental roadmap of the analytical tools that support the network. In this manner, a DRN follows a learning cycle with a goal of continuous improvement, which can be linked to a number of metrics, such as overall data quality or query throughput. To that end, this panel will describe the Learning Cycle that informs the activities of the Distributed Research Network Operations Center (DRN OC) of the PCORnet Coordinating Center.
PCORnet is a distributed network of networks, comprised of 13 Clinical Data Research Networks (CDRNs) and 21 Patient-Powered Research Networks (PPRNs). PCORnet’s CDM network partners include health systems, ambulatory care clinics, and health plans, which contribute electronic health record data, billing system data, and other electronic health data such as clinical registries and health insurance claims. Altogether, these CDMs operate more than 80 individual datamarts, which contain data on tens of millions of patients standardized to the PCORnet CDM. The DRN OC of the PCORnet Coordinating Center is responsible for characterizing the datamart(s) of each network partner and evaluating the data’s fitness-for-use across a broad research portfolio through a foundational data characterization process. Data characterization is designed to complement any of the CDRNs’ internal data quality efforts while generating meaningful, actionable information for the CDRNs, Coordinating Center, sponsor, and other stakeholders. In particular, these results are used to inform the development and deployment of the analytical query tools that support the network, which includes both rapid, prep-to-research queries, as well as observational and comparative effectiveness studies. While the foundational data characterization process does not necessarily determine whether a given datamart is fit-for-use in a specific research study, the results inform study planning decisions, including what to include in the more comprehensive study-specific data characterization routines. These routines are executed prior to the main study analyses, and help assess the quality of the outcomes and variables of interest for the cohort in question.

Panelist presentations

Laura Qualls is a Project Leader at the Duke Clinical Research Institute (DCRI) and is responsible for the implementation of PCORnet’s foundational data characterization process. She will provide an overview of the current process and illustrate how it has evolved over time. In addition, Ms. Qualls will share selected results for the current data characterization cycle and explain how they are used to inform the query activities of PCORnet.

Keith Marsolo is an Associate Professor in the Division of Biomedical Informatics at Cincinnati Children’s Hospital Medical Center. He is a co-investigator in the PCORnet Coordinating Center’s DRN OC where he provides faculty oversight for the efforts related to data characterization and the PCORnet CDM. He will describe how the results of the initial data characterization cycle led to the creation of a document that provides implementation guidance to network partners as they populate the PCORnet CDM. Dr. Marsolo will also cover the process used to update this “Implementation Guidance” and how the material is disseminated and communicated to stakeholders.

Bradley Hammill is a Faculty Statistician at the DCRI. He is a co-investigator on the PCORnet Demonstration Study, Aspirin Dosing: A Patient-centric Trial Assessing Benefits and Long-Term Effectiveness (ADAPTABLE), a pragmatic clinical trial that is comparing the effectiveness of two different daily doses of aspirin to prevent heart attacks and stroke in patients with heart disease (12) as well as a study that transformed Medicare fee-for-service claims data into the PCORnet CDM and tested linkage rates between PCORnet patients and Medicare beneficiaries. Dr. Hammill will provide a high-level overview of these activities and describe how his findings led to changes in the PCORnet CDM specification and implementation guidance, as well as the foundational data characterization process.

Jeff Brown is an Associate Professor in the Department of Population Medicine at the Harvard Pilgrim Health Care Institute. He is a co-lead of the DRN OC and oversees the team responsible for the fulfillment of PCORnet queries and for the development and support of the PCORnet analytical query tools. He will provide an overview of the current suite of analytic tools and describe how the findings of data characterization and the initial PCORnet studies have influenced the development roadmap of future tools, as well as plans to increase the throughput of query activities, such as allowing query fulfillment activities to occur outside of the PCORnet Coordinating Center.

Lesley Curtis is a Professor in the Department of Medicine and Director of the Center for Population Health Sciences in the Duke University School of Medicine. She is a co-lead of the DRN OC along with Dr. Brown. Dr. Curtis will serve as moderator for the panel, setting the stage for the presentations of the other panelists by providing a high-level overview PCORnet, DRN OC activities and a summary of queries issued to-date. She will also discuss the learning cycle that underpins the work of the DRN OC and how it ties together the different workstreams.

Importance of topic

DRNs provide an important platform for the conduct of large-scale observational and comparative effectiveness research. Understanding the relationship between the CDM, data characterization and analytic tool utilization/development is important, as it can have an impact on the reproducibility of the research results and the ability of the network to meet the needs of its stakeholders, which is crucial for long-term sustainability. In addition, the general principles articulated in this panel have applicability beyond distributed research networks. Any project that seeks to use EHR data that have been transformed into a CDM will face similar issues. Raising awareness is the first step towards minimizing error and variability, ultimately leading to more high-quality research.
Target audience
Informatics professionals who generate EHR or CDM datasets; researchers who analyze EHR or CDM datasets.

Discussion questions
• Given the inter-relatedness between the CDM, data characterization and the use/development of analytical tools, how can findings be best communicated to all stakeholders?
• As the quality of the underlying data and sophistication of the query tools increases, should efforts be made to repeat earlier studies / queries?
• Can a ‘scorecard’ be created that describes the state of network/data at the time of analysis?
• How can these processes be translated to run outside of a DRN, so that they can be employed at a single institution?

Statement of participation
All panelists have agreed to take part in the panel.

References
Getting Hooked on CDS: Toward an Open Standard Architecture for Clinical Decision Support in Leading Electronic Medical Records

Blackford Middleton, MD, MPH, MSc¹, Aziz Boxwala, MD, PhD², J. Marc Overhage, MD, PhD³, James Doyle⁴, Mary Sung⁵

¹Apervita, Inc., Chicago, IL; ²Meliorix, Inc., San Diego; ³Cerner Corp., Kansas City, MO; ⁴Epic Systems, Inc., Verona, WI; ⁵athenahealth, Inc., Watertown, MA

Abstract

This interactive panel will discuss the current state-of-the-art with respect to implementing clinical decision support in commercial EMRs, the state of standards used in clinical decision support, and newer approaches to externalizing knowledge management and clinical decision support as web services consumed by an EMR. The discussion is timely given the explosive growth of new knowledge creation and the need for enhanced decision support for clinicians, and their patients, in the areas of precision medicine, care coordination, chronic disease management, and in the setting of rapidly changing clinical practice or acute disease outbreaks (e.g. Zika). The goals of this panel are to facilitate a discussion with leading EMR vendors, application builders, and CDS platform vendors to describe about their approach to CDS and the challenges above.

Significance

Clinical decision support is has been demonstrated to impact the quality, and cost of care when appropriately implemented and effectively used. While most US healthcare delivery organizations (HDOs) have successfully adopted electronic health record technologies due to the ARRA/HITECH stimulus bill (“Meaningful Use”), it is clear that the value proposition predicted by many has yet to be achieved. This may in part be due to challenges HDOs have in optimizing their EHR implementations – the process wherein clinical workflows, care communications, user interactions, and clinical knowledge for decision support is fine tuned to make it most beneficial to the end-user clinician, and positively impact the outcomes of the clinical encounter. Part of this optimization process – the part relating to clinical decision support -- is stymied by the complexity involved for some EHR systems to readily access and use knowledge-based tools and services external to the EHR. These may be in the form of standard knowledge representation formats that should theoretically allow import and use within an EMR, or external knowledge services which interact with an EHR via a standards-based web service to provide clinical decision support. In this interactive panel, one panelist will present a summary of the state-of-the-art with respect to standards for externalized CDS services, and four different perspectives will be provided by members of the vendor community: those who must design and deliver EHRs that can functionally access externalized CDS tools and services, and those that design and deliver such externalized CDS tools and services that must work with all EHRs.

Description of the Panel

The panel will be organized as follows:

<table>
<thead>
<tr>
<th>Time</th>
<th>Speaker</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>5 min</td>
<td>Middleton</td>
<td>Introductions, framing of the discussion</td>
</tr>
<tr>
<td>10 min</td>
<td>Boxwala</td>
<td>CDS standards, and applications– current state, challenges and opportunities</td>
</tr>
<tr>
<td>10 min</td>
<td>Overhage</td>
<td>Cerner CDS integration – current state, challenges and opportunities</td>
</tr>
<tr>
<td>10 min</td>
<td>Doyle</td>
<td>Epic CDS integration – current state, challenges and opportunities</td>
</tr>
<tr>
<td>10 min</td>
<td>Sung</td>
<td>athenahealth CDS integration – current state, challenges and opportunities</td>
</tr>
<tr>
<td>10 min</td>
<td>Middleton</td>
<td>Apervita PaaS for Delivering CDS, and Marketplace for CDS</td>
</tr>
<tr>
<td>35 min</td>
<td>All</td>
<td>Discussion, Q&amp;A with Audience</td>
</tr>
</tbody>
</table>
Learning Objectives:

1. Understand current approaches among leading EMR vendors to accessing and/or incorporating external knowledge-based tools and services in their EMR.
2. Understand current standards relating to the delivery of externalized knowledge-based tools and services.
3. Understand practical approaches to delivery of externalized knowledge-based tools and services to an EMR.

Individual Speaker Contributions

- **Blackford Middleton**: Dr. Middleton will provide a brief introduction for the panel participants, and framing of the discussion.
- **Aziz Boxwala**: Dr. Boxwala will provide a summary of the state-of-the-art with respect to standards for externalized CDS services. A brief synopsis of recent work including HL7 efforts in the FHIR “Clinical Reasoning Module”, and the “CDS-Hooks” workstreams will be highlighted. Assessment of the opportunity for a unified approach harmonizing these workstreams will be discussed, including current state of the effort. Highlights of his company’s efforts in building CDS tools and services using these standards will also be described.
- **James Doyle**: Mr. Doyle will provide a perspective on access and use of externalized CDS tools and services from the vantage point of a commercially successful EHR – Epic. Particular attention will be drawn to what requirements the Epic EHR may have that are unique, or limit the ability of using current standards, or a harmonized standard as outlined by Dr. Boxwala.
- **Marc Overhage**: Dr. Overhage will provide a perspective on access and use of externalized CDS tools and services from the vantage point of a commercially successful EHR – Cerner. Particular attention will be drawn to what requirements the Cerner EHR may have that are unique, or limit the ability of using current standards, or a harmonized standard as outlined by Dr. Boxwala.
- **Mary Sung**: Ms. Sung will provide a perspective on access and use of externalized CDS tools and services from the vantage point of a commercially successful EHR – athenahealth. Particular attention will be drawn to what requirements the athena EHR may have that are unique, or limit the ability of using current standards, or a harmonized standard as outlined by Dr. Boxwala.
- **Blackford Middleton**: Dr. Middleton will provide a perspective on providing standards-based externalized CDS tools and services from the vantage point of a commercial entity providing CDS in a ‘platform as a service’, and a marketplace for the exchange of knowledge artifacts, to multiple disparate EHRs.

Expected Discussion

Given the growing interest in clinical decision support in the evolution of clinical practice toward precision medicine and value-based care, we anticipate this will be a lively discussion about the various standards being developed and deployed for clinical decision support in EMRs. We anticipate questions form the audience regarding the methods by which clinical decision support may be deployed – hosted in EMR, vs. delivered by an externalized knowledge-based service. The moderator will facilitate discussion with each presenter, and sufficient time will be reserved for general audience Q&A.

Intended Audience

- **C*IOs**: clinicians in the role of CMIO, CNIO, and other, with responsibility and/or oversight for clinical decision support
- **CIOs**: responsible for the success of EHR implementations
- **CMO/CQO/CNO**: responsible for achieving high quality care, reduced unwarranted clinical variation, reduced patient harm
- **Physicians, Nurses, all EHR end-users**
- **IT Developers**: development staff building EHR and clinical decision support tools and services
- **IT Implementers**: implementation staff responsible for implementation and effective use of EHR tools, and system optimization
- **Knowledge vendors** supplying content for CDS services
Attestation

The Panel Moderator has assurances from all participants that they will be available to participate in the panel at AMIA’17.
Panel: Clinical NLP in Languages Other Than English

Aurélie Névéol1, PhD, Noémie Elhadad2, PhD, Sumithra Velupillai3, PhD, Hua Xu4, PhD, Guergana Savova5, PhD

1 CNRS, LIMSI, Université Paris Saclay, F-91405 Orsay
2 Columbia University, New York, NY, USA
3 School of Computer Science and Communication, KTH, Sweden
4 The University of Texas Health Science Center at Houston UTHealth, Houston, TX, USA
5 Children's Hospital Boston and Harvard Medical School, Boston, Massachusetts, USA

Abstract

Natural Language Processing (NLP) of clinical free-text has been an active area of research in the past decades. Clinical documents are routinely created across health care providing institutions and are generally written in the official language(s) of the country these institutions are located in. As a result, free-text clinical information is written in a large variety of languages. While many efforts for clinical NLP have focused on English, there is a steady interest to extend this work to other languages in order to gain medical information about patient cohorts in geographical areas where English is not an official language. Furthermore, adapting current NLP methods developed for English to other languages may provide useful insight on the generalizability of algorithms and lead to increased robustness. This panel aims to provide an overview of clinical NLP for a wide variety of languages including Germanic, Romance, Sino-Tibetan and Afro-Asiatic languages. We will also discuss initiatives to support clinical NLP research and sharing of clinical resources for NLP in multiple languages.

General Description of the Panel

The goal of this didactic panel is to engage the medical informatics and clinical Natural Language Processing community in a discussion about ways to advance research through languages other than English. We will report on the current state of clinical NLP in a wide variety of languages including Germanic, Romance, Sino-Tibetan and Afro-Asiatic languages. We seek to provide the audience with insight on how to develop a clinical NLP system for a language other than English. We identify strategies to leverage existing resources and tasks in English and other languages to address new challenges. Finally, we identify the best opportunities for a contribution to the field and propose that new resources be integrated in a state-of-the-art repository.

Clinical NLP in languages other than English.

Natural Language Processing (NLP) of clinical free-text has been an active area of research in the past decades, demonstrating its potential to analyze large quantities of documents rapidly and accurately (Demner-Fushman and Elhadad, 2016). The ability to analyze clinical text in languages other than English allows to unlock medical data concerning cohorts of patients who are treated in countries where English is not the official language. In this context, data extracted from clinical texts in languages other than English adds another dimension to data aggregation. Further, with international initiatives like OHDSI (Observational Health Data Sciences and Informatics) which brings together clinical repositories from more than 60 countries under the same data model (Hripcsak et al, 2016), there is an unprecedented opportunity to foster robust NLP research across the globe.

We follow-up on a panel addressing this topic at the AMIA Fall Symposium in 2014. We widen the scope of the languages covered in the panel to include non Indo-European languages such as Chinese and Hebrew. As the importance of clinical NLP gains recognition, clinical corpora become available to researchers in languages other than English, prompting work that sometimes builds on methods validated for English. Adapting systems that work well for English to another language is a difficult task that may be carried out with varying level of success depending on the task and language (Grouin et al., 2009; Velupillai et al. 2014; Täckström et al., 2012; Tourille et al. 2017). These experiments prompt a reflection on how to carry out clinical NLP in a more global context: should methods be developed for one language and then ported to other languages? Can the source language method benefit from the porting? Can algorithms be more robust if they are initially designed with a multi-language perspective?

North Germanic languages, although closely related to English, exhibit properties that require adaptation and specific solutions for accurate processing of clinical narratives. In recent years, approaches have been developed for
e.g. named entity recognition in Swedish (Skeppstedt et al. 2014, Weegar et al. 2016), dictionary construction for adverse drug event detection in Danish (Eriksson et al. 2013), within several large-scale national and cross-country research initiatives in north Europe.

**Romance languages** benefit from a large coverage in the UMLS. Automatic de-identification is becoming quite advanced for French (Grouin & Névéol, 2013), paving the way for targeted clinical information extraction tasks (Pham et al. 2014). Notable work on Spanish has addressed the analysis of social media and clinical records for pharmacovigilance (Oronoz et al. 2015). For Portuguese and Italian, approaches rely on rule-based or unsupervised methods to avert the greater lack of domain resources in these languages (Silva et al. 2013; Alicante et al. 2016).

In Israel, there is an emerging need for processing the clinical notes for point of care decision support, as well as for supporting biomedical discovery. There are no established Hebrew medical terminologies however. As such, a lot of the work in Hebrew medical NLP is on language adaptation for such resources through transliteration of medical concepts and linking to existing English terminologies (Cohen et al, 2011; Cohen and Elhadad, 2013). Further, effective word segmentation (a critical task for Hebrew) has been shown to improve information extraction from clinical notes (Cohen et al, 2010).

**Chinese clinical text** has been accumulated rapidly due to wide implementation of EHRs in China. Recently researchers have developed various approaches for diverse NLP tasks on Chinese clinical text, such as named entity recognition (Lei et al. 2014, Wu et al. 2015), speculation detection (Zhang et al, 2016), negation detection (Kang et al. 2017), and temporal information extraction (Zhou et al 2011).

**Panelists**

**Dr. Noémie Elhadad** (Associate Professor at Columbia University) will present work in schemas for NLP in the international initiative OHDSI as well as research directions in processing of medical texts in Hebrew. She led the OHDSI clinical NLP data modeling effort, and is co-founder of the hNLP Center with Profs. Palmer and Savova.

**Dr. Sumithra Velupillai** (Postdoctoral Research Fellow at KTH, Sweden) will present work on North Germanic languages including Danish, Norwegian and Swedish. Dr. Velupillai has worked on Swedish clinical narrative data and been involved in several past and ongoing Nordic research collaborations in this research area.

**Dr. Aurélie Névéol** (Staff Scientist at LIMSI-CNRS, France) will present work on Romance languages including French, Italian, Portuguese and Spanish. Dr. Névéol has been leading a project addressing the automatic understanding of French clinical narratives for translational research.

**Dr. Guergana Savova** (Associate Professor at Harvard Medical School) will act as a moderator and will present the hNLP center for health data sharing, which she co-founded. Dr. Savova has been leading the development of the core Clinical Text Analysis and Knowledge Extraction System (cTAKES; ctakes.apache.org).

**Dr. Hua Xu** (Professor at UTHealth) will present work on Chinese medical language processing. Dr. Xu has conducted several projects on named entity recognition from Chinese clinical text using supervised machine learning and deep learning approaches.

**List of Discussion Points**

After the introductory presentations, the moderator will ask questions as well as solicit questions from the audience, to prompt discussion among the panelists. Potential topics and questions include:

- What are your needs for clinical NLP methods, tools and resources in a language other than English?
- To fill existing gaps, are there alternatives to replicating what is currently done for English?
- How can we share data across languages for (1) methodological developments, (2) translational studies?

**References**


Statement of Participation
The first author affirms that all panel participants have agreed to participate and have contributed to the preparation of this document (as of March 9, 2017)
Panel Sponsored by Evaluation and People & Organizational Issues Working Groups

Innovation in Workflow Methods for Consumer Health Informatics

Mustafa Ozkaynak, PhD1 (organizer), Rupa Valdez, PhD2, George Demiris, PhD3, Laurie Novak, PhD4, Yong Choi, MPH3, Charlene Weir, RN, PhD5 (moderator)

1University of Colorado, Aurora, CO, USA; 2University of Virginia, Charlottesville, VA, USA; 3University of Washington, Seattle, WA, USA; 4Vanderbilt University Medical Center, Nashville, TN, USA; 5University of Utah, Salt Lake City, UT, USA

Abstract

Consumer health information technologies are becoming an essential part of wellness and management of chronic conditions. The literature on health informatics emphasizes that workflow studies are key to the design and implementation of health information technologies in clinical settings. Consumer health informatics interventions should also be informed by workflow in daily living settings, however, additional challenges exist such as, privacy, variability in living styles, preferences and daily routines. Workflow methods that have been utilized in clinical settings should be thoughtfully adapted to daily living settings. Compared to traditional approaches in institutional health care settings, novel methods may be needed in the home that account for the unstructured and personal nature of daily living settings. The success of workflow studies depend on the robustness of the methods used to capture, analyze and re-design workflows. This panel will begin with a discussion by each panelist of their experience in workflow research for consumer health informatics interventions and their use of innovative methods. This will be followed with an interactive session, where the audience and panelists will create a list of challenges. Finally, the audience and panelists will discuss effective strategies to address these challenges.

Intended Audience

The intended audience for this panel includes all AMIA attendees with various levels of experience, interested in designing, implementing and evaluating consumer health informatics interventions. Designers, implementers and researchers of these interventions can benefit from this session, by understanding the current state and challenges of workflow methods, and shaping the development of future methods.

Introduction of the Topic

Consumer health information technologies have become an essential part of wellness and management of chronic conditions. Design and implementation however, requires a fit between technology characteristics and the context of the patient’s environment, as well as their abilities and preferences regarding self-management practices. The importance of understanding workflow in clinical settings has been emphasized in the informatics literature. Workflow studies reveal established routines as well as variations that should be accommodated by informatics interventions. Understanding workflow in daily-living (home and community) settings is required to ensure that consumer health informatics technologies do not disrupt the day-to-day life of the individual in unintended ways or confuse patients, yet take full advantage of their potential in terms of supporting self-management. Workflow in daily-living settings represents temporal organization of health-related activities performed by patients and their proxies. There are at least three reasons for innovative workflow research in these settings: (1) Traditional methods are insufficient in capturing high levels of variability; (2) Daily living settings are private environments and intrusive methods may not be appropriate; (3) The context (e.g. physical, organizational, social and cultural) for each patient is unique, and should be studied concurrently with their health-related activities.

Aim of Discussion

The aim of the discussion is to present the state of workflow research in daily-living settings and identify major weaknesses for the purpose of producing insights for future directions. To gain the broadest insight, we will enlist the exchange of ideas and experiences with both the expert panel and audience.

Timeliness of the Topic

This panel is timely because daily-living settings are becoming critical environments to consider in health management. Increasing interest in consumer health informatics by the public, researchers and policy makers does
not necessarily translate into better health outcomes. The current state of workflow research and its associated, limitations in daily-living settings, can be one of the roadblocks. This panel aims to bring panelists and the audience from various disciplines, to address and mitigated this roadblock. Because of its interactive nature, this panel will serve as a forum in which members of the audience and panelists can collaboratively advance the dialogue at the intersection of workflow research and consumer health informatics. The audience will have the opportunity to learn from the panelists and vice versa about experiences and new developments at this intersection. The sharing of experiences stimulated by this panel will serve as a foundation for generating and prioritizing future design and research initiatives informed by effective workflow studies that utilize innovative methodologies.

**Contribution of Each Panelist**

**Mustafa Ozkaynak:** Traditional methods of capturing workflow, present challenges in our understanding of health-related activities in the daily living environment of individuals. To address this challenge, we designed a multi-method qualitative study to better understand self-management of chronic disease. Data were collected through patient interviews (pre & post journaling), the journals and provider interviews. Journaling was accomplished by tablet computers that included software that allowed for voice entry. The qualitative analysis allowed us to develop workflow diagrams for self-management, examine the impact of context, identify challenges and facilitators of self-management as well as gaps between clinical settings (where the therapy plan was developed) and daily living settings (where the therapy plan was implemented). Our analysis also provided insights on the design, implementation and evaluation of consumer health informatics interventions in terms of needed functional requirements. Dr. Ozkaynak will discuss the effectiveness and limitations of our design in terms of richness of data, how data were incorporated from different sources, and concurrent analysis of health-related activities and the embedded context. The discussion will also include strategies to establish validity and reliability in studying workflow in this setting.

**Rupa Valdez:** To help guide design consumer health, IT designers should be informed by an understanding of patient’s existing health information management workflows. Additionally, redesign efforts should be guided by an understanding of the challenges patients face using health IT solutions within the context of existing workflows. As a whole, health information management is comprised of multiple workflows related to information-seeking, information organization, and information communication. The latter of these, although recognized as burdensome for patients, has received little formal attention from consumer health IT designers. The purpose of this research was to explicate health information communication workflows, to serve as a foundation for consumer health IT design for one patient population – individuals with disabilities. This study focused on workflow for both design and redesign efforts. In the first part of the study, interviews were conducted with individuals living with physical, cognitive and sensory disabilities to understand current practices of communicating health information to members of their social network. Emphasis was placed on determining to whom, how, why, and what health information was communicated in addition to when and where this communication occurred. In the second part of the study, task analyses and journaling exercises were used to determine how this workflow was impacted when participants were asked to use three existing consumer health IT apps for health information communication (Microsoft HealthVault, Epic’s MyChart, and CaringBridge). Task analyses took place in a controlled setting, while journaling took place in participants’ daily living settings. In this talk Dr. Valdez will discuss how these methods were triangulated to understand workflow both with and without a specific consumer health IT intervention and the challenges and lessons learned when applying these methods with a range of individuals with disabilities.

**George Demiris:** The emergence of Internet of Things (IoT) technologies whereby everyday household objects become interconnected via the Internet and able to send and receive data, is starting to define new opportunities for home-based monitoring of patients who aim to maintain independence and safety at home. Family caregivers who are essential to the delivery of home health and rehabilitation services for loved ones, are called to carry out various caregiving tasks and often oversee complex medication administration regimens and other care tasks and processes. We demonstrate how the use of IoT technologies in the home enables family caregivers to better understand patterns of daily living for their loved one and accordingly adjust their caregiving workflow. IoT tools have the potential to unintentionally, disrupt and redesign the caregiving workflow. Dr. Demiris will discuss practical, methodological, ethical and other challenges and opportunities associated with the introduction of this emerging consumer informatics tool.

**Laurie Novak:** Adherence to recommended therapy and self-care practices improves clinical outcomes in diabetes, yet carrying out doctors’ recommendations remains a challenge for patients. An emerging science of “patient work”
complements existing adherence research by using established frameworks and methods for studying activities of workers (i.e. in business/industrial settings) to study the work of patients and caregivers, and how that work is organized and made reliable. Understanding patient workflow can produce opportunities for innovation in consumer health informatics design. In the study presented, the primary research questions were: 1) How are diabetes self-care routines aligned with other activities in everyday life?; and 2) What is the relationship between strongly aligned routines and diabetes self-care success?

To answer these questions, we documented from the perspective of 50 patients with diabetes and their caregivers, the actors, actions and structures for each of four primary routines: 1) diet and nutrition, 2) medication and insulin management, 3) physical activity, and 4) physiological monitoring. Examples of structures include artifacts, events, places, people, cultural norms, and other factors that facilitate or impede adherence. The study also involved the application of a novel method to score the routines’ level of embeddedness in other, everyday life routines. Findings will be presented, along with a discussion of the implications for consumer health informatics design.

Yong Choi: Residential settings with embedded technologies to facilitate passive monitoring of residents, also known as smart homes, have the potential to support aging in place and allow for detection of trends and delivery of early interventions for older adults who live alone and want to maximize the quality of their life. Such systems are based mostly on various sensors that capture data related to activities of daily living and also environmental parameters such as humidity and temperature. Mr. Choi will describe an ongoing smart home initiative that utilizes smart home sensors installed in the homes of adults over the age of 65 years who live alone. The vast data generated by this initiative needs to be summarized and presented effectively, in order to provide added value to the stakeholders (namely older adults, clinicians, and family). Mr. Choi will showcase various visualization approaches and present conceptual scenarios to demonstrate how the visualization of behavioral and environmental sensing data affect the decision making workflow for both clinicians and consumers.

Interactive Session

After the introductory presentations the panel and audience will engage in a discussion of (1) Challenges of studying workflow for consumer informatics interventions and (2) Strategies to overcome these challenges, facilitated by the moderator. The moderator will identify rules of discussion; establish a safe environment for participants, and summarize the exchange when appropriate. Prioritized lists of challenges and strategies will be developed for research and design domains. The moderator will engage the audience, stimulate discussion, and maintain a balance between audience and panelists. If time permits, the panel and audience will also engage in a discussion of workflow considerations for linking consumer informatics interventions to clinical informatics interventions.

All panelists have agreed to take part on this panel.

References

Engaging Patients with Health Technologies to Improve Quality of Care and to Reduce Preventable Harm

Wanda Pratt, Ph.D.¹, Patricia Dykes, Ph.D.², Ryan Greyson, M.D.³, Cornelia Ruland, R.N., Ph.D.⁴,⁵, David Bates, M.D.²

¹University of Washington, Seattle, WA; ²Brigham & Women’s Hospital, Harvard Medical School, Boston, MA; ³Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA; ⁴Oslo University Hospital & ⁵University of Oslo, Oslo Norway

Abstract

With recent studies highlighting the pervasiveness and severity of preventable medical errors, renewed efforts have turned to health technologies to improve quality of care and patient safety. However, few efforts target patients and caregivers as primary users of quality and safety-oriented technologies. In this panel, we will describe four new efforts in developing health technologies to engage patients and their caregivers to improve quality of care and to reduce preventable harm. These efforts include co-designing with both a pediatric and adult population, using sensors to promote safe and healthy mobility, engaging hospitalized patients and caregivers in the safety plan, and using gamification techniques to engage patients in chronic illness management. Attendees will learn new approaches for engaging patients and their caregivers in ensuring they receive high-quality, safe care while they are in the hospital. The panel will provide insights that information technology designers, developers, and researchers can build upon to address caregivers’ and patients’ safety needs, while realistically fitting within the workflow of an organization.

Introduction

Although many efforts have called attention to the prevalence of medical errors, recent studies have shown the problem to be much more widespread than previously thought. James’ analysis estimated that 440,000 deaths a year in the U.S. alone are caused by preventable medical errors and 10-20 times that number receive serious harm.¹ In a more recent study, Makary and Daniel found medical errors to be the third leading cause of death.²

Many efforts have used technology to reduce this harm from these medical errors, but few inpatient efforts engage patients and their caregivers. Although patients having access to health information is positively associated with their ability and willingness to get involved with their own safety³, in a hospital setting, medical personnel and system processes receive most of the attention for preventing harm while patients receive little support. Patients and their caregivers offer a distinctive perspective and set of skills that could help reduce preventable harm if they are supported through appropriate technologies. Enhancing patients’ and caregivers’ access to and engagement with meaningful information can help address communication challenges that are often the root cause of many undesirable events. In this panel, we will present recent efforts to design and develop information technology to engage patients and caregivers to prevent harmful errors. Moreover, we will highlight integration of bio-behavioral and positive psychology approaches into technology using gamification⁴ to assist patients with activating their personal strengths as a means to enhance healthy behaviors, motivate positive change, and improve health outcomes.

Learning Objectives

• The attendee will be able to summarize the challenges related to development and use of consumer-facing technologies to engage patients in the quest for high quality and safe care.
• The attendee will describe challenges and opportunities associated with integration of innovative technologies into the clinical workflow of an organization.
• The attendee will describe strategies to improve patient engagement related to high quality and safe care using consumer facing health technologies.

Moderator: David Bates, M.D., M.Sc.

Professor of Medicine, Harvard Medical School, Professor of Health Policy and Management, Harvard School of Public Health, Boston, MA, USA. David Bates is a Professor of Medicine at Harvard Medical School, and a Professor of Health Policy and Management at the Harvard School of Public Health, where he co-directs the Program in Clinical Effectiveness. He directs the Center for Patient Safety Research and Practice at Brigham and
Women’s Hospital, and serves as external program lead for research in the World Health Organization’s Global Alliance for Patient Safety. He is the former board chair of AMIA, former North American representative to IMIA, and the immediate past president of the International Society for Quality in Healthcare (ISQua).

Dr. Bates will provide a brief overview and scope of the problem of preventable medical harm and the promise of health information technology to reduce the problem. He will introduce each panelist and moderate the question answer session.

**Panelist: Wanda Pratt, Ph.D.**

Dr. Wanda Pratt is a Professor in the Information School with an adjunct appointment in Biomedical & Health Informatics in the Medical School at the University of Washington. Dr. Pratt serves on the board of directors for AMIA, and chaired the 2016 AMIA Annual Symposium. Her work in consumer health informatics has received multiple best paper awards from AMIA, the ACM CHI Conference on Human Factors in Computing Systems, and the Journal of the American Society of Information Science & Technology. She is the principal investigator of the Patients as Safeguards Project—a project funded by the Agency for Healthcare Research and Quality (AHRQ) to understand the information needs of and design technology for hospitalized patients’ and their caregivers to prevent, detect, and recover from medical errors.

Dr. Pratt will describe results from co-design efforts with pediatric patients, adult patients, caregivers, and clinicians. She will present highlights from multiple co-design sessions and discuss technology functionalities that participants thought were necessary to support patients and caregivers in their important safeguarding role. Dr. Pratt will include design examples, and summarize how they are building upon those designs to develop in-hospital technologies for patients and caregivers.

**Panelist: Patricia Dykes, Ph.D., M.A., R.N.,**

Dr. Patricia Dykes is research program director in the Center for Patient Safety Research and Practice at Brigham and Women’s Hospital and associate professor at Harvard Medical School where she has a program of informatics and patient safety research. She is past Chair of the AMIA Nursing Informatics Working Group and on the Board of Governors for the Alliance for Nursing Informatics. Dr. Dykes’ program of research includes using health information technology to provide the tools, knowledge, and decision support for patient activation in their recovery plan during an acute hospitalization.

Dr. Dykes will review a suite of tools designed developed through an AHRQ-funded patient safety learning lab to engage hospitalized patients and family in their safety plan. She will review the benefits and challenges of providing bedside access to personal health and safety information through the use of mobile technology and clinical decision support to promote patient activation during an acute hospitalization. Dr. Dykes will review the patient engagement value of the technology and preliminary results evaluating its effectiveness.

**Panelist: Ryan Greysen, M.H.S., M.A., M.D.**

Dr. Ryan Greysen is Chief of Hospital Medicine in the Division of General Internal Medicine at University of Pennsylvania. Dr. Greysen’s quality, safety, and research efforts focus on improving hospital and transition care for vulnerable patients, particularly older adults with multiple morbidities. His program of research focuses on patient-centered stressors such as poor functional status, mobility, sleep, and engagement in care and the potential for personalized mobile technologies to help alleviate these stressors and increase patient and caregiver engagement in acute and recovery care.

In his panel presentation, Dr. Greysen will talk about his research using mobile sensors with goal of reducing falls and other complications including hospital acquired pressure ulcers and hospital acquired infections. He will describe on ongoing pilot program that “prescribes” mobility goals to patients in the hospital. In addition, Dr. Greysen will discuss the benefits and challenges of linking the sensor data with Epic MyChart interface.

**Panelist: Cornelia Ruland, R.N., Ph.D.**

Dr. Ruland is professor and past Department Head of the Center for Shared Decision Making and Collaborative Care Research at Oslo University Hospital in Oslo Norway where she initiated and directs an innovative, externally-funded program of research focusing on user-centered design and evaluation of informatics tools to improve shared decision making, illness management, self-management and patient-centered collaborative care. These tools have shown through randomized controlled trials to be highly effective in improving health outcomes. As a PI of
numerous international research studies, Dr. Ruland has widely disseminated her research and received substantial external recognition, including several best paper awards from AMIA.

Dr. Ruland will highlight her current project that uses gamification to support patients in chronic illness management. Dr. Ruland’s presentation will explore how information and communication technology can be used to help and engage patients in discovery of their personal strengths and provide support and guidance on how their strengths can be used in more efficient health management. She will explore how gamification mechanisms can enhance user engagement, motivation and positive health behavior, but also to foster reflection, self-discovery and capacity building as part of the self-management process. Including patients’ personal strengths into clinical consultations aligns patient care with people’s personal values and preferences and thus is an important factor for patient safety.

**Discussion Questions**

- What are some common and unique features from presented projects/participant projects that serve to engage patients and their care partners while hospitalized in improving the quality and safety of healthcare?
- What are key operational and implementation challenges that were encountered? From the patient/caregiver perspective? From the provider perspective? How are clinical, technical, and operational workflows reconciled?
- How does is technology used to enhance the experience of patients and family caregivers while engaging them in high quality and safe practices? How is information shared across patient and provider platforms to ensure efficiency?

**Summary**

In summary, this interactive panel will engage participants in a range of key topics to share lessons learned from their work related to the design, develop, and implement health technologies to engage patients and caregivers in improving the quality and safety of healthcare and to identify additional lesson learned from participants.

**Statement of the panel organizer**

All panelists have agreed to take part in this important panel.

**References**

2. Makary M & Daniel M. Medical error—the third leading cause of death in the US. BMJ 2016;353;i2139
Latin American e-Health Programs and WHO/PAHO Planning Toolkit

Yuri Quintana, PhD\textsuperscript{1}, David Novillo-Ortiz, PhD\textsuperscript{2}, PhD, Heimar de F. Marin, RN, PhD\textsuperscript{3}, Marcelo A. Logetegui, MD, MS\textsuperscript{4}, Mario R. Cubillo, MBA\textsuperscript{5}, Paula Otero, MD\textsuperscript{6}

\textsuperscript{1}Harvard University, Boston, MA, USA, State; \textsuperscript{2}Pan American Health Organization, Washington, DC, USA, \textsuperscript{3}Alumni Professor, Federal University of São Paulo, Brazil, \textsuperscript{4}Clinica Alemana de Santiago, Santiago, Chile, \textsuperscript{5}Hospital San Vicente de Paul de Heredia, Heredia, Costa Rica, Hospital \textsuperscript{6}Italiano de Buenos Aires, Buenos Aires, Argentina

Abstract

The goal of this panel is to provide an overview major initiatives in Latin America, provide an overview of the World Health Organization (WHO) national e-health strategic planning toolkit, challenges in implementation of e-health systems, overview of the PAHO e-health strategy for the Americas, and provide approaches for implementing programs using cooperative models.

Learning Objectives

1. Identify challenges for the development of e-health in the Latin American region
2. Describe the priority areas for e-health development and informatics training
3. Discuss challenges in deployment and support of informatics systems in clinical and public health sectors

Panel Description

This goal of this panel is to provide an overview major initiatives in Latin America, provide an overview of the Pan American Health Organization/World Health Organization (PAHO/WHO) national e-health strategic planning toolkit, challenges in implementation of e-health systems, overview of the PAHO e-health strategy for the Americas, and provide approaches for implementing programs using cooperative models.

Key Topics

1. Present examples of e-health Initiatives and strategies in Latin America
2. Provide overview of the PAHO/WHO National e-health strategic planning toolkit
3. Discuss models of public-private cooperation for global health challenges
4. Discuss strategies for managing non-communicable diseases
5. Discuss human-centric approaches to healthcare services
6. Discuss ways for managing the risks and impact of current and future epidemics

Speakers

- Session Moderator Yuri Quintana, PhD
- The PAHO/WHO National e-health strategic planning toolkit, WHO Third Global Survey on eHealth (Region of the Americas), and e-health Initiatives and strategies in Latin America, David Novillo, PhD
- E-Health Strategies in Brazil, Heimar Marin, RN, PhD
- Digital Strategy at Clinica Alemana, Santiago Chile, Marcelo Lopetegui, MD, MS
- Digital Strategies in Costa Rica and RACSEL Red Americana de Cooperación para el desarrollo de la Salud Electronica, Mario Ruiz Cubillo, MD
- Clinical Informatics Training Programs, Paula Otero, MD,

Why this topic is important

- Latin America has a growing e-health sector and academic informatics programs
- There synergies and common experiences that can be shared among these programs
- There is an opportunity for AMIA members to be more involved in Latin America via informal networks and formal collaborations
Questions

- What are the priorities in each country and institution?
- What process was used to define the priorities?
- What role can national and international informatics organizations can have in the training and development of informatics in the region?
- What opportunities are there for AMIA members to collaborate in Latin America?

All panelists have agreed to participate.
Design and Implementation of a Structured Sequencing Report Format: A Multi-Stakeholder Perspective from eMERGE

Luke V. Rasmussen, MS\textsuperscript{1}, Darren C. Ames, MS\textsuperscript{2}, Samuel J. Aronson, ALM, MA\textsuperscript{3}, Lawrence J. Babb\textsuperscript{4}, Casey L. Overby, PhD\textsuperscript{5,6}

\textsuperscript{1}Northwestern University Feinberg School of Medicine, Chicago, IL; \textsuperscript{2}DNAnexus, Inc., Mountain View, CA; \textsuperscript{3}Partners HealthCare, Cambridge, MA; \textsuperscript{4}GeneInsight a Sunquest Company, Boston, MA; \textsuperscript{5}Johns Hopkins University, Baltimore, MD; \textsuperscript{6}Geisinger Health System, Danville, PA

Abstract

With continued interest in adopting genomic medicine practices, an ongoing barrier has been around enabling the transmission of genomic laboratory results in a structured manner that allows a receiving institution to perform local computations on those results, including those for clinical decision support. The electronic Medical Records and Genomics (eMERGE) network, now in its third phase, has implemented a structured representation of results for a custom sequencing panel of over 100 genes and 1500 SNPs. This panel will describe the selection of the structured report format, the experiences of two clinical sequencing centers in aligning their laboratory systems to the format, how the format is received by eMERGE institutions, and the lessons to date that have potential relevance to other institutions exploring the integration of structured genomic results.

Panel Overview

In recent years, several research consortia have been exploring the use of genomic results as part of routine clinical practice. These groups, including the NHGRI-sponsored electronic Medical Records and Genomics (eMERGE) network\textsuperscript{1,2}, Implementing GeNomics In pracTicE network\textsuperscript{3} and the Clinical Sequencing Exploratory Research Consortium (CSER)\textsuperscript{4}, have implemented and evaluated novel methods by which genomic results may be better incorporated into existing electronic health record (EHR) systems for pharmacogenomics and genomic medicine. Lessons learned have included considerations for implementing structured results to facilitate clinical decision support (CDS), storage and handling of the results, and strategies for reporting results to patients and healthcare providers.

While much has been learned, there are still many unanswered questions. The eMERGE network, for example, is now in its third phase (eMERGE 3) with nine institutions that will receive results from one of two central sequencing and genotyping centers (CSGs; Baylor College of Medicine and Partners HealthCare with Broad Institute). Results from each of the CSGs will include a customized sequencing panel (eMERGE-Seq) of over 100 genes and 1500 single nucleotide polymorphisms (SNPs) developed for this study. Samples and results are managed in accordance with the Clinical Laboratory Improvement Amendments (CLIA), and include formatted reports (represented as PDF documents) as well as structured representations of the information contained in the reports.

Given the unique infrastructure at each sequencing center, as well as the heterogeneity of EHRs at each of the nine participating institutions, a major consideration has been how to facilitate collaboration across network sites with respect to methodology and technology. A critical underpinning of this has been to achieve consensus on a format for the structured representation of sequencing results. While several candidate formats exist – including SMART on FHIR Genomics\textsuperscript{5} and HL7 Clinical Sequencing\textsuperscript{6} – the involvement of Partner HealthCare, which developed the GeneInsight platform\textsuperscript{7} prompted consideration of the underlying XML format used in GeneInsight as an alternative. The GeneInsight XML format was reviewed by the CSGs and eMERGE network sites, and was accepted with some additional constraints applied on elements. A computable definition of the format has been made available online at: https://github.com/emerge-ehri/results-schema.

While the selected XML format is not a ratified standard (although it has been shared with the HL7 Clinical Genomics Workgroup), and is currently used by GeneInsight users and now at eMERGE network institutions, we believe the experiences in implementing this format across the network will provide generalizable lessons of interest and importance to other institutions. This panel (as described in more detail in the Panelists section) includes a variety of stakeholders from the project, representing perspectives from project leadership, laboratory operations, technical implementation and end-users of the sequencing results. A broad range of AMIA attendees that identify with these groups should find interest in the panel, and will garner real-world experiences of two laboratories aligning to a common standard, considerations for the technical implementation and workflow of these labs when
delivering data for both discovery science and clinical care, as well as the experiences of institutions receiving and processing structured results for return to providers and patients, and for use in CDS. Although a particular structured format will be described, the discussion will focus on the importance of structured results regardless of underlying standard, allowing attendees to apply these lessons at their own institutions for whichever structured representation is available.

Panelists:
Casey Overby – Dr. Overby will act as the moderator for the panel, and as co-chair of the eMERGE EHR Integration workgroup will provide a brief overview of the current objectives of the eMERGE network with respect to EHR integration. This will frame the rest of the panelists as they describe their work within the context of these objectives.

Larry Babb – Mr. Babb will describe the rationale and design behind the original GeneInsight XML format, and the considerations that were made for extending it for the eMERGE project. Given his involvement in HL7, Mr. Babb will also discuss how the efforts within eMERGE align towards standards for clinical sequencing.

Daren Ames – Dr. Ames will discuss the role of DNAnexus® and Baylor College of Medicine (BCM) within eMERGE, and the work performed to align DNAnexus and BCM infrastructure with the selected XML format. Topics will include considerations for mapping existing results to the structured report template, and automation of delivery for both identified and de-identified results.

Sandy Aronson – Mr. Aronson will describe his work as co-chair of the eMERGE EHRI Working Group (with Dr. Overby) in aligning the sequencing labs and participating institutions in the design, review and implementation of the selected XML format. Mr. Aronson will describe considerations from his viewpoint as a project facilitator, IT Director of a laboratory producing results, as well as a consumer of the results at Partners HealthCare.

Luke Rasmussen – Mr. Rasmussen will provide the perspective of a consumer of the selected XML format, describing early experiences in review of the proposed format, through preliminary testing and current implementation work at Northwestern University. This will also include approaches taken across the network to share tools and methods for integrating results across disparate EHR systems.

Discussion Questions
The following discussion questions will be used to initiate further audience discussion following the panelist presentations:
1. Given that standards for clinical sequencing exist, but the reality that not all labs may not currently return results in that format, what seems to be the minimal set of information that should be transmitted?
2. What still needs to be addressed to make structured genomic results more accessible to institutions not involved with consortia such as eMERGE?
3. Given the heterogeneity of EHR systems and the complexities of sequencing results, are structured results worth the hassle to both generate and consume?
4. How do changes in the interpretation of a particular variant factor into the implementation decisions made within eMERGE, and what broader considerations are there for other institutions?

Disclosures
Mr. Aronson works for Partners HealthCare, which receives a royalty on sales of GeneInsight.
Mr. Babb is employed by GeneInsight, a Sunquest Information Systems company, which reports grants from eMERGE III in conjunction with Partners Healthcare as well as independently.

Confirmation of Agreement
The panel organizer (LVR) confirms that all listed participants have agreed to take part in this panel.

References


20 Years Of Digital Pathology-An Overview Of The Road Travelled And What Is On The Horizon

Joel Saltz, PhD, MD¹, Ashish Sharma, PhD², Alexis B. Carter, MD, FCAP, FASCP³, Liron Pantanowitz, MD⁴, Tahsin Kurc, PhD¹

¹Biomedical Informatics Department, Stony Brook University, Stony Brook, NY; ²Biomedical Informatics Department, Emory University, Atlanta, GA; ³Pathology and Laboratory Medicine, Children’s Healthcare of Atlanta, Atlanta, GA; ⁴Pathology and Bioinformatics, University of Pittsburgh Medical Center, Pittsburgh, PA

Abstract

Over the past 20 years, Digital Pathology has developed into a thriving field with both clinical and research applications. FDA approval is anticipated this year to allow use of digital whole slide imaging for primary Pathology diagnosis. Once this occurs, we expect digital Pathology to become as ubiquitous in clinical practice as digital Radiology. While the concept of the Virtual Microscope, a software tool that implements the operation of a whole slide microscope and enables viewing of large digitized tissue samples, is now two decades old, the adoption of Digital Pathology informatics tools in clinical practice is still a work in progress. This panel will describe the history, state of the art, promise and challenges in digital Pathology Informatics. Informatics related challenges include the need for interoperability and standards, integration of analytic pipelines and results into clinical decision support.

Introduction

The true promise of digital pathology lies in precision medicine – leveraging image-derived information to generate Pathology Imaging biomarkers that can be combined with Radiology information and molecular data to steer treatment. The creation of meaningful ways to validate, classify, display and explore these results, integrated into the entire clinical picture will be a crucial Pathology- and Informatics-based contribution to precision medicine. An approach that employs survival curves, patient demographics, genomic data is critical to allow this information to be used to craft individualized patient care plans and to evaluate response to treatment and is not trivial given the very big data sets involved.

Building upon successful work showing that Pathology imaging biomarkers predict survival in Glioblastoma and can be used to sub-classify tumor types, the NCI has funded the development of tools for lung cancer which discover reproducible pathology imaging biomarkers that have prognostic significance. By interrogating large data sets of multiple tumor types, it is expected that certain Pathology imaging biomarkers will be predictive within a tumor type and others across multiple tumor types.

This panel will discuss past and present efforts to develop infrastructure to scan, catalog and store large collections of whole slide images (WSI), both for clinical work and research. Methods and tools to carry out large-scale, integrated analysis with visualization, interrogation and mining of WSI will also be discussed. The panel participants will describe example of institutions and consortia that launched into large scale digital Pathology data acquisition efforts as well as efforts to integrate images and annotations into the Electronic Medical Record through Vendor-Neutral Archives. The panel will also speak on how challenges in digital pathology relate to challenges in other fields of biomedical informatics and how work in digital pathology can enhance and benefit from work in other biomedical informatics fields.

Panelist Presentations

Joel Saltz (Organizer and Participant) will give a lightning overview of the last 20 years of digital Pathology beginning with 1) early speculations about digital Pathology and its telepathology predecessors, 2) initial prototypes, 3) current enterprise level and open source digital software systems, and 4) recent advances in research and development of methods and tools presented in Pathology and imaging conferences and journals. He will address informatics issues around leveraging digital Pathology for clinical decision support, health care quality and precision medicine. He will then describe how digital Pathology will become a ubiquitous component of clinical workflow.
due to imminent FDA approval for primary diagnostic use of whole slide imaging and how Pathology Informatics will become a pivotal component in all areas of biomedical informatics.

Alexis Carter (Participant) will describe the advantages and barriers of Vendor Neutral Archives (VNAs) both within and outside Pathology for storage and retrieval of medically important Whole Slide Images. She will present the potential power of integrating VNA whole slide images with Genomic, Radiologic and other clinical data for research and discovery.

Liron Pantanowitz (Participant) will describe how digital pathology has evolved over time as a research and clinical tool. He will present the current regulatory trajectory in the US and the impact of changes in regulations and FDA approval on the broader applications of digital pathology in research and clinical settings.

Ashish Sharma (Participant) will present challenges and software systems for whole slide tissue image management, visualization and analysis. He will describe the challenges of whole slide tissue image (WSI) management, as the sizes of image datasets scale to federated archives of population-scale image collections. We will present the challenges of processing WSIs at large scales and integrating analysis results with other imaging, molecular, and clinical datasets. He will then go on to discuss the role of visual analytics in the development of robust pathology imaging biomarkers.

Tahsin Kurc will serve as the panel moderator.

**Timeliness and Significance of Panel**

FDA approval is anticipated this year to allow use of digital whole slide imaging for primary Pathology diagnosis. FDA approval and introduction of digital Pathology into the clinical workflow will mark the advent of digital Pathology as a sub-area of clinical informatics, and will allow the integration of analytics from the research arena to clinical practice.

The first Virtual Microscope cost about $300K to set up was one-of-a-kind, and took over 12 hours to scan a single slide. Today, 20 years later, WSI devices are important tools in supporting diagnosis, education and scientific discovery in Pathology. The next 20 years promises to see this tool at the center of Anatomic Pathology through its ability to collect and present objective, reproducible and accurate data derived from WSI, thus enhancing Pathologists’ analyses and patient care.

By developing tools to leverage digital pathology, beyond routine clinical applications, a deeper understanding of tumor pathophysiology will be possible. Integrating this information into the landscape of the entire spectrum of clinical information will drive both disease specific and patient specific information which can be used to drive high risk high reward cancer trials to better results faster. To bring this vision to reality it will be necessary to develop and deploy infrastructure to scan, catalog and store extremely large collections of WSI, both for clinical work and research. We will describe a number of examples of institutions and consortia that launched into very large scale digital Pathology data acquisition, storage and integration efforts. Integration is the key to correlate WSI data with clinical, radiologic and genomic information for clinical care and discovery.

Today and going forward WSI analytics will be powered by artificial intelligence, feature extraction and powerful algorithms to allow reproducible and quantifiable results, such as assessment of the percentage of tumor-infiltrating lymphocytes, important in assessing the body’s immune response to cancer, or to produce a virtual flow cytometry indexing and cataloging cell types present in a WSI. Beyond what is visible to the human eye, image analysis can produce powerful Pathology imaging biomarkers, which can be integrated with genomics, survival curves, patient demographics, public health data, environmental data and census data to drive personalized cancer diagnosis, prognosis, and therapy and also to better understand and predict response to treatment. This pool of personalized data, when scaled up, collected and linked with cancer registry data will provide a radically detailed view of cancer trends and has the potential of significantly impacting treatment regimens and the broader policy around the NCI Cancer Moonshot Program.

**Discussion Questions**

- How will the ubiquitous use of whole slide images impact clinical decision making?
- How can ubiquitous availability of whole slide imaging be employed to improve health care quality?
• In what way can whole slide image information complement information obtained from Pathology reports?
• Should patients have access to their Pathology images? If so, what kind of explanation/education should we provide for patients?
• How can whole slide images be used to better target therapy?
• Crucial decisions hinge on Pathology interpretations. Availability of images in electronic health records will allow non-Pathologists to easily examine the data that goes into Pathologist interpretations. Will this democratize decision-making?
• How can deep learning and artificial intelligence be used to create new Pathology classifications that may do a better job of targeting therapy?
• How should ontologies and controlled vocabularies be leveraged to describe Pathology image features?
• How will regulatory approval for deep learning (AI) with WSI proceed in the USA?
• Can collections of population-scale pathology archives affect our understanding of disease and treatment efficacy?

**Statement from Panel Organizer**

All participants have agreed to take part on the panel.
Evaluating Impacts of Patient-Facing eHealth Technologies: Connected Care in the Veterans Health Administration

Stephanie L. Shimada, PhD1,2,3, Timothy P. Hogan, PhD1,3, Thomas K. Houston, MD MPH1,3, D. Keith McInnes, ScD MSc1,2, Neil C. Evans, MD4,5
1Center for Healthcare Organization & Implementation Research, Department of Veterans Affairs, Bedford, Massachusetts, USA; 2Department of Health Law, Policy, and Management, Boston University School of Public Health, Boston, Massachusetts, USA; 3Division of Health Informatics and Implementation Science, University of Massachusetts Medical School, Worcester, Massachusetts, USA; 4Office of Connected Care, Department of Veterans Affairs, Washington, DC, USA; 5Washington DC VA Medical Center, Department of Veterans Affairs, Washington, DC, USA

Abstract
Patient-facing eHealth technologies such as patient portals, mHealth applications, home telehealth devices, and automated texting systems are increasingly common. Rigorously evaluating the impacts of these technologies is thus increasingly important. These ‘impacts’ can include both the intended positive effects on access, patient engagement, and health outcomes as well as other unintended adverse consequences. In this panel, we will present new data on the impacts of a range of patient-facing technologies being implemented across the Veterans Health Administration (VHA), including the My HealtheVet patient portal, Secure Messaging, and VHA’s automated, tailored text-messaging system. Attendees will learn about current and upcoming patient-facing technology implementation in the VHA, approaches for gauging different stakeholder experiences, and to select appropriate measures for positive and negative impacts of patient-facing technology. Implications for future evaluations will also be discussed.

The panel will be moderated by Dr. Neil Evans, Chief Officer for the Office of Connected Care for the Veterans Health Administration. We will present findings from a randomized trial of an intervention to encourage adoption of secure messaging, an evaluation of the implementation of an automated texting system to support patient self-management, and a multi-modal study to identify and categorize unintended consequences of multiple patient portal features.

Introduction
Engaging and supporting patients through technology has been a national priority for over a decade and is actively encouraged through policy initiatives, including meaningful use requirements focused on access to and use of patient portals and secure messaging. Systematic reviews in the area of Consumer Health Informatics have reported on patient and provider attitudes toward these technologies and on adoption and utilization. Many opinion papers also cite the potential of these technologies, but fewer rigorous studies have assessed impacts on specific outcomes, including patient access, disease outcomes, or costs. Further, most research to date has focused on the expected positive benefits of patient-facing eHealth technologies.

VHA’s Office of Connected Care is responsible for the implementation and evaluation of patient-facing virtual eHealth technologies across the VHA healthcare system. We consider eHealth as a model for the delivery and receipt of healthcare services that has great potential to improve a variety of outcomes, including patient access to healthcare services, care coordination, and the broader transition from episodic to continuous care. In the eHealth model, patients, their families, and clinical team members use patient-facing eHealth technologies that support functions (i.e., communication, behavior support, transactions) that are critical to disease prevention and health management, which, in turn, influence behaviors, processes, and outcomes. In an effort to successfully fulfill its mission, the Office of Connected Care has invested in the eHealth Partnered Evaluation Initiative, an evaluation center that draws upon the theories, strategies, and methods of implementation science, to support the further integration of patient-facing eHealth technologies into VHA care.

Our panel encompasses recent and ongoing work from investigators affiliated with the VA eHealth Partnered Evaluation Initiative and will offer a balanced perspective, demonstrating positive impacts but also presenting
important unintended consequences. The VHA Offices of Connected Care and Health Services Research & Development Service both support this research with the goal of enhancing healthcare quality and assuring patient safety. The results will guide future VHA implementation efforts, educate patients, guide providers, and ultimately enhance these technologies. Our panel is timely as adoption of patient-facing technologies is experiencing exponential growth. Much work remains to be done to fully understand the implementation of these technologies, and we anticipate that our presentations and the subsequent discussion will be of interest to researchers in Consumer Health Informatics and to those in leadership positions who are committed to the careful evaluation of patient-facing technologies in their respective healthcare organizations.

An Overview of Connected Care in the VA (Dr. Neil Evans)
We will describe the mission of the Office of Connected Care and proceed to review some of the key patient-facing technologies currently being rolled out by VHA. These include, but are not limited to, the My HealtheVet patient portal with integrated patient-to-clinical-team secure messaging, the Annie automated text messaging system, VHA’s expanding suite of mHealth applications, and telehealth programs that support real-time video-audio interactions among patients and members of their care team. Recognizing the speed with which eHealth technologies evolve and the implications of broader organizational changes for technology adoption, this review will emphasize key functions – education and information sharing, communication, behavior support, and transactions with the healthcare system – that VHA’s patient-facing eHealth technologies are intended to support. Our focus on technologies and the functions they support will serve as a framework for the subsequent research and evaluation project presentations.

Supported Adoption Intervention for Secure Messaging (Dr. Thomas K. Houston)
We will present findings from a randomized trial of a supported adoption intervention for Secure Messaging. Within the My HealtheVet patient portal, Secure Messaging is a way for Veterans to asynchronously communicate requests for information and requests for action to their healthcare teams. My HealtheVet users from three VA Medical Centers who had not previously sent Secure Messages to their clinical teams (N = 1,196) were randomized to be encouraged to use Secure Messaging, or to receive no intervention. For those randomized to intervention, we established a supported adoption intervention (SAI) curriculum that included motivational messages, reminders, and educational instruction about how to use Secure Messaging. Those in the SAI arm received the curriculum in a stepped approach, beginning with educational mailings and proactive Secure Messages from the facility informing patients of the ease and benefits of Secure Message use, the steps needed to send a Secure Message, and numbers to call for assistance. Next, those SAI patients not yet engaged in Secure Messaging after the educational mailings and Secure Messages received a motivational telephone call during which the Veteran was asked to think about how they could personally leverage Secure Messaging for their own healthcare, and during which research assistant trained in motivational interviewing techniques could troubleshoot any potential barriers with the patient.

Demographic characteristics, age, gender, socio-economic status, were balanced by randomization across the two groups. Among those randomized to the intervention, 14.3% began using Secure Messaging during the six-month follow-up, compared with only 4.6% of those Veterans randomized to control. Among those patients randomized to SAI who subsequently did send a Secure Message, 44% sent the first message prior to the motivational call, and 56% sent the first message after the motivational call. In an intent-to-treat analysis, Veterans randomized to the intervention were more likely to report that it was easy or very easy to communicate with their doctor when they needed to (65.6%), compared with the Veterans randomized to no intervention (58.5%, p = 0.052). This difference was driven by the subset of intervention Veterans who began using Secure Messaging, 79% of whom reported it was easy to communicate with their doctor. Intervention patients also reported higher scores on the Health Care Climate Questionnaire (p = 0.01 compared with control).

Veterans in the SAI group had near-triple the rate of adoption of Secure Messaging, and reported higher rates of perceived access and self-reported health care climate. Veterans reported that the SAI reminders and education about when to use Secure Messaging with their team influenced their use of the system.
Automated Patient Text Messaging System – Formative Evaluation to Assist with Wider Adoption
(Dr. Timothy P. Hogan and Dr. D. Keith McInnes)

The “Annie” text messaging system was developed in the VHA based on a similar system that was implemented in the United Kingdom’s National Health Service. Annie uses baseline patient profiles and expert-written rules, called “protocols”, to deliver targeted text messages and elicit responses from individual patients. We present results of a formative evaluation based on 48 qualitative interviews with stakeholders including patients, clinicians, and clinical staff from 4 VHA Medical Centers where the Annie system was being tested.

Patients reported feeling more accountable for their behaviors, and perceiving Annie as a kind of coach to assist them with chronic disease management. Providers appreciated the ability to help patients continue to be engaged with care between office visits, and utilized the customization features to have the appropriate kinds and volume of text messages sent to patients. Our interviews revealed various barriers to (e.g. challenges fitting into clinical workflow) and facilitators of (e.g. quality improvement culture in the adopting clinic or hospital) implementing the Annie text messaging system which in turn have implications for understanding both the intended and unintended consequences of the system for different stakeholder groups. We will present the results of this evaluation and extrapolate broader lessons learned that can inform the implementation of similar automated texting systems as well as other patient-facing eHealth technologies.

Unintended Consequences of eHealth Implementation (Dr. Stephanie Shimada)

Too often, health informatics studies are designed only to find differences in the hypothesized positive effects of that technology, ignoring the greater context and potential unintended adverse consequences which may have far-reaching impacts such as on patient-provider relationships, clinical workload and workflow, and even patient safety. This narrow window of evaluation has had important implementation consequences.

We sought to develop a taxonomy of unintended consequences of patient-facing technology use, by focusing on the unintended consequences of My Health eVet implementation in VA. We gathered data on actual and potential positive and negative consequences related to patient portal use from the literature, from prior studies of patient portal use in VA, and an in-depth qualitative evaluation with key VA stakeholders including My Health eVet operational staff, providers, and patients. We will present the results of this evaluation, including both negative and positive unintended consequences of patient portal use, discuss some of the factors involved in unintended adverse consequences of patient-facing eHealth implementation, and make recommendations for including consideration of unintended consequences in evaluations of other patient-facing technologies moving forward.

Discussion Questions to Enhance Audience Participation

Thinking of your role as a patient, clinician, or informatician, how do the presented results relate to your own experiences with patient-facing technologies?

If you are a clinician, how are you leveraging patient-facing technologies with your patients? If you are not, what are some of the barriers to doing so?

What are the best ways you have found to incorporate the patient voice into your research along with other stakeholder input?

What are other appropriate measures for evaluating patient-facing technologies? What data sources can or should we leverage to gather the data necessary for accurate and timely measurement of impacts?

What types of unintended consequences related to patient-facing technologies (portals, telehealth, mobile apps, etc.) have you encountered in your work or research?

The moderator and all presenters have agreed to take part in the panel.

This work is supported by the U.S. Department of Veterans Affairs, Grants # PPO 15-395, PEC 15-470, IIR 15-101. The views expressed are those of the authors and do not necessarily reflect the position or policy of the Department of Veterans Affairs or the United States government.
Interactive Panel: Participatory Health Informatics for Data-Driven Precision Medicine

Jaideep Srivastava\textsuperscript{a}, Luis Fernandez-Luque\textsuperscript{b}, Fernando Martin-Sanchez\textsuperscript{c}, Kenneth D. Mandl\textsuperscript{d}

\textsuperscript{a} Computer Science Department, University of Minnesota, Minneapolis, MN
\textsuperscript{b} Qatar Computing Research Institute, Hamad Bin Kalipha University, Qatar Foundation, Doha, Qatar
\textsuperscript{c} Department of Healthcare Policy and Research, Division of Health Informatics, Weill Cornell Medicine, New York, NY, USA
\textsuperscript{d} Computational Health Informatics Program, Boston Children’s Hospital, Boston, Massachusetts

Abstract

Lifestyle plays a major role in the prevention and management of chronic conditions. Sleep, physical activity and nutrition behaviours are being found to play a major inter-related role in physical and mental wellness of all of us. Worldwide empowered citizens are using new technologies such as social media, mobile apps and wearables to track their health and also get support from digital health services and peers. In that process, big and heterogeneous data sources are being generated. This panel will explore how data generated by empowered patients can facilitate the development of precision medicine. The panelist will address, among others topics, a) the use of social media to monitor lifestyles, b) data analytics of health wearables, c) the linkage between Genotype/Phenotype/Expotype data, and d) integration with other biomedical data sources.. Using an interactive approach with the audience the panel will aim to identify key elements for the development of a roadmap for research and practice.

Keywords: Social media; wearables, participatory health; precision medicine

1. Introduction

Precision Medicine is foreseen as a drastical transformation of our healthcare systems. This transformation is based on the availability of new individual health data that can be used to personalize health prevention and management in an unforeseen manner. Particularly promising is the integration of personal health data, such as lifestyles, which is often necessary to understand the health context of the patients in their real life. This integration encompasses data sources ranging from environmental data, such as sleep or physical activity (expotypes), to genomics (genotypes) and clinical data (or derived phenotypes). Further, the data are highly heterogeneous and spread among many different stakeholders. How to “link” personal health data with more “traditional” biomedical data is one of the major challenges ahead to achieve the full potential of precision medicine\textsuperscript{1}, and some initiatives are emerging to reduce those data-gaps\textsuperscript{2}.

Health consumers are increasingly becoming more proactive (empowered) stakeholders, which is drastically transforming the personal health data ecosystem. Empowered patients and participatory medicine as a movement are catalyzing an explosion of person-generated health data, often as part of the quantified-self movement\textsuperscript{3}. Currently, wellness and fitness devices are becoming massively adopted. If we consider participatory technologies, such as smartphone apps, Direct-To-Consumer services and wearable devices as sources of individual health data for potential clinical use we have to acknowledge that person-generated data is still largely an under-utilized health resource.

The use of these type of data for precision medicine raises many technical, legal and practical questions (e.g. privacy, liability, ownership). Yet alone, the integration and linkage with clinical data open many other questions such as the creation of new data and patient-driven health information economical models. A machine learning and data-driven ecosystem also brings many ethical questions, and some opponents of machine learning are even suggesting that “Artificial Intelligence is the White Guy Problem” as “smart” models are built inherently biased due to the type of data used by the machine learning algorithms. This panel is aiming at discussing those aspects and starting collaborative work to draft a possible roadmap for research and practice in this area.

2. Learning and Discussion Objectives

After participating in this session, the audience should be better able: 1) to analyze and understand complex socio-ethical challenges arising in the use of patient-generated data for precision medicine, 2) to study technical challenges and potential precision medicine solutions for the integration of personal health data in precision medicine applications, 3) to define strategies for deploying data-driven solutions based on personal health data.

Further, this panel is aiming to create debate and promote discussion on controversial aspects, which are currently hot topics in both research and practice. The audience is expected to engage in this dialogue to provide their insights. Among others, a key element for the discussion will be the effect of data-driven precision medicine in participatory health (i.e. patient empowerment, doctor-patient communication) and vice versa.

3. Panel Description

This panel provides a ground for discussing how Participatory Medicine, and particularly patient empowerment, can play a role in the development of data-driven precision medicine. This panel brings together members of AMIA, the Society of Participatory Medicine and two IMIA Working Groups (Big Data Mining and Analytics and Social Media). The panelists have extensive experience in various facets of social media, wearable and mobile health, participatory health informatics, data mining, and the analysis of biomedical data for health research.

4. Panel Characteristics

**Intended audience:** Patient advocates, healthcare professionals, data scientist, and medical informatics community at large.

**Aim of the discussion:** To create an inclusive roadmap for the research and development agenda of precision medicine based on participatory health informatics.

**Interactivity:** the panel presentations will be shared in Slideshare (with license creative commons) prior to the conference and distributed among working groups mailing lists. We will seek comments with regards of questions for debate using a Google Form.

During the conference we will engage with the online and offline audience. The presentation will be shared using Facebook Live Video and Twitter using the hashtag of AMIA 2017. The debate questions raised in the panel will be put under vote using a real-time voting system (e.g. Poll everywhere). These polls will be also used for getting information on the background and experience of the participants. The polls will distinguish between online and in-situ participants.

After the panel the audience will be invited to continue the debate offline aiming at developing a white paper on the topic of the panel. This will follow the example of a previous experience. This paper will be circulated among related AMIA/IMIA Working Groups. Funds for Open Access of the paper will be seek from the Qatar Foundation Open Access Fund.

---

6 Pre-selected tool for polling https://www.polleverywhere.com/
5. Panel Organization and Participants

Moderator: Dr. Luis Fernandez-Luque is co-Chair of IMIA Social Media Working Group and Scientist at Qatar Computing Research Institute and member of the Society for Participatory Medicine. He is also co-founder of the digital health startup Salu media which develops mobile applications for empowering patients with cancer.

1) Introduction to the Panel and Topics (15 minutes)
by Luis Fernandez-Luque (Scientist at Qatar Computing Research Institute)

He will initially provide a brief 15-min overview of the panel’s aims, format, structure and tools for participation. This introduction will provide a background and help to frame each individual panel member’s presentation and examples. The following questions will be used as a guide for discussion:

- What advantages/benefits does participatory medicine offer to the development of precision medicine?
- What are the challenges and negatives aspects that may arise to achieve the full potential of precision medicine?
- Which are the socio-ethical challenges ahead?

He will then introduce each panel member and thereafter, each will provide a short presentation (10 minutes each) of a select topic relevant to the discussion.

2) Data Mining of Personal Health Data: Leading and Misleading? (10 minutes)
By Prof. Jaideep Srivastava – Computer Science Department, University of Minnesota

- Short bio: Jaideep Srivastava is Professor at the University of Minnesota with over 30 years of experience researching in data mining, both in academia and industry.
- Panel Focus: He will focus on the technical challenges ahead, in special the role of deep learning and wearable data as study case. He will also explain cases in which machine learning can lead to erroneous and dangerous results.

3) Breaking the Silos of Digital Health Data for Precision Medicine (10 minutes)
By Prof Kenneth Mandl MD, MPH FACMI– Department of Biomedical Informatics, Harvard Medical School, Harvard University

- Short bio: Prof. Kenneth D. Mandl directs the Computational Health Informatics Program at Boston Children’s Hospital.
- Panel focus: Prof. Mandl will focus on how new approaches to interoperability and innovation in healthcare, including of the SMART on FHIR apps ecosystem can facilitate the integration of personal health data from consumer health applications with other biomedical data sources.

4) Precision Medicine requires linking Genotype/Phenotype and Expotype Data (10 minutes)
By Fernando Martin PhD FACMI, Professor of Healthcare Policy and Research, Weill Cornell Medical College, Weill Cornell in NYC.

- Short bio: Prof Martin-Sanchez is Co-chair of the IMIA Big Data Working Group, member of the Society for Participatory Medicine and a researcher in the All of Us Research Program (US Precision Medicine Initiative). His current research focuses on the application of informatics methods in Participatory Health and Precision Medicine.
- Panel focus: He will cover aspects related to methods for linking Genotype/Phenotype/Expotype data. He will also open the discussion on socio ethical issues.

5) Debate and Roadmap Discussion (45 minutes)
In this part of the panel the moderator will unveil the different questions polled during the session. These will be addressed using the three major areas for discussion highlighted in the introduction.

The audience will be encourage to think on a timeframe of 5-10 years in the future for the discussion to identify major aspects with long term impact. Further, the audience will be encouraged to suggest additional topics for discussion.

5. Statement of the Panel Organizer

All panel members agree to take part in the panel.
Advancing electronic case reporting (eCR) to enable public health disease control and emergency response: getting into the technical weeds!

Catherine J Staes, BSN, MPH, PhD1, John W. Loonsk, MD FACMI2, Kathy Turner, PhD, MPH3, Noam Arzt, PhD, FHIMSS4, Patina Zarcone-Gagne, MPH5

1University of Utah, Salt Lake City, UT; 2CGI Federal, Fairfax, VA; 3Idaho Department of Health and Welfare, Boise, ID; 4HLN Consulting, Palm Springs, CA; 5Association of Public Health Laboratories, Silver Spring, MD

Abstract

Many healthcare institutions grapple with the legally required mandate to report information to public health authorities when patients present with conditions of public health importance, such as pertussis, salmonella, Zika virus infection. Existing reporting of communicable diseases by health care providers is incomplete, untimely and predominantly a manual, paper-based process. Electronic case reporting (eCR) is relevant for over 200 conditions reportable in over 3,000 public health jurisdictions in the US. As such, eCR requires an ecosystem (knowledge, decision support, infrastructure, and standards) that can meet unique jurisdictional and clinical needs. The panelists represent leaders in the public health and informatics community involved in building the platform that supports standards-based services and defining the eCR-specific HL7 standards for reporting and response. The panel objectives are to a) describe the need for a scalable, flexible strategy for eCR that meets both routine and emergency reporting needs; b) describe the decision support services, infrastructure, and HL7 document standards that support bi-directional communication for eCR; and c) explore the strengths and limitations of the current eCR strategy. The Zika infection case reporting scenario will be used to illustrate key findings, and experience from pilot testing during 2017 will be shared.

Background

Tracking and responding to common and emerging public health threats (e.g., gonorrhea, measles, Zika infection, lead poisoning) requires communication between clinical settings and public health authorities that conduct surveillance and control efforts. Currently in most US communities, case reporting by health care providers is dependent on manual processes (phone and fax) and as a result, reporting is incomplete and untimely1. Providers are often not aware of reporting requirements.2 Electronic laboratory reporting has significantly improved the completeness of case finding for those conditions for which a laboratory result is necessary, however, there are other conditions for which laboratory results are not timely enough for case detection or not appropriate or available, and laboratory information is only part of the information required for initiating public health investigations. Due to the complexities around case reporting, many healthcare institutions grapple with the legal mandate to report required information to public health.

Electronic Case Reporting (eCR) is enabled by the widespread adoption of electronic health records (EHRs),3,4 and the emerging technical and standards-based ecosystem that supports bi-directional communication between public health and healthcare systems. Such an ecosystem must support routine eCR and be able to scale up during public health emergencies. Establishing new communication channels in the midst of an outbreak can be difficult and disruptive. The AMIA informatics community has, and can, contribute significantly to the implementation of the nationwide scalable, and standards-based eCR solution.

Objectives

The objectives of the panel are to a) describe the need for a scalable, flexible strategy for eCR that meets both routine and emergency reporting needs; b) describe the infrastructure, decision support services, and HL7 document standards that support bi-directional communication for eCR; and c) explore strengths and limitations of the current eCR strategy.

Panel description

This didactic panel is organized so key leaders in the public health, informatics, and technical aspects of eCR can discuss the ecosystem (including knowledge, decision support, infrastructure, and standards) for bi-directional communication to support eCR. The Zika infection case reporting scenario will be used to illustrate key findings.
Several panelists are involved in the eCR Digital Bridge initiative (http://www.digitalbridge.us/about/) and will be able to share experiences based on pilots performed during 2017.

The panel will be moderated by Dr. Kathryn Turner, Chair of the Surveillance and Informatics Committee for the Council of State and Territorial Epidemiologists (CSTE), and member of the Governance Committee for the Digital Bridge initiative. CSTE represents state and local epidemiologists throughout the US engaged in public health surveillance and emergency response.

**Panel presentation 1: Framing the problem addressed by electronic case reporting (eCR)**

eCR is dependent upon public health laws and programmatic needs for disease surveillance and investigation that vary by jurisdiction. These needs may change in response to specific public health threats. In addition, eCR should support clinicians who desire bi-directional electronic communication with the public health community. During public health emergencies (e.g., Ebola and Zika outbreaks), the eCR solution must be scalable and flexible so it can be leveraged for newly identified conditions or be adapted for higher sensitivity to detect events during emerging outbreaks.

Implementation of eCR requires new automated processes for initiating and determining the reportability of clinical events. Determining reportability requires knowledge that is being systematically documented with input from epidemiologists across the US. To date, default logic has been defined for 69 of the over 200 conditions reportable in at least one jurisdiction. Some of the 69 conditions, upon preliminary review, include criteria that add complexity, such as epidemiologic criteria (41% of the conditions), symptoms (39%), preliminary lab results (30%), lab orders (20%), lab value comparisons (20%), and immunization data (5%). For 41% of the conditions, optional logic was requested to be available for use during emergencies or to accommodate reporting rules that sometimes vary by state. The standard codes required to automate the reporting logic informed the development of the event-based triggers that will be used by an EHR to initiate the case reporting workflow.

Dr. Catherine Staes, an informaticist and co-lead of the CSTE Reportable Condition Knowledge Management System (RCKMS) project, will provide background information to frame the problem addressed by eCR and briefly describe the knowledge developed to trigger the EHR to initiate the case reporting workflow and determine reportability. Trigger codes and reporting rules for Zika infection will be used to illustrate key concepts.

**Panel presentation 2: Defining a long-term public health clinical decision support (CDS) strategy**

As the knowledge required to support case reporting evolves from unstructured to more structured and standardized formats, it becomes suitable for electronic clinical decision support (CDS). CDS for case reporting confronts two challenges: a) While EHRs are moving toward local CDS capabilities, it will take several years for EHR systems to consistently support this capability; and b) public health-related CDS knowledge, such as Zika infection detection and reporting rules, may differ from jurisdiction to jurisdiction. Therefore, there is an ongoing need to manage reporting rules in a distributed manner. The Reportable Condition Knowledge Management System (RCKMS) is a component of the current eCR solution. RCKMS has the following capabilities that serve key the above requirements: a centralized CDS service that can be accessed by EHRs until they all have local CDS capabilities; and a knowledge authoring environment that allows ongoing distributed rule authoring (whether these rules are accessed through a central service or eventually exported for local EHR implementation).

Dr. Noam H. Arzt, president of HLN Consulting, LLC - the entity building the RCKMS service and authoring tool for CSTE and CDC - will describe how RCKMS supports the strategy for public health knowledge management, and how it will evolve over time to provide the systems and services to satisfy short-, mid-, and long-term public health CDS requirements. In addition, Dr. Arzt will reference emerging technical standards that support this work with which RCKMS is compliant. CDS concerning Zika infection will be used to illustrate the ideas presented.

**Panel presentation 3: Building the ecosystem on a scalable, flexible services platform**

Knowledge and CDS to enable automated case reporting must reside in an ecosystem that supports collaboration between clinical and public health partners and includes the necessary tools and services. Centralization of services is a focal theme to ease the implementation burden for both jurisdictions and reporters. Part of the eCR solution includes hosting key services on the Association of Public Health Laboratories (APHL) AIMS platform. AIMS provides a secure and flexible infrastructure for both messaging and public health decision support that determines the reportability of the condition, and to which jurisdiction the case report should be sent. AIMS is built on top of Amazon Web Services (AWS) which provides extreme scalability as demands and needs grow over time. The platform serves to connect clinical EHRs through trust networks to the RCKMS and finally to public health agencies.
Patina Zarcone-Gagne, Director of Informatics at APHL, is a member of the Digital Bridge Governance Group and chairs the Digital Bridge Sustainability Workgroup. Ms Zarcone-Gagne will discuss a) how the AIMS Platform is being used as part of the national ecosystem for routine public health surveillance, b) describe services and validation and transport functions supported by the platform, and c) briefly describe the process for clinical systems to on-board to AIMS and the status and timeline for the national eICR project.

**Panel presentation 4: Establishing standards for bi-directional exchange to support eCR**

Finally, the eCR ecosystem requires standards for bi-directional information exchange. Public health case reporting had been conspicuously absent from Meaningful Use criteria, and thus has not benefitted from the HITECH incentives that have met other public health needs (e.g., immunization registries and electronic laboratory reporting). The absence was partly because the standards for public health case reporting have been historically problematic even though standards are central to the application of incentives. In the past, standards development efforts suggested having individual standards for each of the more than 200 conditions reportable to a US public health agency. Another past effort suggested harmonizing the case reporting content for all programs and jurisdictions into one standard. Both approaches were problematic for policy and practical reasons, but laid the groundwork for the current approach.

The current approach for standards development supports a two-step automatically-initiated reporting process. To develop that standard for the first step, epidemiologists with CSTE identified a set of data (all of which should be available in certified base EHR products) that they expect will be sufficient to meet most, if not all, public health needs for initiating a case investigation. The HL7 Public Health and Emergency Response (PHER) working group used the data set to develop the electronic Initial Case Report (eICR) using the Consolidated Clinical Document Architecture (C-CDA) document. The second step of the eCR solution involves providing information from a public health agency back to clinical care. The information must be specific to the patient and the condition and indicate the reportability status of the information shared in the eICR. To develop the standard for this second step, the HL7 PHER workgroup is developing a Reportability Response (RR) using the HL7 CDA document structure.

John W. Loonsk, MD FACMI, Executive Sponsor and lead developer of the two HL7 eCR standards, will discuss the approach, challenges, and limitations of the eICR and RR standards, again using the Zika infection scenario to illustrate specific issues. In addition, Dr. Loonsk will share information about the experience from testing and prototyping in a number of settings, and convey plans for enhancing eCR standards.

**Audience Discussion**

At least 15 minutes will be reserved at the end for a discussion with and among the audience. Questions will be posed, such as: From the health care enterprise perspective, do the requirements and systems for implementing eCR leverage your existing infrastructure and make case reporting more efficient? Are there unintended consequences that the informatics and public health community should be thinking about?

**Conclusion**

This panel will allow the AMIA audience to get updated information about the knowledge, infrastructure, and standards supporting electronic case reporting (eCR) and experiences gleaned from pilot implementations supported by the Digital Bridge initiative. Reaching the goal of bi-directional communication between clinical and public health settings will require engagement from the AMIA community.

**References**

The State of Interoperability and Health Information Exchange in the U.S.: Advancing Steadily or Treading Water?

Walter Sujansky, M.D., Ph.D.¹, Julia Adler-Milstein, Ph.D.², Ross Martin, M.D., M.H.A.³

¹Sujansky & Associates, LLC, San Carlos, CA; ²Department of Medicine, University of California, San Francisco; ³Chesapeake Regional Information System for our Patients (CRISP), Baltimore, MD

Abstract
Prompt and efficient access to clinical data, regardless of its source, is one of the key promises of electronic health record technology as EHRs approach universal adoption in the United States. But in the nation’s complex and fragmented health care system, that goal requires interoperability and health information exchange across disparate information systems and business entities. In this panel, experts from academia and industry explore whether the U.S. healthcare system is now finally nearing the objective of broad-based technical and organizational interoperability for health data or still struggling against formidable barriers that have frustrated past interoperability initiatives. The panel presentations and discussion will assess the current state of interoperability and health information exchange (IHIE), as well as the potential contributions of contemporary IHIE initiatives, such as the FHIR standards and the Carequality and CommonWell industry consortia, to advance IHIE. The speakers will demystify these initiatives, highlight their novel attributes, and identify their remaining limitations in delivering on the promise of interoperability and health information exchange, including lessons learned from personal experience. The discussion will engage the audience and explore specific actions that informatics and health I.T. stakeholders can take to advance interoperability and HIE goals.

Learning Objectives
1. Review the goals of clinical data interoperability and health information exchange and characterize the current state of interoperability and exchange from an empirical and realistic perspective.
2. Analyze several contemporary technology efforts to advance IHIE, including the FHIR standards and the Carequality and CommonWell industry consortia, with a particular emphasis on what may be new and different in these efforts versus similar to past unsuccessful approaches.
3. Describe the primary remaining challenges to achieving true interoperability and health information exchange based on applied research and the experiences and lessons learned from a specific health information exchange organization.
4. Highlight the most important actions that can be taken by the medical informatics and health information technology communities to accelerate progress today.

General Description
This didactic panel will explore whether the U.S. healthcare system is advancing steadily toward the objective of broad-based technical and organizational interoperability for health data or still struggling against formidable barriers that have frustrated past interoperability initiatives. The panel presentations and discussion will assess the current state of interoperability and health information exchange (IHIE), including through the presentation of empirical research findings. The panel will also assess the potential contributions of contemporary IHIE initiatives, such as the FHIR standards¹ and the Carequality² and CommonWell³ industry consortia, to advance IHIE. The speakers will demystify these initiatives, highlight their novel attributes, and identify their remaining limitations in delivering on the promise of interoperability and health information exchange, including lessons learned from personal experience. The discussion will engage the audience and explore specific actions that informatics and health I.T. stakeholders can take to advance interoperability and HIE goals.

Panel Presentations
The panelists will each address the following aspects of this topic: 
Julia Adler-Milstein, PhD will present current data on the state of IHIE in the U.S. She will describe challenges to IHIE to date, including technical, regulatory, and market barriers. She will describe current policy efforts to advance IHIE.

Walter Sujansky, MD, PhD, will describe the potential for FHIR standards to address long-standing technical barriers to IHIE, the current state of FHIR, and remaining challenges in realizing the expected benefits of FHIR. He will also describe two large initiatives to advance IHIE among healthcare organizations – Carequality and CommonWell – including their current state and future challenges.

Ross Martin, MD, MHA, will describe the current and planned capabilities of CRISP – the state-designated health information exchange for Maryland, which also serves the District of Columbia and West Virginia. He will highlight how the current policy, standards and certification processes impact real-world efforts to bring relevant clinical information to the point of need in support of care coordination and clinical decision making and will discuss ways in which the medical informatics community can shape our path forward.

**Timeliness and Value of This Topic**

This topic is timely and of broad interest because achieving the vision of IHIE is a top policy priority and a goal for which both the public and private sector are devoting substantial attention and resources. Today, there is a flurry of touted IHIE initiatives, resources, organizations, and policies. Yet, it is difficult to understand where we stand with respect to IHIE progress, and whether we are now on a fast upward trajectory due to recent developments, or stalled out because we have not yet addressed the key obstacles. Such an understanding is critical to guide ongoing efforts. Our panel will bring empirical evidence and observations based on deep expertise, both theoretical and practical, to help audience members separate reality from hype and understand where we have made progress, where we expect to make progress, and where substantial barriers remain. We will explain key “hot topics” in the IHIE space – FHIR, Carequality, and CommonWell – and give a realistic assessment of how they advance IHIE and where stumbling blocks remain. The panel members are well-suited to address this topic, in that they are knowledgeable and directly involved, yet unencumbered by vested interests in any particular approach, resource, or commercial endeavor.

We expect that this panel will draw a wide array of attendees – from academic, practitioner, industry, and policymaker backgrounds – who understand the importance of IHIE but are not fully informed on the topic. Because we focus on the latest advances, we expect that we will draw a large number of participants who want to understand topics such as FHIR in an accessible and not overly-technical way.

**Potential Discussion Questions**

1. What is the level of IHIE that we’re trying to achieve? Is there a consensus?
2. Why do we not yet have IHIE today, as generally envisioned, despite widespread EHR adoption? What are the key barriers holding us back?
3. What is FHIR and how is it different from previous interoperability standards? Which challenges to IHIE could FHIR address? What work remains for FHIR to deliver on its potential?
4. What are the Carequality and CommonWell consortia and what specific resources do they provide to advance IHIE? What are their similarities and differences? What are their remaining challenges in supporting IHIE?
5. What real-world obstacles do provider organizations, state entities, and health information organizations face in achieving envisioned levels of IHIE? Are their expectations realistic? Are they knowledgeable about the nature of the barriers? What can the medical informatics community do to help them advance their IHIE knowledge and goals?
6. What specifically could the informatics research community do to advance IHIE? Policy makers? Medical informaticists working in industry? NIH/NLM?

**Statement of Participation**

All participants have agreed to take part on the panel.

**References**


340
Panel Proposal: Computational Phenotyping on Diverse Data Sources

Jimeng Sun, PhD¹, Bradley Malin, PhD², Abel Kho, MD³, Joydeep Ghosh, PhD⁴,
Mark Craven, PhD⁵

¹Georgia Institute of Technology, Atlanta, GA, USA, ²Vanderbilt University, Nashville,
TN, USA, ³Northwestern University, Chicago, IL, USA, ⁴University of Texas Austin, TX,
USA, ⁵University of Wisconsin Madison, WI, USA

Abstract

Phenotypes are the measurable biological, behavioral and clinical markers of a condition or disease. The process of deriving research-grade phenotypes from clinical data using computer-executable algorithms is called computational phenotyping (or phenotyping for short) [1]. Phenotyping includes a range of approaches, ranging from specifying existing phenotype using expert-derived rules and discovery of new phenotypes through novel computational methods that account for the interplay of many features conditions simultaneously. Traditionally, phenotyping has focused on algorithms for extracting clinical phenotypes from electronic health records (EHR). However, there is an increasing need to broaden the scope of, as well as standardize the principles underlying, computational phenotyping: 1) to capture and process more diverse health related data (e.g., *omic and mHealth); 2) to extract more sophisticated phenotypes based on temporal and spatial patterns; and 3) to speed up algorithm development, validation and deployment to support clinical and research applications. The learning objectives include various topics in phenotyping for diverse data sources, which include:

- Advanced machine learning methodologies for phenotyping
- Phenotyping development on national health networks
- Multi-scale phenotype development from multiple data sources
- Privacy considerations and knowledge sharing about phenotyping

Intended Audience: Medical informatics researchers and practitioners

Introduction

The national Precision Medicine Initiative aims to enroll one million members in a national cohort that will integrate data from biospecimens, sensor and mobile technologies, and healthcare, largely from electronic health record (EHR) data, to advance biomedical discovery and improve health. The realization of this vision will require efficient and effective methods to convert data from diverse data sources into specific and reliable phenotype characterizations that can be used to predict an individual’s risk of disease or response to drug therapy.

The quantity and quality of biomedical data that is growing. It is being collected across an increasing collection, and diversity, of technologies and mechanisms (e.g., EHR, mHealth, genomic data), which are distributed across institutions. Thus, it is critical for the medical informatics community to appraise the state of affairs in computational phenotyping methods, provide guidance on how best to translate such methods (and the phenotypes themselves) into practice, and define an agenda for the future of research to make phenotyping scalable and reusable. We believe such a panel is badly needed to provide necessary background and to inspire discussion of research in phenotyping for diverse data sources.

Topics for discussion:

Advanced machine learning methodology for phenotyping (by Prof. Ghosh) Raw EHR data collected across diverse populations and multiple care-givers tend to be extremely high dimensional, unstructured, heterogeneous, and noisy. Manually querying and interpreting such data is a formidable challenge for healthcare professionals. In the past few years, however, there has been much progress in obtaining high-throughput “silver standard” phenotypes via computational/machine-learning approaches that extract insights from clinical notes, procedures and diagnoses pertaining to large patient databases. Dr. Ghosh will highlight some of the key approaches and findings, and discuss how one may further such progress by incorporating even more diverse data sources, including multi-scale m-health sensors and ICUs, the additional challenges posed, and some new machine learning methods to deal with such challenges.

Phenotyping development on national health networks (Dr. Kho) We will describe national network phenotyping, and the effect of multi-institutional and external to EHR data on phenotypes. Several large national networks leverage
Electronic Health Record data across multiple institutions to define clinical phenotypes for research and quality improvement. Additional sources of data, for example provided directly by patients or inferred by geocoded data linked to home address, can add additional depth and breadth to EHR data. Dr. Kho will describe phenotyping efforts from national networks such as PCORnet, the NIH-sponsored Electronic Medical Records and Genomics (eMERGE) network), Next-D2, and EvidenceNOW and outline how multi-institutional data, and data captured outside of the EHR contribute to more fully informing computed phenotypes.

Multi-scale phenotype development from multiple data sources (Prof. Craven): We will talk about phenotyping from multiple rich data types, including EHR, imaging, 'omics, and mobile sensor data. Methods for risk assessment, disease subtype identification, and disease trajectory profiling can be potentially improved by employing approaches that characterize diseases in terms of their signatures in multiple data types. Moreover, there are technical challenges that are common across many of these data sources, such as high dimensionality and sparse longitudinal measurements. We will discuss applications of such approaches to Alzheimer's disease, breast cancer, and asthma.

Privacy consideration and knowledge sharing (Prof. Malin): As the diversity of data sources grow, it becomes increasingly challenging to define a priori hypotheses to test. As such, biomedical investigators benefit when they have access to resources either to validate what they see in their own data and to generate new ideas. In this respect, it is critical to ensure that data from diverse studies are made available on a wide scale. At the same time, there remain concerns that making such data available, even in a summarized form, can lead to the privacy compromises for the patients from which the data was collected. Over the past year, we have made substantial strides in demonstrating how summary data derived from genomic and phenomic resources (namely electronic medical records) can be made accessible while accounting for adversarial capabilities. In this section of the panel, Dr. Malin will report on the results of the iDASH Genome Privacy Competition, recent advances in sharing genomic data summary statistics for the Sequence and Phenotype Integration Exchange (SPHINX), which consists of data on over 8000 subjects in eMERGE [2]. He will further provide intuition into the privacy models that are being set in place in the All of Us Cohort program of the NIH Precision Medicine Initiative.

Discussion questions
- What are the challenges in phenotyping diverse biomedical data sources?
- What are the rational privacy consideration for phenotyping?
- How can advanced machine learning methods help improve phenotyping development?
- What are different phenotyping initiatives across the US?

Panel lineup:
- **Jimeng Sun (moderator),** Ph.D., is an Associate Professor of School of Computational Science and Engineering and a co-director of center of health analytics and informatics at College of Computing at Georgia Institute of Technology. His research focuses on health analytics and medical informatics, especially in applying large-scale predictive modeling and computational phenotyping on biomedical applications. He has published over 100 papers, filed over 20 patents (5 granted). He has received SDM/IBM Early Career Data Mining Research Award 2017, ICDM Best Research paper in 2008, SDM Best Research Paper in 2007, and KDD Dissertation Runner-up Award in 2008. Dr. Sun received his B.S. and M.Phil. in Computer Science from Hong Kong University of Science and Technology in 2002 and 2003, and PhD in Computer Science in Carnegie Mellon University in 2007. Prior to joining Georgia Tech, He was a research staff member at IBM TJ Watson Research Center.

- **Mark Craven,** Ph.D. is a professor in the Department of Biostatistics and Medical Informatics at the University of Wisconsin, and an affiliate faculty member in the Department of Computer Sciences. He is the Director of the Center for Predictive Computational Phenotyping, one of the NIH Centers of Excellence for Big Data Computing. He is also the Director of the NIH/NLM-funded Computation and Informatics in Biology and Medicine (CIBM) Training Program, and a member of the Institute for Clinical and Translational Research, the Carbone Cancer Center, and the Genome Center of Wisconsin. The focus of his research program is on developing and applying machine-learning methods to the problems of inferring models of, and reasoning about, networks of interactions among genes, proteins, clinical and environmental factors, and phenotypes of interest.

- **Abel Kho** MD, MS is Associate Professor of Medicine and Preventive Medicine in the Feinberg School of Medicine at Northwestern University and Director of the Center for Health Information Partnerships (CHiP www.healthinformationforall.org). He has been PI or Co-PI of over $50 million in funding to the Chicagoland
or surrounding region including the ONC funded Chicago Health IT Regional Extension Center (www.chitrec.org), the PCORI funded Chicago Area Patient Centered Outcomes Research Network (http://capricorn.cdrn.org/), and the AHRQ funded Health Hearts in the Heartland consortium (http://www.healthyheartsintheheartland.org/) within the EvidenceNOW initiative. His research focuses on developing regional Electronic Health Record (EHR) enabled data sharing platforms for a range of health applications including high throughput phenotyping, cohort discovery, estimating population level disease burden, and quality improvement.

- **Bradley Malin**, Ph.D., is Vice Chair and Professor of Biomedical Informatics at Vanderbilt University. He is also a Professor of Biostatistics, a Professor of Computer Science, and is Affiliated Faculty in the Center for Biomedical Ethics and Society. He co-directs the Health Data Science Center, the Center for Genetic Privacy and Identity in Community Settings (an NIH Center of Excellence in ELSI Research), and the Big Biomedical Data Science Ph.D. program. Since 2007, he has led a data privacy consultation service for the Electronic Medical Records and Genomics (eMERGE) network, an NIH consortium. He is also the co-chair of the Data Privacy and Security Working Group of the All of Us Research Program of the U.S. Precision Medicine Initiative. His investigations on the empirical risks to health information re-identification have been cited by the Federal Trade Commission in the Federal Register and certain privacy enhancing technologies he developed have been featured in popular media outlets and blogs, including Nature News, Scientific American, and Wired magazine. He is an elected fellow of the American College of Medical Informatics and was honored as a recipient of the Presidential Early Career Award for Scientists and Engineers (PECASE).

- **Joydeep Ghosh** is currently the Schlumberger Centennial Chair Professor of Electrical and Computer Engineering at the University of Texas, Austin. Dr. Ghosh’s research interests lie primarily in data mining and their applications to a wide variety of complex real-world problems. He has received 14 Best Paper Awards over the years, including the 2005 Best Research Paper Award across UT and the 1992 Darlington Award given by the IEEE Circuits and Systems Society for the overall Best Paper in the areas of CAS/CAD. Dr. Ghosh has been a plenary/keynote speaker on several occasions such as ICHI’15, ICDM’13, Health Informatics workshops at KDD14, ICML13 and ICHI13; and has widely lectured on intelligent analysis of large-scale data. He served as the Conference Co-Chair or Program Co-Chair for several top data mining oriented conferences, including SDM’13, SDM’12 and KDD 2011.

We affirm that all panel participants have agreed to participate and have contributed to the preparation of this document.

**References**


Social and Behavioral Determinants of Health: Fundamental Informatics Challenges for Enriching Health IT Systems

Panelists
Albert W. Taylor, MD, Daniel J. Vreeman, PT, DPT, MS, Matteo Verzola, MS, Jon Puro, MPA

Abstract
In 2015, the Institute of Medicine released Capturing Social and Behavioral Domains and Measures in Electronic Health Records: Phase 2. This document identified core domains that were recommended for inclusion in patients’ health histories, and endorsed specific measures for each of these core domains. This work reflected the substantial impact of these factors on patients, families and communities.

This panel is designed to share the experience of a diverse set of stakeholders who have been engaged in the recognition of social determinants of health (SDH) factors on health status and the current and future state of SDH domains and measures within electronic health records. The panel will present the current ONC (Office of the National Coordinator) perspective on the 2015 HIT (Health Information Technology) certification criteria that included many of the IOM endorsed domains and measures, highlight the critical work that a standards development organization (LOINC) plays both currently and in the future, implementation challenges identified and addressed within a health care system (Verzola), and the integration and use of SDH data within primary health care (Puro). The panel presentations will emphasize lessons learned as well as current and future barriers that can be addressed to foster more inclusion of SDH factors in clinical care.

This panel is sponsored by the AMIA Social Determinants of Health (SDH) discussion group.

Introduction and Background
Good health results from interactions with the healthcare system, and also from factors that include exposure to early childhood events, the environment, neighborhoods, and community and family cultural influences. Recognizing, understanding, and including these factors in health care delivery are an important part of helping individuals and communities improve their health status. Data on these ‘social determinants of health’ (SDH) can be used on their own or through predictive modeling to help identify barriers to care and prognosis for recovery, inform clinical decision-making and the tailoring of interventions, enable care coordination with appropriate services, and support development of tailored clinical decision support in electronic health records. SDH are also gaining recognition as factors that should be incorporated into new schemas for managing patients in the framework of value-based care and accountable care organizations. Data related to SDH are of great interest to large-scale, cooperative research projects like PCORnet (cf. PCORnet’s SDH research interest group). Currently, there is no accepted, comprehensive standard for capturing data related to SDH in the electronic health record (EHR).

Panel Description
In this panel, we will discuss the present opportunities and challenges in the current state of SDH information, including standard development for SDH information, ways to improve the identification and use of these data areas, and lessons learned from implementation pilots. The best approach for capturing the range of SDH that can be
meaningfully intervened upon by clinical providers and have substantive impact on patients’ health is still unclear, but this current work will inform the future pathway. Jon Puro will moderate the current and future role and impact of SDH from multiple perspectives, including the federal government, a major EHR vendor, and two clinical perspectives. The experiences and lessons learned from this early SDH work will help set the stage for ongoing work in this area as we move towards integration of these data sets in health care systems. Attendees will have an opportunity to understand the critical and difficult nature of including SDH information within the clinical care continuum and the urgent need for increasing attention to the inclusion of these factors in health care delivery models.

The panel will focus on issues including:

1. A review of various development and regulatory activities from the federal perspective (ONC)
2. Representing SDH in vocabulary standards
3. Implementing IOM (Institute of Medicine) recommendations within the EHR
4. Integrating SDH data sets into primary care and research: an IT perspective

Presenters

ONC: Current and Future (Taylor)

Dr. Taylor will discuss ONC’s perspective and approach to advancing the use of SDH in HIT. ONC recognizes and appreciates the work of Institute of Medicine (now National Academy of Medicine) and other organizations in the recognition of the direct health linkages of SDH and the degree of consensus and adoption of standardized representation of these domains and measures. This work was critical in the inclusion of a number of specific domains in ONC’s 2015 Edition HIT Certification Criteria and subsequent Interoperability Standards Advisories. Many of these measures, as well as those more directly representative of clinical domains, are trademarked or copyrighted property. These protections must be taken into consideration when negotiating the new use case using standardized terminology such as LOINC, including them in certification criteria and in commercial EHR products. Dr. Taylor will discuss ONC’s experience in navigating these issues so that measure developers and intellectual property owners can advance their own instruments in a HIT-enabled manner.

Representing SDH content in vocabulary standards (Vreeman)

Dr. Vreeman will discuss the fundamental need for representing individual- and community-level SDH variables in common vocabulary standards. He will discuss how enriching existing standards with representations of SDH measures can improve interoperable sharing, aggregation, and computation in clinical, population, and research applications. He will describe the growing momentum for incorporating SDH variables in LOINC (http://loinc.org), a freely available international standard for identifying health measurements, observations, and documents. As a widely adopted vocabulary standard in EHRs, LOINC codes enable interoperable data exchange and promotes reuse of this data through aggregation, analytics, and cross-study comparisons. Dr. Vreeman will provide an overview of LOINC’s current model [11], coverage for representing SDH content, and how implementers can use this content. He will also discuss the current efforts to catalyze addition of SDH content in LOINC and how users can propose new individual or community level variables for inclusion.

Implementing IOM Recommendations within the EMR (Matteo Verzola)

Mr. Verzola will discuss the approach to implementing the IOM recommendations around capturing social determinants of health within the Epic EMR, sharing valuable lessons learned for all HIT systems. He will cover approaches for ensuring data is captured as part of the already existing clinical workflows, as well as what tools physician, nurses, and care managers can use for reviewing a patient's individual domain risk and overall risk related
to SDH. The discussion will then move to how SDH information can be leveraged for other workflows, including research, analytics, and referral workflows to community resources to address SDH challenges.

Primary Care Experience from an IT perspective (Puro)

Mr. Jon Puro will discuss the technical methods used at OCHIN and the ADVANCE Clinical Data Research Network for importing and integrating geospatial data related to community level social determinants of health ("Community Vital Signs") into the ADVANCE Research Data Warehouse. ADVANCE and OCHIN regularly receive over 70 sets of geospatial metrics in partnership with the Robert Graham Center to use for patient care as well as community-focused research. Jon will describe these data sets, explain the technical extract, transform and load processes, and discuss current efforts and future plans for integrating these data into research and patient care. Jon serves as the PI of the ADVANCE Clinical Data Research Network.

Discussion Questions

1. What are the top three benefits and challenges for the inclusion of SDH domains and measures in EHRs?
2. How can the health informatics community leverage the lessons learned from prior EHR development efforts to help accelerate the development, deployment and integration of SDH standards?
3. What are appropriate ways to address IP issues that result from using SDH measures?
4. What can AMIA do to assist with the integration of SDH into clinical care?

Panel Organizer Statement: All participants have agreed to take part in the panel and discuss the topics as outlined above.

References

Optimizing Patient Care through Clinical Decision Support: Identification of Opportunities and Call to Action by the National Academy of Medicine

James Tcheng, MD\textsuperscript{1}, Kensaku Kawamoto, MD, PhD, MHS\textsuperscript{2}, Blackford Middleton, MD, MPH, MS\textsuperscript{3}, Jonathan M. Teich, MD, PhD\textsuperscript{4}, Scott Weingarten, MD, MPH\textsuperscript{5}

\textsuperscript{1}Department of Medicine, Duke University, Durham, NC, \textsuperscript{2}Department of Biomedical Informatics, University of Utah, Salt Lake City, UT, \textsuperscript{3}Apervita, Inc., Chicago, IL, \textsuperscript{4}Departments of Medicine and Emergency Medicine, Harvard University, Boston, MA, \textsuperscript{5}Cedars-Sinai Medical Center, Los Angeles, CA

Abstract

With support from the U.S. Office of the National Coordinator for Health IT (ONC), the National Academy of Medicine (NAM) convened a collaborative effort with health care leaders to identify untapped opportunities and practical strategies for improving clinical decision support (CDS) practice and adoption. Through a series of NAM-facilitated working meetings and work groups focused on four key CDS areas (content authoring, platform integration, operations, and dissemination), expert authorities reviewed current and emerging CDS practices and identified collaborative opportunities to accelerate national progress in the real-time application and use of CDS to improve health and health care decision making. In this panel, leaders of this NAM effort will summarize the key findings identified by this multi-stakeholder collaboration of clinicians, academics/researchers, industry, patient advocates, and the government. The panelists will also present cross-cutting Priorities for Action developed at the conclusion of this work and engage the audience in an interactive discussion on how we can work together to pursue these priorities with the goal of optimizing health and health care at scale.

Description

The panel will be organized as follows:

<table>
<thead>
<tr>
<th>Time</th>
<th>Speaker</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>5 min</td>
<td>Tcheng</td>
<td>Need for bridging the gap between the promise and reality of CDS</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Overview of NAM effort to advance CDS-enabled care improvement</td>
</tr>
<tr>
<td>10 min</td>
<td>Kawamoto</td>
<td>CDS content authoring – current state, challenges and opportunities</td>
</tr>
<tr>
<td>10 min</td>
<td>Weingarten</td>
<td>CDS platform integration – current state, challenges and opportunities</td>
</tr>
<tr>
<td>10 min</td>
<td>Teich</td>
<td>CDS operations – current state, challenges and opportunities</td>
</tr>
<tr>
<td>10 min</td>
<td>Middleton</td>
<td>CDS dissemination – current state, challenges and opportunities</td>
</tr>
<tr>
<td>10 min</td>
<td>Tcheng</td>
<td>Cross-cutting Priorities for Action</td>
</tr>
<tr>
<td>30 min</td>
<td>All</td>
<td>Call to action and interactive discussion with audience</td>
</tr>
</tbody>
</table>

Dr. Tcheng, who oversaw the NAM’s efforts as the Chair of the Planning Committee, will serve as the moderator and provide an overview of the motivation for the NAM effort and how the effort was carried out. Then, the leaders of each of the convened Work Groups will describe the current state, challenges, and opportunities identified in the areas of CDS content authoring (Kawamoto), platform integration (Weingarten), operations (Teich), and dissemination (Middleton). Dr. Tcheng will then present the cross-cutting Priorities for Action identified by the effort. Finally, Dr. Tcheng will moderate an interactive discussion with the audience on how the informatics community can work together with each other and their cross-disciplinary colleagues to pursue the action priorities and achieve the full promise of CDS-enabled care optimization.

Tcheng: Dr. Tcheng is a practicing interventional cardiologist and Professor of Medicine and Community and Family Medicine (Informatics) at the Duke University School of Medicine. He chairs the Informatics and Health IT
Task Force of the American College of Cardiology, in which role he seeks to advance evidence-based cardiac care at a national scale. He served as the Chair of the Planning Committee for this NAM initiative to advance CDS. In opening the panel discussion, Dr. Tcheng will provide brief overview of the motivation for the NAM CDS effort and how the effort was carried out.

Motivation for Effort: The foundational problem that motivated the NAM CDS effort is the large gap between the immense promise and current reality of CDS. In the last decade, electronic health record (EHR) adoption rates have skyrocketed; nearly every provider and hospital in the United States uses some form of an EHR, paving the way for increased use of CDS tools that leverage EHR data to provide decision support to clinicians and patients. These tools are directed toward reduction of errors and adverse events, promotion of best practices for quality and safety, cost profile improvement and more. Many government and payor programs, such as the Merit-Based Incentive Payment System (MIPS), the Protecting Access to Medicare Act (PAMA) and Accountable Care Organization (ACO) programs, include specific or implied requirements for the use of CDS.

At the same time, consumers are increasingly generating and collecting data via wearable devices and mobile health apps that supplement the information individuals provide about their preferences, lifestyle factors and goals of care. New models of care delivery and payment that aim to reward quality and value of care rather than volume of services cannot succeed without active engagement of patients and family caregivers. CDS interventions that leverage multiple data types, bring the most current and relevant evidence to bear on clinical decisions, produce actionable recommendations from the mounds of data, and account for conflicting influences and values affecting decision making are necessary to care for patients with multiple chronic illnesses or comorbidities.

For a variety of reasons, however, CDS implementation and actualization remain nascent. To realize the potential of these tools to improve the quality, safety and efficiency of healthcare we must identify key priorities for action needed to realize the full potential of CDS within real-world environments.

Approach Taken: With support from the ONC, the NAM (formerly the Institute of Medicine) convened a collaborative effort with health care leaders to better understand potential opportunities and practical strategies for improving CDS practices and adoption. Expert authorities first met in-person (March 16, 2016) to describe current and emerging CDS practices, identify collaborative opportunities to accelerate progress in the real-time application and use of CDS to inform health and health care decision making, and provide guidance on implementation challenges and strategies at a national scale. The second meeting in the series (October 27, 2016) provided an opportunity to review, discuss, and refine actionable next steps to address CDS content, technical implementation, operations, and distribution. In a final meeting (February 10, 2017), key public and private stakeholders were engaged to discuss implementation and next steps for continued collaboration. The project was guided by an external planning committee and was not a product of the ONC.

In addition to the meeting series, the NAM CDS Planning Committee initiated small work-groups to address: content (CDS authoring), platform integration (technical implementation), clinical functionality and measurement (operations), and dissemination and assessment (distribution). Each small work-group met virtually to help develop action plans on their designated topics, which were consolidated into cross-cutting Priorities for Action by the work group members and the larger community of stakeholders convened through this effort.

Kensaku Kawamoto: Dr. Kawamoto is Associate Chief Medical Information Officer, Assistant Professor of Biomedical Informatics, Director of Knowledge Management and Mobilization, and Chair of the CDS Committee at the University of Utah. Dr. Kawamoto is also co-chair of the HL7 CDS Work Group and co-Initiative Coordinator for the ONC/CMS-supported Clinical Quality Framework (CQF) initiative to develop and validate a harmonized set of interoperability standards for CDS and electronic clinical quality measurement. Dr. Kawamoto led the NAM CDS work group on CDS content authoring and will present on the current state, challenges, and opportunities associated with the creation, management, and curation of computable, effective CDS.

Scott Weingarten: Dr. Weingarten is senior vice president and chief clinical transformation officer at Cedars-Sinai. He is a professor of Medicine at Cedars-Sinai and a clinical professor of Medicine at the David Geffen School of Medicine at UCLA. Dr. Weingarten was the co-founder, president and chief executive officer of Zyox Health, which is the leader for order sets and care plans for EHRs, and he is the co-founder and Chairman of the Board at Stanson Health, a leading CDS solutions provider. Dr. Weingarten led the NAM CDS work group on CDS platform integration and will present on the current state, challenges, and opportunities associated with standards-based integration of CDS into EHR platforms.
Jonathan Teich: Dr. Teich is assistant professor of medicine and emergency medicine, and a board-certified attending physician at Harvard Medical School. He has served as Chief Medical Informatics Officer for Elsevier, a leading CDS content and solutions provider, and serves as Clinical Architecture Lead for Bahmani and for OpenMRS, which provides critical open-source health IT solutions to over 50 low- and middle-income countries. Dr. Teich was the founder and first director of the Clinical Informatics Research and Development department at Partners Healthcare System and has authored or co-authored over 100 papers, books, and other publications in the field of medical informatics and healthcare information systems, with a particular concentration in CDS. He co-directed the HHS Roadmap for National Action on CDS, and has served as the CDS subject matter expert for ONC. Dr. Teich led the NAM CDS work group on CDS operations and will present on the current state, challenges, and opportunities related to CDS functionality, usability, implementation, and impact evaluation.

Blackford Middleton: Dr. Middleton is Chief Informatics & Innovation Officer at Apervita, Inc., the provider of the largest marketplace for the creation, exchange, and use of data and analytics in healthcare. He is immediate Past-Chair of the Board of Directors of the American Medical Informatics Association (AMIA), and Instructor in the Harvard TH Chan School of Public Health in the Departments of Health Policy and Management, and Policy Translation and Leadership Development. Previously, he was a Professor of Biomedical Informatics, and/or of Medicine, at Stanford, Harvard, and Vanderbilt Universities, and he held executive leadership roles at MedicaLogic/Medscape, Partners Healthcare System, and at Vanderbilt. Dr. Middleton’s work is focused on clinical informatics – the applied science surrounding strategy, design, implementation, and evaluation of clinical information systems in complex environments. Dr. Middleton led the NAM CDS work group on CDS dissemination and assessment in the marketplace, and will present on the current state, challenges, and opportunities related to widely disseminating effective CDS, including discussions of the CDS marketplace, the role of public and private knowledge sources, business rules in ensuring a vibrant and successful marketplace, constructs for feedback loops to inform value, and the financial business case for CDS.

James Tcheng: Following the presentation on findings from the four NAM CDS work groups, Dr. Tcheng will share the 13 cross-cutting Priorities for Action defined across 3 key areas: (i) developing, testing, establishing, validating and applying standards; (ii) encouraging adoption, use, and assessment at the delivery system level; and (iii) establishing a national CDS infrastructure. Dr. Tcheng will then moderate an interactive discussion with the audience on how we as an informatics community can work together to move forward on these priorities and realize the full promise of CDS-enabled health care improvement.

Significance of panel topic and anticipated audience

The anticipated audience is biomedical informatics professionals from diverse backgrounds seeking to improve the quality, safety, and/or value of health care. This panel topic is important because CDS represents one of the most promising informatics-enabled strategies available for improving health and health care, and this panel provides a comprehensive analysis of the CDS landscape as well as a concrete guide for priority actions that the informatics community can take to finally realize the promise of CDS-enabled care optimization.

Discussion questions

How can you or your organization contribute to the Priorities for Action?

What can ONC, NAM, and others do to accelerate progress?

What barriers do you see, and how do you think we should overcome them?

Participation statement

All proposed panelists have agreed to participate in the panel if the proposal is accepted.
The Personal Journey – Informaticists Confront Their Own Health Issues

Jessie Tenenbaum, Ph.D. 1, Ed Hammond, Ph.D. 2, Ross Koppel, Ph.D. 3, Michael Kamerick, BSCS 4
1Duke University, Durham, NC; 2Duke University, Durham, NC; 3University of Pennsylvania, Philadelphia, PA, 4Sequoia Consulting, San Anselmo, CA

Abstract
As informaticists, we often create tools and manage data to assist patients in making difficult treatment decisions. What happens when one of us becomes a patient? How do we approach the dilemma of our own complex disease and possible response, especially when multiple treatment options are available? What is our decision-making modality? Do we engage in the same calm, rational process we espouse to others, or does our new-found sense of vulnerability overwhelm our analytic faculties? How has our own experience changed the way we approach the goals of our profession? A panel of informaticists discuss their personal encounters with disease, decision making, treatment, and outcomes.

Introduction
The concept of a patient-centric healthcare system has gained great currency in recent years, especially since the creation of PCORI. How has the informatics community responded to the idea of putting the patient at the center of the treatment process? In this panel discussion, we will see how several informaticists took the most direct path one can take to a patient centered worldview -- by becoming patients themselves.

Each of the panelists will present their personal experience of having become a patient confronted with a serious health condition, one that required a thoughtful and thorough approach to the decisions about which treatment path to follow and what outcomes to hope for. The panelists will discuss their own subjective responses to their diagnoses. They will also discuss the tools and resources they made use of to assist them in their decision-making processes, as well as suggested improvements to those tools. Individual presentations will be 15-20 minutes each, with sufficient time allowed for audience engagement after the presentations and during the last part of the session.

Presentations
1. Prostate Cancer – What I Knew Then and What I Know Now
   Michael Kamerick, BSCS

   In 2005, Mr. Kamerick was diagnosed with prostate cancer. At the time, he was employed at the UCSF Breast Care Center as the head of a team building various information systems to support cancer care and research. Some of those systems were built to help breast cancer patients understand their risk and possible treatments, and some were built to help physicians confer on the issues of specific patients. In that context, working alongside a team of oncologists, he had a wealth of information and support available to assist him. His presentation will discuss the steps and tests taken leading up to his final diagnosis, the tools he used to understand the nature and implications of his disease, the decision process he went through to determine a preferred course of treatment among those available at the time, and the outcomes of that treatment course.
2. Experiences in Aging
Ed Hammond, Ph.D.

My first serious experience with the healthcare system happened in 2012. Sitting at my computer terminal at home in the evening, I suddenly felt heat in the back on my head and suffered unbelievable pain. I live alone, so I will share what I did, and what happened to me over the next few days. The diagnosis was subarachnoid hemorrhage, and although my care was excellent, the experience was terrifying. I also want to share my experience with a diagnosis of Type 2 diabetes – what I did and am doing and the positive results. Finally, I will share my experiences with a wearable sensor and the reporting of that data into an Epic flowsheet as part of MyChart to my personal physician. Aging can be great...

3. I get to keep my hair, right? Therapeutic decision making for stage 0 cancer from the patient perspective
Jessie Tenenbaum, Ph.D.

In December 2016 Dr. Tenenbaum was diagnosed with ductal carcinoma in situ, or DCIS. Even for this non-life-threatening condition, a whole host of treatment decisions had to be made from medication to surgery to radiation. Tools and concepts that had previously been professional, including the patient portal, family health history, risk alleles, pharmacogenomics, and drug-drug interactions, were suddenly made personal. Dr. Tenenbaum will discuss the many decisions to be made throughout the treatment process, how informatics resources helped or fell short, and how the experience has informed her perspective on what we do as informaticists.

4. HIT and the patient experience – culture and workarounds in two nations
Ross Koppel, Ph.D.

Professor Koppel had the unfortunate and very recent advantage of experiencing both the NHS’ hospital service and a major American hospital’s service. In both cases he was able to observe the healthcare information technology as it affected clinical workflow and clinical functionality. While the systems were different in level of penetration, sophistication and what we shall charitably call interoperability, the interplay of clinical work with the HIT is both the subject of his observations and the focus of his talk at this panel.

Many of the themes are the obvious: usability, data accessibility, quality of data presentation, integration with workflow, responsiveness to clinical need and clinical efficiency, and IT staff’ responsiveness of the clinicians’ requests. Some of the themes, however, are perhaps not obvious: workarounds – both to the HIT and the rules, the effect of the technology on patients' understanding of their treatment and prognosis, the role of the technology in enhancing and reducing frustration by the medical staff, and the role of the HIT in expectations of improvements.

It may be of note that Koppel is a PhD sociologist and has been teaching research methods and ethnography, in addition to statistics and biomedical informatics.

Conclusion
The individual’s own experience of disease can bring abstract professional concepts into the personal realm. Such events present informaticists with an opportunity to employ their own tools, to validate their work, and to challenge themselves to improve and change what they have already built upon.
And Now for Something Completely Different: Successful Career Transformations in Biomedical and Health Informatics

Amy Y. Wang, MD1, Douglas B. Fridsma, MD, PhD, FACP, FACMI2, Judy Murphy, RN, BSN, FACMI, FHIMSS, FAAN3, Chintan O. Patel, PhD4, Hillary Ross, JD5

1Northwestern University Feinberg School of Medicine, Chicago IL; 2American Medical Informatics Association, Bethesda, MD; 3IBM, Washington, DC; 4Applied Informatics, Inc., New York, NY; 5Witt/Kieffer, Oak Brook, IL

Abstract

The health and biomedical career landscape continues to grow and evolve as new technological opportunities and challenges arise. As professionals pursue new opportunities, careers take winding paths across unknown terrain. In an interactive panel discussion, informatics leaders who have experienced or guided colleagues through dramatic career transformations will address this topic. Panelists have transitioned between academic, clinical, research, educational, government, nonprofit, and industry realms. The panel will discuss the development of “soft” skills, such as vision, strategy, perseverance, resilience, communication, mentorship, and emotional intelligence, which are valuable for all professionals. Panelists will also address skills that are relevant to specific domains and roles. The presentation will begin with defined questions and then discuss questions from the audience. The purpose of the panel is to help informatics professionals identify and develop skills to understand the full breadth of career opportunities and paths and prepare and plan for career transformations. The intended audience includes informatics students or professionals in all career stages who intend to transform their careers or mentor colleagues. This panel is jointly sponsored by the Education and Student Working Groups and supports AMIA’s core mission of professional growth and leadership development for our members.

Learning Objectives

• Recognize the diversity of paths, organizations, roles, and activities in health informatics careers
• Identify, develop, transfer, and apply essential skills for entering and succeeding in different roles and domains in health and biomedical informatics
• Prepare and plan for new career opportunities and transformations in health and biomedical informatics

Introduction

As the field of health and biomedical informatics continues to evolve, there is a growing need for competent informatics professionals to translate health information into improved outcomes. Professional growth and leadership are top priorities in AMIA’s vision and strategy1. Professional organizations have defined competencies for health informatics, and there are many educational programs and certification pathways for developing and demonstrating health informatics expertise2. As health informatics is an interdisciplinary, collaborative field, informatics professionals must develop multiple areas of interest and expertise3, 4 and plan to shift roles and areas of focus throughout their careers3, 4. Desired qualities include critical thinking skills and the ability to “think and learn” on the job. Specific areas of competency include health care, health terminology, clinical workflow, health information systems, information management, privacy, security, regulations, quality improvement, root cause analysis, computer science, databases, data analytics, software engineering, and mobile technologies3. Health information technology employers have also identified advanced and interpersonal skills, such as communication, negotiation, management, strategic thinking, leadership, and planning3. Organizations report significant needs and challenges in recruiting skilled informatics professionals4, especially those competent in clinical workflow, technology, politics, and interpersonal skills3. There are unique challenges for entry-level and mid-career professionals entering the field as well as for seasoned informatics professionals desiring career advancement.

Our speakers demonstrate diverse interests, experiences, and meandering paths across different domains and functions in health informatics. They represent the rich depth and breadth of the field of health and biomedical informatics. Panelists have served in academic, commercial, nonprofit, and governmental organizations in diverse endeavors such as product design and implementation, standards development, health system operations, public
Panelists will introduce themselves and share their career experiences, focusing on types of opportunities, how and where to start, challenges with transitions, and how to succeed before, during, and after transitions. Panelists will answer an initial set of questions and then engage in an interactive discussion with the audience. Topics will include:

- What are key success factors, such as education, competencies, experience, and mentorship?
- What is the role of educational programs, advanced certification, support groups, and other resources?
- How do roles and essential skills shift through different career stages and levels of leadership?

Panelists

Amy Y. Wang, MD (Panel Organizer and Moderator) is an assistant professor at Northwestern University Feinberg School of Medicine in Preventive Medicine and Family and Community Medicine. Her expertise is in standards and interoperability, health terminology, electronic health records, and informatics education. She was previously at Intelligent Medical Objects (IMO), where she developed and implemented terminologies for EHRs to help health care organizations transition to ICD-10-CM and meet Meaningful Use requirements. Prior to IMO, she worked for Greenway Medical and focused on EHR development, implementation, and clinician support. While at the College of American Pathologists (CAP), she served on the international team that developed SNOMED CT.

Douglas Fridsma, MD, PhD, FACP, FACMI (Panelist) is President and Chief Executive Officer of AMIA and an expert in interoperability, standards, and health IT. Dr. Fridsma was previously Chief Science Officer for the Office of the National Coordinator for Health Information Technology (ONC), responsible for the portfolio of technical resources for Meaningful Use and interoperability. At ONC, he developed the standards and interoperability framework for accelerating development of technical specifications for interoperability, and collaborated with the NIH and other agencies to establish priorities in the PCOR Trust fund. Prior to ONC, he held academic appointments at the University of Pittsburgh, Arizona State University, University of Arizona, and Mayo Clinic, and practiced medicine at the Mayo Clinic Scottsdale. As a former board member of HL7 and the Clinical Data Interchange Standards Consortium (CDISC), and current board member of SNOMED international, he has been instrumental in developing standards bridging clinical care and research. He is a Fellow of the American College of Medical Informatics (ACMI).

Judy Murphy, RN, BSN, FACMI, FHIMSS, FAAN (Panelist) is Chief Nursing Officer at IBM Global Healthcare, where she is responsible for building relationships and expanding business across the healthcare industry. She is a strategic advisor to clients and helps develop health IT solutions for providers to improve health and healthcare, and lower costs. Prior to IBM, she was CNO and Deputy National Coordinator for Programs and Policy at the ONC in Washington D.C. In these roles, she led federal efforts to assist health care providers in adopting health information technology to improve care and to promote consumers’ greater understanding and use of health information technology for their own health. She came to ONC with more than 25 years of health informatics experience at Aurora Health Care in Wisconsin, an integrated delivery network with 15 hospitals, 120 ambulatory centers, and over 30,000 employees. As Vice President-EHR Applications, she led their EHR program since 1995, when Aurora was an early adopter of health IT. She has served on the AMIA and HIMSS Boards of Directors and is a Fellow in the American Academy of Nursing, ACMI and HIMSS. She has received numerous awards, including the AMIA 2014 Don Eugene Detmer Award for Health Policy Contributions in Informatics, the HIMSS 2014 Federal Health IT Leadership Award, and the HIMSS 2006 Nursing Informatics Leadership Award.

Chintan Patel, PhD (Panelist) is Chief Technology Officer and cofounder of Applied Informatics and Trialx. Previously, he worked at IBM’s T.J. Watson Research, on ontology reasoning and information extraction that then became the infamous Jeopardy winning AI software “Watson”. He holds a PhD in biomedical informatics from Columbia University. His areas of expertise include software architecture, mobile technologies, semantic technologies, biomedical ontologies, machine learning and natural language processing.

Hillary Ross, JD (Panelist) is Managing Director and Leader of the Healthcare IT Practice with executive search firm Witt/Kieffer. She specializes in the recruitment of senior-level physicians and informaticians, including Chief Medical Information Officers and Chief Health Informatics Officers for academic medical centers, hospitals, health delivery systems, vendors, and consulting firms across the country. She is nationally recognized for her expertise in recruiting highly specialized and executive leadership positions, such as Chief Research Informatics Officer and Directors of Bioinformatics, Biomedical Informatics, Data Analytics, Personalized Medicine, Quality, and
Innovation. She is a frequent speaker and published author on informatics, IT, and executive recruiting and is based in Witt/Kieffer’s Oak Brook, Illinois office.

Panelist Contributions and Presentations

Amy Y. Wang, Moderator, will introduce the panel, drawing from her experience in clinical practice and health informatics, including major career transformations. She went from clinical practice into standards and terminology development at CAP, a nonprofit. She then implemented standards in EHRs through Greenway and then at IMO, which she helped grow from startup to mature company. She transitioned into full-time academia in 2016. Dr. Wang will touch on universal skills and those that apply to specific areas and roles in health informatics.

Douglas Fridsma, Panelist, has represented the voice of health informatics professionals at the highest levels of leadership as President and CEO of AMIA and Chief Science Officer at ONC. His scientific understanding, practical experience, and policymaking in informatics give him a depth of knowledge well suited to the critical challenge of transforming health and health care. Dr. Fridsma will discuss how to navigate budgetary, regulatory, technological, and political challenges in informatics. He will also discuss AMIA’s vision on professional development and leadership and how members can embrace new opportunities for advancement.

Judy Murphy, Panelist, will share her experiences as she transitioned from clinical practice to information technology and nursing informatics, including major shifts in career venues. She went from leading the EHR implementation at a large, non-profit healthcare organization to administering the Meaningful Use Program at the Federal Government to supporting healthcare clients at a large for-profit, publicly-traded IT company. In each venue, she drew from a core set of leadership and management skills which she will share with the audience. Ms. Murphy will also describe the importance of her involvement in organizational work at AMIA and HIMSS and how that contributed to her career advancement.

Chintan Patel, Panelist, will bring his many years of computer science, applied informatics, and industry experience to the panel. He cofounded and led a successful health information solution company, which spun off another successful technology company. He has also developed innovations with a large, multinational technology company. He will discuss the necessary skills and essentials on starting new companies and solutions, delivering value, and leading a business. He will share his insights on fostering innovation, creativity, and entrepreneurship in health information technology amid a rapidly evolving and competitive marketplace.

Hillary Ross, Panelist, with her leadership role in Witt/Kieffer’s Healthcare IT Practice and dedicated focus on informatics, counsels healthcare professionals with a passion for informatics and technology on diverse career paths. She will share a top-level recruiting perspective on informatics careers. There has been an explosion of interest and an expansion of leadership roles in a variety of areas. Informatics professionals may enter their positions from many paths and backgrounds. Ross will describe current expectations among healthcare organizations for various positions, how the roles have evolved, and what types of career paths leaders take. She will explore career advancement, leadership positions, and other issues critical to understanding and thriving in today’s new roles.

Participation Statement

I, Amy Wang, hereby confirm that all panelists listed in this proposal have agreed to participate in this panel. Panelists are aware that there are no travel funds available. Panelists are also aware that the Working Groups are unable to reimburse registration costs.

References

Collaborative Pharming
Mark G. Weiner, MD, FACMI1, Jennifer Boehne, PharmD, MHI, MPH2, Terese Kornet, MSN3, Ross Koppel, PhD, FACMI4
1Lewis Katz School of Medicine at Temple University, Philadelphia, PA
2CHPSO, a Division of the Hospital Quality Institute, Sacramento, CA
3School of Nursing, University of Pennsylvania, Philadelphia, PA
4Sociology Dept and LDI Senior Fellow (Wharton) at the University of Pennsylvania, Philadelphia, PA

Abstract
One EHR activity that significantly affects patient safety is medication ordering and administration, where the interaction and communication among physicians, pharmacists and nurses is crucial. The different information requirements and priorities of these clinical team members and EHR users creates design challenges in the development of a pharmacy interface in both the inpatient and ambulatory environments. Failure to achieve these goals creates a system that, while functional, is neither fully optimized nor conducive to medication safety.

This interactive panel will bring together stakeholders from the physician, usability/workflow analysis, nursing, and pharmacy communities to foster dialog across the groups. Participants will discuss their respective priorities and pain points in terms of addressing information needs, communication, workflow, and data recording requirements when ordering, processing and administering medications. Improved understanding of the other stakeholders’ needs should provide valuable insight into what information is needed at various points in the process, and the engagement of human factors expertise on the panel will help bridge the gap between existing medication ordering and administration interface design. The ideas and findings of this panel will help inform the design of a better pharmacy interface.

Learning Objectives:
1. To increase awareness of the broad nature of real-world errors in medication orders and administration
2. To understand different stakeholder perspectives and priorities related to medication ordering, processing and administration
3. To develop a workflow for medication ordering, processing and administration that better encompasses all stakeholder perspectives
4. To explore informatics solutions that optimize medication ordering, processing and administration

The increasing adoption and utility of Electronic Health Records (EHR) has enabled better patient care through enhanced availability of clinical information, improved decision support, and communication between care team members. However, despite these successes, human factors and usability issues continue to hamper the optimal use of these systems, resulting in well-publicized complaints about electronic information technology and pandemic workarounds. One of the focal points of EHR activity with significant impact on patient safety is medication ordering and administration, where the interaction and communication among physicians, pharmacists and nurses is particularly critical. The different information requirements and priorities of these clinical team members and EHR users creates design challenges in the development of an ideal pharmacy interface in both the inpatient and ambulatory environments. The net result is a system that is functional, but still not fully optimized and continues to impact patient safety and be a source of medication errors.

This interactive panel will bring together stakeholders from the physician, nursing, usability, and pharmacy communities to foster dialog across the groups about their respective priorities and pain points in workflow, information needs, and data recording requirements when ordering, dispensing, and administering medications. The improved understanding of the other stakeholders’ needs should provide valuable insight into what information is needed at various points in the process, and the engagement of human factors expertise on the panel will help bridge the gap between existing order entry and administration interface design and a more optimized design suggested by the ideas raised through the panel.

While common issues such as alert fatigue and wrong patient/wrong drug orders continue to be problematic, this panel is particularly timely as many institutions are facing additional practical issues in EHR interface design that
transcend these traditional topics. Among the vignettes of medication ordering/administration issues with difficult informatics/workflow solutions that we will offer for discussion are as follows:

Errors at Transitions:
- Admission orders are written by the floor team while the patient is in the ED. The patient’s transfer to the med/surg floor is delayed. Should pharmacy deliver the medications to the med/surg floor or to the ED? If meds are delivered to the ED, but the patient is transferred before the meds are delivered to the ED, how do the medications follow the patient?
- A patient is transferred to the med/surg floor from the ED before admit orders are written. Some scheduled medications had been written while the patient is in the ED. How does the nurse know which medications from the ED should be continued on the floor?

Errors in medication timing/dose/discontinuation:
- A physician writes for multiple intravenous medications that are reasonable to order at the same time, but are not compatible for simultaneous administration. Is there a foolproof way for the nurse to know which medications can be given at the same time and, presuming she knows this, how does she know what order to administer the medications?
- An intravenous medication is supposed to be delivered over 2 hours, with a peak blood level drawn after the infusion to help assess the appropriate dose. Because of recurring occlusions, the medicine is delivered over 4 hours, and the peak level is drawn an hour late. How are these common, but often invisible variations addressed in the calculations to determine the correct dose?
- A patient is transitioned from a medication patch for pain control to a PCA. The order for the patch is discontinued after 24 hours of therapy but the initial frequency duration was 72 hours which is when the task reminder for the nurse appears. Patient receives two narcotics for pain control as the nurse forgets to remove the patch and the task does not appear to remind the nurse till 48 hours later.

Errors in visibility of special instructions:
- An epidural catheter is planned to be removed the second post-op day in the morning. Patient is scheduled to receive a dose of subcutaneous heparin at 6am. Instead of placing a medication order to hold the medication the prescriber writes a nurse communication order that is not seen by the nurse. The morning dose of heparin is administered and a six hour delay results before the epidural is discontinued.
- In order for a bladder irrigation to be ready for administration during an OR procedure, the resident writes the order the evening before. In the order the resident types in a note to the pharmacist that the irrigation is for OR the next morning – this note is not visible to the nurse and the nurse hangs the irrigation solution.
- A patient requires intubation due to respiratory failure. The patient is due for an antihypertensive medication – since the patient has an nasogastric tube the nurse crushes the medication and administers the medication via the NGT. The medication is an extended release tablet and the patient experiences hypotension requiring vassopressors. Usual practice on the unit is to not require the prescribers to re-write medications for the correct route (oral verses NGT) so the pharmacy is able to dispense the correct formulation of the medication.

Errors at the interface between IT systems and policy issues:
- Hospital policy requires that certain meds be given for only a few doses before a mandatory automatic discontinuation. How should IT be configured to avoid overuse of medication without leading to unintended gaps in medication administration?
- Automated dispensing machine (ADM) issues. A nurse removes a medication from an ADM as an “override” to handle an urgent situation. How does that override transmit back to the EHR to be processed as a verbal order, and how is appropriate “waste” of unused medication documented?

Interactive Component
The panel will consist of four brief presentations by the panel organizers, each representing physician, nursing, and pharmacy perspectives on the medication ordering, dispensing and administration process as well as insights from a usability/workflow expert on the priorities that have influenced the current state of information technology related to pharmacy processes. After the presentations, the audience will be divided into small groups each with representation from the different stakeholders. The groups will then engage in discussions, facilitated by the panel leads, who help define pitfalls in the medication ordering and administration process, and interface solutions that
could help to address the pitfalls. We expect the participants to continue the dialog outside of the meeting, and work to develop manuscripts based on the work initiated in the panel.

Panelists

Mark Weiner, MD, FACMI - Lewis Katz School of Medicine at Temple University
Dr. Weiner is a practicing general internist who will represent the physician perspective on the panel. His interests relate to medication reconciliation across the transitions of care at the boundaries between inpatient and ambulatory care as well as transitions across inpatient care settings. He is also concerned how issues like the frequency and timing of medications affect patient care, especially for medications where peak and trough values are critical in determining the correct dose.

Jennifer Boehne, PharmD, MHI, MPH -- CHPSO
Dr. Boehne will represent the pharmacy perspective on the panel. She brings insight into the information requirements needed to create a safe and reliable medication use process. Her research focuses on the unintended consequences of various forms of health information technology. She just completed an analysis of incidents involving opiates, benzodiazepines and their reversal agents.

Terese Kornet, MSN, RN -- University of Pennsylvania Health System
Ms. Kornet is the Director of Clinical Nursing Systems at the Hospital of the University of Pennsylvania. She has provided strategic and operation nursing leadership in the development, deployment, re-engineering and integration of clinical information systems to support clinician practices and patient care services. Her special interest related to medication prescribing and administration is BCMA, medication label font/content and weight management in the EHR

Ross Koppel, PhD, FACMI -- University of Pennsylvania
Dr. Koppel is a Senior Fellow at Wharton’s LDI and a Professor of Sociology at the University of Pennsylvania. His work focuses on healthcare IT, and of the interactions of people, computers and workplaces. At Harvard, Dr. Koppel is co-PI on the FDA-funded study of prescribing errors related to patient data displays. In addition, he served as an Internal Evaluator of their project that created the new HIT architecture (SMART) at Harvard. His work combines ethnographic research, extensive statistical analysis, surveys, and usability studies. His work in medical informatics reflects his 46-year career as researcher and sociologist of work and organizations, statistics, ethnographic research, survey research, and medical sociology. He will discuss several issues of visualization of the medication ordering screens--based in part on his recent work with the FDA-funded study of 5 hospital systems and 4 EHRs that focused on ordering screens associated with medication prescribing errors. Koppel will also discuss the interaction of workflow and the several UXs involved in the medication process.

Who Should Attend?

Physicians, nurses, pharmacists, and human factors experts with an interest in interface design, use and workflow related to medication ordering, dispensing, and administration.

All participants have agreed to take part on the panel.
Abstract

As the amount of health-related information being generated and collected by medical organizations continues to increase, the amount of data available to healthcare providers within the electronic health records (EHRs) continues to grow at an unprecedented rate. This vast amount of health-related data poses cognitive and comprehension challenges for medical professionals to understand patients' medical histories and conditions at the point of care and for researchers studying individual- and population-level exposure and outcome data. Information visualization and visual analytics have recently received a significant amount of attention as potential techniques to support the cognitive processes, understanding, and decision-making for anyone gathering and using vast amounts of digital health data.

In this didactic panel, we will review the current state-of-the-art visualization techniques that have been incorporated within EHRs, discuss the areas of healthcare in which better tools and visualization techniques are needed, and examine the need for user evaluations to better understand the effectiveness of different visualization techniques and dashboards in clinical settings. The panelists will also summarize some of the work of the AMIA VIS working group in increasing the awareness of the importance of effective visualization techniques when analyzing large collection of clinical data.

Introduction

Data visualization and visual analytics is an emerging discipline that combines elements of visualization, data mining, human-computer interaction, data analytics, and human cognition to facilitate the exploration, reasoning, and interpretation of complex multi-modal data. The mission of the AMIA Visual Analytics Working Group is to promote the development and validation of new data visualization techniques that can be used to address some of the fundamental data and informatics challenges evident in the healthcare domain. The purpose of supporting and motivating more research and development on new visual analytical methods in healthcare is to make information more insightful, accessible, and actionable to a wide range of participants in the healthcare ecosystem including clinicians, administrators, policy makers, researchers, patients, families, and caregivers. This didactic panel will help by fostering an active, interdisciplinary research community; highlighting impactful applications of the data visualization techniques; developing best practices for development and evaluation.

The didactic panel will consist of three 15-minute presentations focusing on these three
areas/topics:

1. EHR Visualization – Clinical Perspective: V. West and D. Dowding will briefly present a review of existing visualization techniques in healthcare [1] as well as recent studies exploring how to provide visualized data at the point of care to home care nurses to support their decision making for heart failure patients. Panelist will discuss key issues that have arisen from the study including the ability of nurses to comprehend data in a visualized format, how to prioritize what data to visualize, and how to provide it in a format and point in the EHR where it will be of most clinical value.

2. Visualization of EHR data for population health: H. Kharrazi will present a review of existing visualization techniques to support population health management and analysis. As the research director of the Johns Hopkins Center for Population Health IT (CPHIT), Dr. Kharrazi provides direct population-based applications to providers, patients, and payers. Some of his work includes how to provide visualization applications to more effectively illustrate data.

3. EHR Visualization and the AMIA VIS working group: J. Caban as the chair of the AMIA VIS WG will present some of the work been done by the AMIA VIS working group in increasing the awareness of visual analytics in healthcare by organizing tutorials, workshops, design challenges, and meet-ups [2].

Discussion Questions

1. What’s the current state of data visualization within EHR?
2. What are areas of healthcare in which better tools and visualization techniques are needed?
3. How can we better connect clinicians, engineers, developers, and patient in designing effective ways to visualize clinical data?

Conclusion

We proposed a didactic panel to share the current state of EHR visualization with AMIA audiences. All the four panelists (VW, HK, DD, JC) and the moderator (DW) have agreed on participating in this didactic panel.

References:


PCMEDICS: Peace Corps’ Home-Grown EMR

Thomas M. Wilkinson MD (moderator), Shannon Behr DeNoble BA PMP, Kimberley Skrtic RN BSN MHA/INF, Monica D. Charles BA MPH(c), Lyndsey Knauf BA CORII; all from United States Peace Corps, Health Informatics Unit, Washington DC.

Abstract:
In 2016 Peace Corps deployed a cloud-based electronic medical record system to support the healthcare of Volunteers deployed throughout the world. The solution was designed, built, deployed, and supported by the Peace Corps’ Health Informatics Unit (HIU) in Washington, DC. Developers were embedded directly into medical operations at Peace Corps, and the resulting EMR has enjoyed significant support (even admiration) from its many end-users. This panel is a presentation of the HIU’s experience. It highlights some of the unique strengths of the EMR itself, explores issues around security in modern Federal IT systems, and describes the particular emphasis which the HIU has placed on a quality training and support program during go-live and in the year since.

Panel:

Wilkinson: Introduction and overview
[Introduction of panel members]
Peace Corps is mandated by the 1961 Peace Corps Act to provide all necessary healthcare to its active Volunteers. The Peace Corps healthcare system is basically a global primary care network, with an average of 2 Peace Corps Medical Officers (PCMOs) in each country where Peace Corps serves, providing direct care to our 7000 Volunteers in the field. In 2016, Peace Corps launched an EMR named PCMEDICS in 63 countries. It was built in-house, using a modified open source platform, and it came in on time and on budget. This panel is a brief presentation of the Health Informatics Unit’s (HIU) experience designing, building, deploying, and supporting a home-grown EMR.

Some of the fundamental requirements of the EMR were unique. It needed to coordinate frequent communications and care between posts, regional hubs, and Washington; it needed tools for mass immunization; it needed to support the collection of epidemiological data; and it had to meet some very specific legal requirements around sexual assault and medical inventory. We built in SNOMED encoding from the ground up to achieve deep indexing within the charts, to facilitate an information retrieval system. But fundamental to it all, PCMEDICS had to be a robust system able to operate on frail infrastructures, for which we built a distributed, partially redundant data model, redundant access points, and asynchronous synchronizations.

Behr DeNoble: Security
Federal information security requirements are governed by a number of laws, including HIPAA and FISMA, and securing electronic personal health information raises important ethical and professional implications as well. Additionally, Peace Corps is unique in that it tracks comprehensive information about our Volunteers, including financial, safety, and personal data. Implementing sound security practices throughout the development lifecycle of PCMEDICS required that project sponsors, project managers, and security leadership coordinate budgets, security resources, and plans, as well as balance multiple competing needs.

IT security boils down to identifying risk, planning its remediation and mitigation, documenting those plans, and then following through on them. Part of the approach involves putting a reliable perimeter around the system. Controls such as password protection and dual authentication, in addition to data
encryption during communications and at rest, help secure that perimeter. Furthermore, managing security in the cloud requires the cloud host’s active participation in implementing and documenting their security controls. Data integrity, which is related to the accuracy and completeness of data within the system, is supported with e-signatures and user audit trails. Finally, physical security includes failover systems and automatic backups.

At Peace Corps, we baked security into our culture. IT security resources from the Office of the Chief Information Officer were paired with PCMEDICS project teams from the beginning, so that security was considered at each new stage of EMR development. Our strategy also included engaging both stakeholders and team members in security awareness and planning. Every member of the HIU is seen as responsible for information security.

**Skrtic: Rollout education**

The HIU developed and facilitated global PCMEDICS training support to the users of the EMR before, during and after deployment. A series of web-based user scripts that provide detailed, step-by-step instructions with screen shots of clinical processes typical of the Peace Corps health system were released electronically on a weekly basis for two months prior to official PCMEDICS user training. This provided the opportunity for self-paced practice using a PCMEDICS web-based training environment. Immediately prior to go-live, Peace Corps required the attendance of all global providers and health unit ancillary staff at a unique, intensive 7–day EMR training program that taught the providers how to use PCMEDICS safely and effectively across the entire life cycle of the Volunteer, blending clinical care with day-to-day business processes and Peace Corps policies. Each Health Unit trained together as integrated teams, so that providers and assistants could learn and practice the handoff between roles in the new electronic system. Learners were required to show competency in PCMEDICS prior to returning to their posts by passing a standardized, role-based competency assessment. Remediation was immediately offered to individuals faltered in their skills or confidence, prior to their return to country. Within one week of returning to their health units, PCMEDICS went live at that post.

**Charles: Post-deployment education**

For post-deployment support, the providers in the field receive as-needed one-on-one support from the HIU, that is, from the same familiar staff that developed and deployed the EMR and facilitated their initial training. HIU staff provide ongoing education to the field using a Helpdesk ticket tracking system, Skype interactions, and frequent email communications regarding new changes to the EMR system and tips on how to use the system more efficiently. Web-based user scripts and policies related to EMR use are updated and housed on a SharePoint site for all users to access. Post implementation refreshers and new feature training are provided through periodic global Webinars and through live training at annual Peace Corps conferences.

At one-year post deployment we conducted a survey of providers using the system to elicit their ongoing educational needs. The vast majority (89%) of end users surveyed reported that they were feeling either pleased or ecstatic with the EMR overall. The open-ended responses provided further insight. In a qualitative analysis, three themes emerged: (A) requests for advice on how to best manage time and workflow during the day, while using the EMR; (B) a desire to leverage the EMR to improve patient care; (C) a desire to better understand charting standards and quality reviews in the context of an EMR system. Our conclusion was that medical professionals’ education needs evolve over time, starting with the basics (e.g. where to click, when to use certain features), then, after they navigate that learning curve, they want to understand more how to use the tool’s features to improve the quality of their work. Our post-implementation education likewise evolved to focus more on using PCMEDICS in the clinical environment – to help manage time, to improve patient care, and to meet charting standards.
Knauf: Tiered support

The HIU’s decision to build an in-house EMR resulted in the HIU owning its own internal IT support system. We built a three-tiered system. Tier 1 is online self-help, including step-by-step written materials accompanied by screen shots and narrated videos. Some Peace Corps countries have limited bandwidth connections to the internet, so we pay close attention to file sizes and accessibility issues. The advantage of online material is that Peace Corps countries have 24 hour access to help. Tier 2 is the public facing personable support described in more detail below; it is where the Helpdesk ticketing system is routed to, and it kicks in when self-help was not sufficient. Tier 3 is managed by the HIU developers, who respond to needs that rise beyond what Tier 2 support can resolve.

Our Tier 2 support has been pivotal in the ultimate success of PCMEDICS. An internal system for Tier 2 (and, to a lesser extent, Tier 3) humanizes IT support, since familiar staff become the “face” of support for the EMR, rather than anonymous technicians from a vendor. One of the most important benefits that the HIU has garnered from an internal support system is the ability for end users and IT support to forge a working relationship. When end users know that their questions and concerns are being managed by staff who also have a stake in seeing the EMR succeed, those end users tend to come forward with more observations and recommendations that improve the application. In-house support allows HIU to not only answer questions about the EMR, but also to observe our end users’ work flow patterns and sticking points. Those observations then provide us with a tight feedback loop, so support staff not only provide more context-sensitive EMR support, but can also steer enhancement recommendations and guide targeted educational efforts that have a big impact on end users.

Discussion Points:
1. Buy vs Build: a perennial question regarding IT systems. Our experience of building was quite positive.
2. What makes a system operationally robust, even while the underlying infrastructure is unreliable?
3. Data integrity is the high-water mark for FIPS 199 security classification of electronic medical record systems, not data confidentiality.
4. Training and support often occur as afterthoughts, relegated to negotiable line items in a contract; when done poorly, this mindset introduces risk to the patient and to the project, and it impedes adoption, integration, and socio-technical evolution.
Quality Assurance of NCI Thesaurus by Mining Structural-Lexical Patterns

Rashmie Abeysinghe¹, Michael A. Brooks, MD², Jeffery Talbert, PhD³, Licong Cui, PhD¹,⋆

¹Department of Computer Science, University of Kentucky, Lexington, KY
²Departments of Radiology and Medicine, University of Kentucky, Lexington, KY
³Institute for Pharmaceutical Outcomes and Policy, University of Kentucky, Lexington, KY

Abstract

Quality assurance of biomedical terminologies such as the National Cancer Institute (NCI) Thesaurus is an essential part of the terminology management lifecycle. We investigate a structural-lexical approach based on non-lattice subgraphs to automatically identify missing hierarchical relations and missing concepts in the NCI Thesaurus. We mine six structural-lexical patterns exhibiting in non-lattice subgraphs: containment, union, intersection, union-intersection, inference-contradiction, and inference union. Each pattern indicates a potential specific type of error and suggests a potential type of remediation. We found 809 non-lattice subgraphs with these patterns in the NCI Thesaurus (version 16.12d). Domain experts evaluated a random sample of 50 small non-lattice subgraphs, of which 33 were confirmed to contain errors and make correct suggestions (33/50 = 66%). Of the 25 evaluated subgraphs revealing multiple patterns, 22 were verified correct (22/25 = 88%). This shows the effectiveness of our structural-lexical-pattern-based approach in detecting errors and suggesting remediations in the NCI Thesaurus.

Introduction

Biomedical terminologies and ontologies serve as a knowledge source for many biomedical applications, including natural language processing applications and decision support systems¹. Quality issues in terminologies, if not addressed, can affect the quality of all downstream information systems relying on them as a knowledge source². Terminology Quality Assurance (TQA) strives to estimate and enhance the quality of terminologies by improving consistency, coverage and completeness, non-redundancy and clarity³. However, it is labor-intensive and time-consuming to discover errors or inconsistencies by manual review of large biomedical terminologies. Therefore, automating TQA has been an active area of research⁴.

Developed and maintained by the National Cancer Institute (NCI), the NCI Thesaurus (NCIt) is a reference terminology used in an increasing number of NCI and other systems⁵, ⁶. It contains over 100,000 concepts related to cancer research, including cancer-related diseases, findings and abnormalities; anatomy; agents, drugs and chemicals; genes and so on⁷. Given the sheer size of the NCIt, it is unavoidable that errors may be introduced in its development, update, and maintenance phases. Moreover, it is impractical for human experts to manually review the terminology to discover quality issues such as missing concepts, concept redundancies, and missing hierarchical relations. Automatic approaches to quality assurance of the NCIt are highly desirable to provide human experts with error candidates for review and verification.

In this paper, we develop an automatic approach to detecting missing hierarchical IS-A relations and missing concepts in the NCIt based on non-lattice subgraphs, which were initially introduced for auditing SNOMED CT⁸. We investigate six structural-lexical patterns exhibiting in non-lattice subgraphs in the NCIt, with each pattern indicating a certain type of potential error and suggesting a potential correction. Human experts reviewed a random subset of non-lattice subgraphs automatically detected using this approach to confirm the uncovered errors and suggested corrections.

1 Background

1.1 NCI Thesaurus (NCIt)

The NCIt is a biomedical terminology for cancer research, covering vocabulary for clinical care, translational and basic research, and public information and administrative activities⁵, ⁶. It was first published in 2000 with the intention to facilitate data sharing and interoperability by different NCI components. Concepts in NCIt are hierarchically organized in 19 domains, including Abnormal Cell; Anatomic Structure, System, or Substance; Biological Process; Disease,
Disorder or Finding; Drug, Food, Chemical or Biomedical Material, Gene, Gene Product, Molecular Abnormality, and Organism. The version 16.12d of NCIt contains over 118,000 concepts.

The NCIt was built using Ontylog, a description logic explicitly for building large complex terminologies. It is published in several formats including Ontylog XML, Web Ontology Language (OWL), and flat files. The NCIt also has defined and inferred versions. The defined version contains the assertions about each concept by the terminology editors. The inferred version includes additional assertions and classifications inferred by DL classifiers. In this paper, we used the inferred version of the NCIt to perform our pattern-based error detection and correction.

1.2 Quality Assurance of Biomedical Terminologies

Researchers have proposed various approaches to auditing biomedical terminologies, such as the NCIt and SNOMED CT. Min et al. proposed the abstraction networks (AbNs) approach to audit NCIt based on area taxonomies and partial-area taxonomies, where area taxonomies are groups of concepts that have exactly same roles, and partial-area taxonomies are further divisions of areas so that they are structurally uniform and singly-rooted. Ochs et al. introduced subject-based AbN methods that summarize a subhierarchy rooted at a subject concept within a large hierarchy, and tribal-based AbN methods that are based on a subhierarchy rooted at a child of a hierarchy root to audit SNOMED CT. Verspoor et al. introduced an automated method for identifying univocality violations in Gene Ontology. Zhang et al. proposed a lattice-based structural auditing method to exhaustively detect non-lattice pairs in SNOMED CT. Cui et al. presented a big data approach to perform lattice-based terminology quality assurance on SNOMED CT. Agrawal et al. used a combination of positional similarity sets and structural indicators to identify modeling inconsistencies in SNOMED CT. Bodenreider introduced a method to identify missing hierarchical relations in SNOMED CT from logical definitions based on the lexical features of concept names. Ceusters et al. assessed the conformity of NCIt to widely accepted principles in terminology construction and ontology building. Mougin et al. presented a semantic web technology method for quality assurance of NCIt. Zhang et al. proposed a lattice-based structural auditing method to exhaustively detect non-lattice pairs in SNOMED CT. Cui et al. presented a big data approach to perform lattice-based terminology quality assurance on SNOMED CT. Agrawal et al. used a combination of positional similarity sets and structural indicators to identify modeling inconsistencies in SNOMED CT. Bodenreider introduced a method to identify missing hierarchical relations in SNOMED CT from logical definitions based on the lexical features of concept names. Ceusters et al. assessed the conformity of NCIt to widely accepted principles in terminology construction and ontology building. Mougin et al. presented a semantic web technology method for quality assurance of NCIt. Zhang et al. introduced a topological pattern-based method to recommend new concepts to include to NCIt.

More recently, Cui et al. have introduced a hybrid structural-lexical method based on non-lattice subgraphs for scalable and systematic discovery of missing hierarchical relations and concepts in SNOMED CT. Four lexical patterns in non-lattice subgraphs were proposed for error detection and correction in SNOMED CT. In this paper, we apply these four patterns to NCIt and introduce two new patterns to identify errors and suggest corrections. To provide better readability, we review the definitions of the four patterns proposed for SNOMED CT in the Methods section and illustrate the patterns with examples in the NCIt.

2 Methods

Our approach leverages both structural and lexical information in the NCIt to systematically detect potential errors and automatically suggest remediations. Firstly, we identify all non-lattice subgraphs in NCIt. Secondly, we mine structural and lexical patterns in the non-lattice subgraphs, where each pattern indicates a potential missing hierarchical relation or missing concept in the NCIt. Finally, human domain experts evaluate a randomly selected sample of the potential errors detected, as well as the proposed remediation. We used the 16.12d version of the NCIt in this work.

2.1 Detecting Non-lattice Subgraphs

Non-lattice pairs. From a structural point of view, lattice is a desirable property for a well-formed terminology. A terminology is a lattice if any two concepts in the terminology have a unique maximal shared descendant, as well as a unique minimal shared ancestor. A pair of concepts is known as a non-lattice pair, if the two concepts have more than one maximal shared descendant (alternatively minimal shared ancestor). A non-lattice pair generates a graph fragment with the nodes (or concepts) between the concept pair and the maximal shared descendants. There could be multiple non-lattice pairs which possess the same maximal shared descendants. In this case, it is not economical to examine each of these separately. If non-lattice pairs possessing the same maximal shared descendants are added together, this is also not economical since there might be concepts with ancestor-descendant relationship, which cause redundant analysis. Therefore the notion of non-lattice subgraphs has been introduced to facilitate effective analysis.
Non-lattice subgraphs. A non-lattice subgraph \( p = (c_1, c_2) \) can be obtained by first reversely computing the minimal common ancestors of the maximal common descendants, \( mcd(p) \); then accumulating the concepts and the edges between (including) any concept in \( mca(mcd(p)) \) and any concept in \( mcd(p) \). The reverse computation obtains the minimal concepts sharing the same maximal common descendants to avoid redundant analysis. The minimal concepts \( mca(mcd(p)) \) are called the upper bounds of the non-lattice subgraph, and \( mcd(p) \) are called the lower bounds. The size of a non-lattice subgraph is the number of concepts it contains. For instance, Figure 1A shows a non-lattice subgraph of size 6 in the NCIt, where Stage III Pharyngeal Cancer and Nasopharyngeal Carcinoma are the concepts in the upper bounds, and Stage III Nasopharyngeal Carcinoma AJCC v6 and Stage III Nasopharyngeal Carcinoma are the concepts in the lower bounds.

In this work, we first parse the NCIt distribution file “ThesaurusInferred.owl” to extract all the concepts and their labels, as well as hierarchical IS-A relations (i.e., Child-Parent relations). Then we leverage the computational pipeline implemented in previous work \( ^8 \) to exhaustively detect non-lattice subgraphs in the NCIt. Each resulting non-lattice subgraph consists of five components: concepts in the lower bounds, concepts in the upper bounds, concepts in the non-lattice subgraph, IS-A relationships in the non-lattice subgraph, and the size of the non-lattice subgraph.

2.2 Mining Structural and Lexical patterns in Non-lattice Subgraphs

Since manual review of non-lattice subgraphs to discover potential errors is labor-intensive and time-consuming, we further take into account of the lexical information (concept labels) to automatically identify structural and lexical patterns in non-lattice subgraphs. Each pattern indicates certain type of errors and suggests a potential remediation.

For lexical information, we consider the label of a concept as a set of words in lower case. For example, the concept label Stage III Pharyngeal Cancer is considered as a set of words \{stage, iii, pharyngeal, cancer\}. For structural information, given a non-lattice subgraph, we use \( U_i \) to denote the set of words for a certain concept in the upper bounds, and \( L_j \) to denote the set of words for a certain concept in the lower bounds.

We define six patterns taking into account of such lexical and structural information in the NCIt: Containment, Union, Intersection, Union-Intersection, Inference-Contradiction, and Inference-Union. The Containment, Union, Intersection, and Union-Intersection patterns were initially proposed in previous work \( ^8 \) for SNOMED CT. The Inference-Contradiction and Inference-Union patterns are newly proposed in this work, incorporating inference into the structural and lexical information.

2.2.1 Containment

A non-lattice subgraph is defined as exhibiting a containment pattern \( ^8 \), if the set of words for one concept \( U_i \) in the upper bounds is contained in the set of words for another concept \( U_j \) in the upper bounds, or the set of words for one concept \( L_i \) in the lower bounds is contained in the set of words for another concept \( L_j \) in the lower bounds. That
is, $U_i \subset U_j$, or $L_i \subset L_j$. This pattern may suggest a missing IS-A relation between the two concepts in the upper bounds (or lower bounds), that is, $U_i$ IS-A $U_j$ (or $L_j$ IS-A $L_i$). Consider the example in Figure 1A, $L_1 = \{\text{stage, iii, nasopharyngeal, carcinoma}\}$ in the lower bounds is contained in $L_2 = \{\text{stage, iii, nasopharyngeal, carcinoma, ajcc, v6}\}$ in the lower bounds. This indicates a potential missing hierarchical relation: $L_2$ IS-A $L_1$, with $L_2$ more specific than $L_1$. The suggested correction is to add the relation Stage III Nasopharyngeal Carcinoma AJCC v6 is a subclass of Stage III Nasopharyngeal Carcinoma (highlighted as a red edge in Figure 1B).

For the containment pattern, we do not consider non-lattice subgraphs with concepts involving negation words such as no, not, without, absence, since that would incorrectly suggest a missing hierarchical relation between a concept with negation and a concept without negation.

### 2.2.2 Union

A non-lattice subgraph is defined as exhibiting a union pattern, if the union of the sets of words for two concepts $U_i$ and $U_j$ in the upper bounds is equal to the set of words for some concept $L_k$ in the lower bounds, that is, $U_i \cup U_j = L_k$. This pattern may suggest a missing IS-A relation between other concepts in the lower bounds and $L_k$. For instance, in Figure 2A, the union of $U_1 = \{\text{testicular, non-seminomatous, germ, cell, tumor}\}$ and $U_2 = \{\text{malignant, testicular, germ, cell, tumor}\}$ in the upper bounds is equal to $L_1 = \{\text{malignant, testicular, non-seminomatous, germ, cell, tumor}\}$ in the lower bound. This indicates a potential missing IS-A relation between the other concept Childhood Testicular Yolk Sac Tumor in the lower bounds and $L_1$. That is, Childhood Testicular Yolk Sac Tumor IS-A Malignant Testicular Non-Seminomatous Germ Cell Tumor (highlighted as a red edge in Figure 2B).

**Figure 2:** (A) A non-lattice subgraph exhibiting the Union pattern. (B) The suggested remediation of the non-lattice subgraph in (A): Childhood Testicular Yolk Sac Tumor IS-A Malignant Testicular Non-Seminomatous Germ Cell Tumor.

### 2.2.3 Intersection

A non-lattice subgraph is defined as exhibiting an intersection pattern, if the intersection of the set of words for two concepts $L_i$ and $L_j$ in the lower bounds is equal to the set of words for some concept $U_k$ in the upper bound, that is, $L_i \cap L_j = U_k$. This pattern may suggest a missing IS-A relation between $U_k$ and other concepts in the upper bounds. For instance, in Figure 3A, the intersection of $L_1 = \{\text{splenic, t, lymphoblastic, lymphoma}\}$ and $L_2 = \{\text{splenic, b, lymphoblastic, lymphoma}\}$ in the lower bounds is equal to $U_1 = \{\text{splenic, lymphoblastic, lymphoma}\}$ in the upper bound. This indicates a potential missing IS-A relation between $U_1$ and the other concept Aggressive Non-Hodgkin Lymphoma in the upper bound. That is, Splenic Lymphoblastic Lymphoma IS-A Aggressive Non-Hodgkin Lymphoma (the red edge in Figure 3B).
2.2.4 Union-Intersection

A non-lattice subgraph is defined as exhibiting an union-intersection pattern, if the union of the set of words for two concepts $U_i$ and $U_j$ in the upper bounds is equal to the intersection of the set of words for two concepts $L_s$ and $L_t$ in the lower bounds, that is, $U_i \cup U_j = L_s \cap L_t$. This pattern may suggest a missing intermediary concept between the two concepts ($U_i$ and $U_j$) in upper bounds and the two concepts ($L_s$ and $L_t$) in the lower bounds. For example, in Figure 4A, the union of $U_1 = \{\text{localized, carcinoma}\}$ and $U_2 = \{\text{adult, liver, carcinoma}\}$ is equal to the intersection of $L_1 = \{\text{localized, non-resectable, adult, liver, carcinoma}\}$ and $L_2 = \{\text{localized, resectable, adult, liver, carcinoma}\}$, that is, $U_1 \cup U_2 = L_s \cap L_t = \{\text{localized, adult, liver, carcinoma}\}$. This indicates a potential missing concept Localized Adult Liver Carcinoma (green node in Figure 4B), which represents the features that are common to $L_s$ and $L_t$ in the lower bounds and inherited from $U_i$ and $U_j$ in the upper bounds.

It is worth noting that if $U_i \cup U_j = L_s \cap L_t$ happens to be equal to $L_s$ or $L_t$, then the non-lattice subgraph falls into the union pattern as well; if it happens to be equal to $U_i$ or $U_j$, then the non-lattice subgraph falls into the intersection pattern as well. In such cases, the suggestion for union pattern or intersection pattern is adopted, since no intermediary concept is needed.

2.2.5 Inference-Contradiction

Given a non-lattice subgraph $G$, we define two types of concept pairs appearing in $G$: related and unrelated. A pair of concepts $(C_i, C_j)$ in $G$ is called related if $C_i$ is a subclass or descendant of $C_j$; otherwise, $(C_i, C_j)$ is called unrelated. For instance, in Figure 5A, the concept pair (Anaplastic Cell, Neoplastic Large Cell) is related; while the concept pair (Anaplastic T-Lymphocyte, Neoplastic Large T-Lymphocyte) is unrelated.

Suppose $R$ is the set of all related concept pairs in $G$, and $\overline{R}$ is the set of all unrelated concept pairs in $G$. We perform a set-difference-based inference to derive contradiction in the following way. For each related concept pair $(B_d, B_a)$ in...
If $B_d - (B_d \cap B_a) \neq \emptyset$ and $B_a - (B_d \cap B_a) \neq \emptyset$, an inferred term pair \( (B_d - (B_d \cap B_a), B_a - (B_d \cap B_a)) \) can be derived. Similarly, for each unrelated concept pair \( (N_i, N_j) \) in $\mathcal{R}$, if $N_i - (N_i \cap N_j) \neq \emptyset$ and $N_j - (N_i \cap N_j) \neq \emptyset$, an inferred term pair \( (N_i - (N_i \cap N_j), N_j - (N_i \cap N_j)) \) can be derived. If there exists some common term pair that can be derived from both a related pair in $\mathcal{R}$ and an unrelated pair in $\mathcal{R}$, we say that the non-lattice subgraph is exhibiting an inference-contradiction pattern. For instance, in Figure 5A, the related concept pair \( \text{Anaplastic Cell, Neoplastic Large Cell} \) derives a term pair \( \text{Anaplastic, Neoplastic Large} \); while the unrelated concept pair \( \text{Anaplastic T-Lymphocyte, Neoplastic Large T-Lymphocyte} \) derives the same term pair \( \text{Anaplastic, Neoplastic Large} \). This pattern may suggest a potential missing IS-A relation between the unrelated concept pair: \( \text{Anaplastic T-Lymphocyte IS-A Neoplastic Large T-Lymphocyte} \) (the red edge in Figure 5B).

**Figure 5:** (A) A non-lattice subgraph exhibiting the Inference-Contradiction pattern. (B) The suggested remediation of the non-lattice subgraph in (A): \( \text{Anaplastic T-Lymphocyte IS-A Neoplastic Large T-Lymphocyte} \).

If $U_s \cup (L_i \cap L_j) = L_k$, this may suggest a missing IS-A relation between other concepts in the lower bounds and $L_k$. For instance, in Figure 6A, the intersection of $L_1 = \{\text{lungs, mucinous, adenocarcinoma}\}$ and $L_2 = \{\text{mucinous, bronchioloalveolar, carcinoma}\}$ in the lower bounds is $\{\text{mucinous}\}$, whose union with $U_1 = \{\text{lungs, adenocarcinoma}\}$ is equal to $L_1 = \{\text{lungs, mucinous, adenocarcinoma}\}$. This indicates a potential missing IS-A relation between the other concept $L_2$ in the lower bounds and $L_1$. That is, \( \text{Mucinous Bronchioloalveolar Carcinoma IS-A Lung Mucinous Adenocarcinoma} \) (the red edge in Figure 6B).

**Figure 6:** (A) A non-lattice subgraph exhibiting the Union, Inference-Contradiction, and Inference-Union patterns. (B) The suggested remediation of (A): \( \text{Mucinous Bronchioloalveolar Carcinoma IS-A Lung Mucinous Adenocarcinoma} \).

### 2.2.6 Inference-Union

A non-lattice subgraph is defined as exhibiting an inference-union pattern, if the union of the set of words for some concept $U_s$ in the upper bounds and the intersection of the set of words for two concepts $L_i$ and $L_j$ in the lower bounds is equal to the set of words for some concept $L_k$ in the lower bounds, that is, $U_s \cup (L_i \cap L_j) = L_k$. This may suggest a missing IS-A relation between other concepts in the lower bounds and $L_k$. For instance, in Figure 6A, the intersection of $L_1 = \{\text{lungs, mucinous, adenocarcinoma}\}$ and $L_2 = \{\text{mucinous, bronchioloalveolar, carcinoma}\}$ in the lower bounds is $\{\text{mucinous}\}$, whose union with $U_1 = \{\text{lungs, adenocarcinoma}\}$ is equal to $L_1 = \{\text{lungs, mucinous, adenocarcinoma}\}$. This indicates a potential missing IS-A relation between the other concept $L_2$ in the lower bounds and $L_1$. That is, \( \text{Mucinous Bronchioloalveolar Carcinoma IS-A Lung Mucinous Adenocarcinoma} \) (the red edge in Figure 6B).
2.2.7 Non-lattice Subgraphs with Multiple Patterns

We also investigate non-lattice subgraphs exhibiting multiple patterns among the above-mentioned six patterns. For instance, the non-lattice subgraph in Figure 1A exhibits both containment and inference-union patterns, and both patterns suggest a missing IS-A relation: *Stage III Nasopharyngeal Carcinoma AJCC v6 IS-A Stage III Nasopharyngeal Carcinoma*. The non-lattice subgraph in Figure 6A is following three patterns: union, inference-contradiction, and inference-union, and all these patterns suggest a missing relation between *Mucinous Bronchioloalveolar Carcinoma* and *Lung Mucinous Adenocarcinoma*.

2.3 Evaluation

For evaluation, we focus on small non-lattice subgraphs (size of 4, 5, and 6) due to two reasons. One is that small ones are relatively easy to visually inspect by domain experts. The other reason is that small non-lattice subgraphs may be contained in larger ones, and fixing errors in small ones will automatically eliminate the same errors propagated in the larger ones (although there might be other errors in the larger ones).

To evaluate the performance of applying different patterns in small non-lattice subgraphs to automatically detect real errors in NCIt and suggest corrections, we randomly selected 25 non-lattice subgraphs with a single pattern, and 25 ones with multiple patterns, respectively. These 50 sample non-lattice subgraphs as well as their suggested remediations were rendered in scalable vector graphics and provided to experts (authors MAB and JT) for evaluation. MAB evaluated cancer-related samples, and JT evaluated drug-related samples.

### Table 1: Number of non-lattice subgraphs exhibiting each of the 24 patterns.

<table>
<thead>
<tr>
<th>Pattern</th>
<th>No. of non-lattice subgraphs</th>
<th>No. of small non-lattice subgraphs (size of 4-6)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Containment</td>
<td>159</td>
<td>84</td>
</tr>
<tr>
<td>Union</td>
<td>7</td>
<td>3</td>
</tr>
<tr>
<td>Intersection</td>
<td>430</td>
<td>166</td>
</tr>
<tr>
<td>Union-Intersection</td>
<td>24</td>
<td>2</td>
</tr>
<tr>
<td>Inference-Contradiction</td>
<td>37</td>
<td>3</td>
</tr>
<tr>
<td>Inference-Union</td>
<td>21</td>
<td>12</td>
</tr>
<tr>
<td>Inference-Contradiction, Containment</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Inference-Union, Containment</td>
<td>19</td>
<td>13</td>
</tr>
<tr>
<td>Inference-Contradiction, Inference-Union</td>
<td>12</td>
<td>9</td>
</tr>
<tr>
<td>Intersection, Containment</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Intersection, Inference-Contradiction</td>
<td>33</td>
<td>9</td>
</tr>
<tr>
<td>Union, Inference-Union</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Inference-Contradiction, Union-Intersection</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Intersection, Inference-Union</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Inference-Union, Inference-Contradiction, Containment</td>
<td>14</td>
<td>7</td>
</tr>
<tr>
<td>Intersection, Inference-Union, Containment</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Union, Inference-Union, Inference-Contradiction</td>
<td>7</td>
<td>4</td>
</tr>
<tr>
<td>Union, Intersection, Inference-Union</td>
<td>13</td>
<td>12</td>
</tr>
<tr>
<td>Intersection, Inference-Contradiction, Containment</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Union, Union-Intersection, Inference-Union, Containment</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>Union, Intersection, Inference-Union, Inference-Contradiction</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>Intersection, Inference-Contradiction, Containment, Union-Intersection</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Intersection, Inference-Union, Inference-Contradiction, Containment</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Union, Union-Intersection, Inference-Union, Inference-Contradiction, Containment</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>Total</td>
<td>809</td>
<td>337</td>
</tr>
</tbody>
</table>

3 Results

3.1 Non-lattice Subgraphs Exhibiting Structural and Lexical Patterns

A total of 8,143 non-lattice subgraphs were identified in the NCIt (version 16.12d), among which 809 exhibits a single pattern or multiple patterns. Of these 809 non-lattice subgraphs, 678 were found exhibiting a single lexical pattern, 131 exhibiting multiple patterns. Of the 809 non-lattice subgraphs, 337 were small ones (size of 4, 5, and 6), among which 270 exhibited a single pattern, 67 exhibited multiple patterns. Table 1 shows the numbers of both non-lattice
subgraphs and small non-lattice subgraphs exhibiting different combinations of patterns (six single pattern, eighteen multiple patterns). For instance, there were 159 non-lattice subgraphs exhibiting a single containment pattern (the first row in Table 1), and 5 non-lattice subgraphs exhibiting multiple patterns: union, union-intersection, inference-union, inference-contradiction, and containment (the last row in Table 1). Figure 7 shows an example of non-lattice subgraph with these five patterns. For the 678 non-lattice subgraphs with a single pattern, the intersection pattern accounted for the largest proportion (430 non-lattice subgraphs). For the 131 non-lattice subgraphs with multiple patterns, the intersection and inference-contradiction patterns accounted for the largest proportion (33 non-lattice subgraphs).

Figure 7: (A) A non-lattice subgraph exhibiting five patterns: Union, Union-Intersection, Inference-Union, Inference-Contradiction, and Containment. (B) The suggested remediation of the non-lattice subgraph in (A): Sustained Release Buccal Tablet Dosage Form IS-A Sustained Release Tablet Dosage Form.

Table 2: Numbers of small non-lattice subgraphs evaluated by domain experts in terms of patterns, as well as numbers of correct suggestions verified by experts.

<table>
<thead>
<tr>
<th>Pattern</th>
<th>No. of non-lattice subgraphs</th>
<th>No. of non-lattice subgraphs with correct suggestions</th>
<th>Correction rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Containment</td>
<td>7</td>
<td>6</td>
<td>85.7%</td>
</tr>
<tr>
<td>Union</td>
<td>1</td>
<td>1</td>
<td>100%</td>
</tr>
<tr>
<td>Intersection</td>
<td>14</td>
<td>2</td>
<td>14.3%</td>
</tr>
<tr>
<td>Union-Intersection</td>
<td>1</td>
<td>1</td>
<td>100%</td>
</tr>
<tr>
<td>Inference-Contradiction</td>
<td>1</td>
<td>1</td>
<td>100%</td>
</tr>
<tr>
<td>Inference-Union</td>
<td>1</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td>Inference-Contradiction, Containment</td>
<td>1</td>
<td>1</td>
<td>100%</td>
</tr>
<tr>
<td>Inference-Union, Containment</td>
<td>4</td>
<td>3</td>
<td>75%</td>
</tr>
<tr>
<td>Inference-Contradiction, Inference-Union</td>
<td>3</td>
<td>3</td>
<td>100%</td>
</tr>
<tr>
<td>Intersection, Containment</td>
<td>1</td>
<td>1</td>
<td>100%</td>
</tr>
<tr>
<td>Intersection, Inference-Contradiction</td>
<td>3</td>
<td>2</td>
<td>66.7%</td>
</tr>
<tr>
<td>Inference-Union, Inference-Contradiction, Containment</td>
<td>2</td>
<td>2</td>
<td>100%</td>
</tr>
<tr>
<td>Intersection, Inference-Union, Containment</td>
<td>1</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td>Union, Inference-Union, Inference-Contradiction</td>
<td>2</td>
<td>2</td>
<td>100%</td>
</tr>
<tr>
<td>Union, Intersection, Inference-Union</td>
<td>4</td>
<td>4</td>
<td>100%</td>
</tr>
<tr>
<td>Union, Union-Intersection, Inference-Union, Containment</td>
<td>2</td>
<td>2</td>
<td>100%</td>
</tr>
<tr>
<td>Union, Intersection, Inference-Union, Inference-Contradiction</td>
<td>1</td>
<td>1</td>
<td>100%</td>
</tr>
<tr>
<td>Union, Union-intersection, Inference-Union, Inference-Contradiction, Containment</td>
<td>1</td>
<td>1</td>
<td>100%</td>
</tr>
<tr>
<td>Total</td>
<td>50</td>
<td>33</td>
<td>66%</td>
</tr>
</tbody>
</table>

3.2 Evaluation

Of the 50 sample non-lattice subgraphs evaluated by domain experts, 33 were verified to contain errors and make correct suggestions (33/50 = 66%). Among these 33 correct cases, 32 were missing hierarchical relations and one was a missing intermediary concept. Table 2 presents the numbers of evaluated non-lattice subgraphs exhibiting each combination of patterns, and the numbers of correct suggestions confirmed by domain experts. Of the 25 evaluated non-lattice subgraphs with a single pattern, 11 were verified correct (11/25 = 44%). Of the 25 evaluated non-lattice
subgraphs with multiple patterns, 22 were verified correct \((22/25 = 88\%)\). This illustrates that non-lattice subgraphs with multiple patterns achieved a better performance than those with a single pattern in terms of the correction rate.

4 Discussion

In this paper, we investigated non-lattice subgraphs in NCIt based on six structural and lexical patterns, with each pattern automatically suggesting a potential missing hierarchical relation or missing concept. Our pattern-based approach leveraging both structural and lexical information is scalable and applicable to other terminologies for quality assurance work, since it generally takes concepts (as well as concept labels) and hierarchical relations of a terminology as the input, and generates erroneous non-lattice subgraphs and potential corrections as the output.

Analysis of failure cases. For the single-pattern non-lattice subgraphs evaluated in Table 2, the suggestions made by the intersection pattern turned out to have a low correction rate \((2/14 = 14.3\%)\). Figure 8A shows a non-lattice subgraph exhibiting the intersection pattern: \(\{\text{gestational, choriocarcinoma}\} \cap \{\text{ovarian, choriocarcinoma}\} = \{\text{choriocarcinoma}\}\). However, its automatic suggestion in Figure 8B is not correct. That is, \textit{Choriocarcinoma} is NOT a subclass of \textit{Malignant Female Reproductive System Neoplasm}, since \textit{Choriocarcinoma} can be a malignant female reproductive system neoplasm, but it can also arise in the male testis. Another example of wrongly suggested case by the containment pattern is: \(\{\text{osteoma}\} \subset \{\text{osteoid, osteoma}\}\). However, despite the similarity in names, \textit{Osteoid Osteoma} and \textit{Osteoma} are two completely different types of tumor, and \textit{Osteoid Osteoma} is thus NOT a subclass of \textit{Osteoma}.

![Diagram](image)

**Figure 8:** (A) A non-lattice subgraph exhibiting an Intersection pattern. (B) The wrongly suggested remediation of (A).

Comparison with previous work. The hybrid approach to mining structural-lexical patterns in non-lattice subgraphs was initially proposed in previous work\(^8\) for quality assurance of SNOMED CT, where four patterns were studied: containment, union, intersection, and union-intersection. In this paper, we applied these four patterns to NCIt, and further proposed two new patterns with implicit inference: inference-contradiction and inference-union. In addition, only single-pattern non-lattice subgraphs were investigated in previous work\(^8\), while in this paper, we not only studied non-lattice subgraphs with a single pattern, but also those with multiple patterns. Non-lattice subgraphs in NCIt with multiple patterns turned out to have a higher error detection and correction rate than those with a single pattern (see Table 2). For SNOMED CT\(^8\), the overall correction rate of the four patterns (by single pattern) was 59\%. For the NCIt in this paper, the overall correction rate of the six patterns (by both single pattern and mixed patterns) is 66\%.

Limitations and future work. A limitation of this work is that we only evaluated small non-lattice subgraphs (size of 4, 5, 6) for experts’ ease to review and validate. It would be interesting to further examine larger-size non-lattice subgraphs for evaluation. In addition, our evaluation was limited on the number of samples and only one domain expert was involved. We plan to evaluate more samples by multiple experts in the future. Another limitation of this work is that the list of negation words used in detecting the containment pattern was manually constructed based on our observation and previous experience. In the future, we expect to use a resource like NegEx for this purpose. The followings are a couple of directions for additional future work. When defining different patterns, we only used the concept labels for lexical information. We plan to take into account of the concept synonyms to complement concept labels, which may obtain more non-lattice subgraphs with patterns. Note that there are non-lattice subgraphs that may be erroneous but are not exhibiting any of the six patterns studied in this paper. New patterns or approaches are needed to uncover potential errors in such non-lattice subgraphs.
5 Conclusions

In this paper, we investigated a hybrid approach to identifying potential errors in the NCI Thesaurus and automatically suggesting remediations, by mining structural and lexical patterns in non-lattice subgraphs. This approach proved an effective way for error detection and correction in the NCI Thesaurus, and is applicable to other biomedical terminologies for quality assurance purposes.

Acknowledgement

This work was supported by the National Institutes of Health (NIH) National Center for Advancing Translational Sciences through grant UL1TR001998 and National Science Foundation through grant IIS-1657306.

References

17. Bodenreider O. Identifying missing hierarchical relations in SNOMED CT from logical definitions based on the lexical features of concept names. International Conference on Biomedical Ontology and BioCreative. 2016;1747.
Building an Informed Consent Tool Starting with the Patient: The Patient-Centered Virtual Multimedia Interactive Informed Consent (VIC)

Fuad Abujarad, MS, PhD¹, Sandra Alfano, Pharm.D.¹, Tiffani J. Bright, PhD² Sneha Kannoth, MS³, Nicole Grant¹, Matthew Gueble¹, Peter Peduzzi PhD³, Geoffrey Chupp MD¹

¹Yale University School of Medicine, New Haven, CT; ²Agency for Healthcare Research and Quality, Rockville, MD; ³Yale University School of Public Health, New Haven, CT

Abstract

Patient safety and quality of care are at risk if the informed consent process does not emphasize patient comprehension. In this paper, we describe how we designed, developed, and evaluated an mHealth tool for advancing the informed consent process. Our tool enables the informed consent process to be performed on tablets (e.g., iPads) utilizing virtual coaching with text-to-speech automated translation as well as an interactive multimedia elements (e.g., graphics, video clips, animations, presentations, etc.). We designed our tool to enhance patient comprehension and quality of care, while improving the efficiency of obtaining patient consent. We present the User-Centered Design approach we adopted to develop the tool and the results of the different methods we used during the development of the tool. Also, we describe the results of the usability study which we conducted to evaluate the effectiveness, efficiency, and user satisfaction with our mHealth App to enhance the informed consent process. Using the UCD approach we were able to design, develop, and evaluate a highly interactive mHealth App to deliver the informed consent process.

Keywords— informed consent; mHealth; user-centered design;

Introduction

Objective: Informed consent (IC) is essential for the ethical conduct of research and medical treatment. The overarching goal of the IC process is to guarantee that the patient acquires a sound understanding of the purpose, risks, and methodology of a clinical trial and/or medical procedure¹,². For patients to fully understand the content of the IC process, the informed consent document should clearly explain the purpose, process, risks, benefits and alternatives to medical procedures or clinical research, while stressing the patient’s rights and responsibilities³. If the IC process achieves patient comprehension, then healthcare costs and patient safety risk can likely be deterred⁴. In regards to the expert consensus on IC tools, the Joint Commission (2007) reported that, “among patients who sign an IC form, 44% do not understand the nature of the procedure to be performed, and 60-70% did not read or understand the information contained in the form.” Thus, the Joint Commission urges reform given the poor potential of the IC process in achieving patient understanding.

Background: Studies have shown that the lack of sufficient comprehension of information included in the IC document negatively impacts patient safety and quality of care¹,². Patients want to be involved in making decisions about their own care. Currently, the standard consent process (paper and/or electronic) does not always guarantee patient comprehension or improve the quality of care. The real challenge is that many patients lack the means to truly comprehend existing IC documents⁵. As a result, many patients are consenting without fully comprehending the content and purpose of the consent form.

To minimize costs and risks, many providers have used electronic IC, but these systems are usually digital versions of standard paper-based systems⁶. Traditional consent forms, both paper and electronic, contain limited and/or inadequate descriptions – this inaccessibility is exacerbated with a high-level scientific vocabulary and a lack of a patient-centered approach. Therefore, the paper/electronic IC forms often require the medical provider or researcher to provide a simplified verbal laymen translation. This may also pose strain for patients, given the provider may not be the most adept at explaining the IC form contents. Many patient-centered outcomes research (PCOR) studies have advocated for providing more efficient, informative, and useful patient-centered informed consent processes²,⁷.
**Related works:** Research findings suggest Health Information Technology (HIT) interventions are successful tools for improving patient safety and quality of care. Virtual coaching, using mobile apps, shows great promise in encouraging general motivation to achieve healthy outcomes. Such interventions enabled through mobile health (mHealth) are completed by patients using apps on tablets (e.g., iPad), while maintaining the patient’s anonymity. mHealth architecture can be adopted as an IC medium in emphasizing patient-centered IC processes. mHealth achieves the planned aspect in that it allows patients to collect and report their own exposure history in real time on an app, rather than having to recall it after time has passed. Furthermore, in advocating for open and shared mHealth architectures, patients will not have to needlessly switch between apps to check exposure history or medical records. With an open architecture, the coherence and integration of mHealth benefits are emphasized. In addition to consolidating data interface in the mHealth tool, multimedia may also assist patients’ understanding and satisfaction with a research clinical trial or medical treatment process. Multimedia can also be used to tailor the mHealth tool to be more patient-centered, which is the primary goal of undertaken by this study.

**Study Development:** Given that the standard consent process does not always guarantee patient comprehension or improve the quality of care, many patients are consenting without the necessary knowledge of IC form contents. To improve the traditional IC process, we developed the Patient Centered Virtual Multimedia Interactive Informed Consent tool (VIC) (See Figure 1). Our approach uses virtual coaching to conduct a brief and virtual interview with patients using tablet computers (e.g., iPads) with a comprehensive multimedia library (e.g., video clips, animations, presentations, etc.) to explain the risks, benefits, and alternatives of the proposed treatment or clinical study, to enhance patient awareness. Our tool presents the IC materials to patients with the option of additional review to increase information grasp. The VIC tool has an option to assess patient comprehension with automated quizzes, which can help patients assess their own comprehension of the information presented. VIC provides many features and functions including Internet access to the consent, a retrievable electronic record of IC, electronic signature, and potential for seamless integration with the electronic health record (EHR). Moreover, VIC includes extensive security strategies that will guarantee the confidentiality and privacy of the patient and the clinical information. It also provides access to the IC content via the Internet before, during and after the study or procedure, allowing patients to benefit from supplemental resources as well.

In this paper, we present how we developed the VIC mHealth App based on prior work and evidence in IC research to support the patient-centered IC as well as enhance patient comprehension. We describe in detail the state-of-the-art software standards and the User-Centered Design approach (UCD) that we used to design, develop, refine, and test the VIC mHealth App. Moreover, we present the usability evaluation approach that we took to ensure that VIC satisfies the highest standard for usability and acceptability along while maintaining effectiveness, efficiency and patient satisfaction. The following sections list key factors that influenced design and development of VIC (Figure 2).

**Methods**

We used several methods to design, build, and evaluate VIC. The Our methods were based on the User-Centered Design (UCD) approach including requirement gathering and analysis, conceptual design, focus groups, design and development of mockups screen and prototype, usability evaluation of the prototype, development of the VIC system, and usability evaluation of VIC App. In this section, we describe these methods with an emphasis on the patient oriented design. The Institutional Review Board (IRB) of Yale University reviewed and approved the study protocol.
User-Centered Design Methodology (UCD).
We adapted UCD to build VIC, that is, an iterative multi-stage design approach that involves the user’s input throughout the development process\(^{18,19}\). UCD gives extensive attention to the user’s needs, wants and limitations at each stage of the design process and allows User Experience (UX) evaluations to be incorporated into the design. We used the data collected throughout various design phases to enhance the overall design iteratively. Specifically, we incorporated UCD techniques and UX evaluations to produce usable and acceptable software that delivers a more satisfying user experience. The result was a VIC system that is built and optimized around how patients can, want, or need to use it\(^{20}\). The following subsections list key factors that influenced design and development of the VIC system.

Conceptual Design.
The VIC system initial concept was developed based on prior work, and literature findings in IC research, patient input, and subject matter expert interviews. Our theoretical framework is based on Mayer’s Cognitive Theory of Multimedia Learning and the principle that the use of multimedia in the presentation of the IC process will improve patient comprehension\(^{15,16,21,22}\). In a recent study on patient-centered IC, Braithwaite et al. concluded that the type of IC that would be most suitable for patient-centered care is the type that is clinically oriented (e.g. decision aids) rather than legally oriented\(^{23}\).

Our goal was to develop a conceptual model with the patient in mind for a more efficient, informative, and effective patient-centered IC processes. We wanted to develop a model where a virtual coach could substitute for a human for most parts of the IC. The virtual coach would explain the IC in a way similar to how the study coordinator or a provider would explain the IC process to a study participant or patient. Therefore, the IC content could be presented in a style that is patient-friendly and easy to understand. The IC content should be brief and designed to quantify participant understanding, offering the participant a customized level of information. In addition, the model should include...
features to assess patient comprehension. Moreover, it should enhance the workflow and provide better access to the IC materials. Finally, it should also maintain the confidentiality and privacy of the patient and the clinical information.

**Front-End Focus Group.**

The goal of this focus group is to aid the initial design of VIC. We conducted a front-end focus group with patients and researchers from the Yale Center for Asthma and Airway Disease Mechanisms (YCAAD) clinic, and one IRB members from the Yale IRB office, as a part of the user requirements analysis. Patients were recruited by a research coordinator at the YCAAD clinic and were screened by the PI over the phone. Letters were then mailed containing directions and a reminder of the date and time of the focus group session. The focus group was held in a closed conference room at Yale University in July 2015 and lasted 80 minutes. All participants were told that their participation was voluntary and their information was confidential/anonymouse. An introductory script was read to explain the purpose and guidelines of the focus group and followed a dialogue of open-ended questions along with follow-up questions (Table 1). A member of the research team moderated the focus group and two observers took notes of the sessions, which was also audio-recorded. Patients in the focus groups were compensated with an incentive payment of $45 for their participation. The two research participants were recruited by several members of the research team and were not compensated. The audio recording was later transcribed for purposes of analyzing themes and comments.

The focus group contained 5 participants: 2 patients, 2 researchers, and 1 IRB member. Because patient input is vital to understanding the IC and how it is perceived from a patient prospective, we have identified two patients from the community to work directly on the project as co-investigators. We engaged the co-investigators in all aspects of the project including design, development, implementation and evaluation. The information gathered from the Front-End Focus Group was used to formulate the Initial User Requirements. Then, the research team then held several design sessions to carefully define and list all accepted technical and functional requirements of VIC.

**Table 1. Semi-structured Focus Group Questions.**

<table>
<thead>
<tr>
<th>Engagement Questions</th>
<th>Exploration Questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>When was the last time you had to go through the informed consent process?</td>
<td>Are you satisfied with how the informed consent process has functioned thus far? Why?</td>
</tr>
<tr>
<td>Did something happen in the procedure that you did not expect or was not clearly discussed in the informed consent?</td>
<td>Why do you think the informed consent is in paper form?</td>
</tr>
<tr>
<td></td>
<td>What do you think is missing from the current paper-based informed consent process?</td>
</tr>
<tr>
<td></td>
<td>Do you believe a better-informed consent solution exists for protecting patients?</td>
</tr>
<tr>
<td></td>
<td>In your opinion would video, audio, and graphical presentation help explain the informed consent process? Why?</td>
</tr>
<tr>
<td>In your opinion, would using an iPad to complete the informed consent process help? [Prompt for discussion for comfort level for using the iPad to consent]</td>
<td>In your opinion would using an iPad to complete the informed consent process help? [Prompt for discussion for comfort level for using the iPad to consent]</td>
</tr>
<tr>
<td>What do you believe are other potential effects of iPad-based informed consent? [Prompt for discussion of both positive and negative effects, including the impact on patient decision]</td>
<td>What do you believe are other potential effects of iPad-based informed consent? [Prompt for discussion of both positive and negative effects, including the impact on patient decision]</td>
</tr>
<tr>
<td>How do you see the role of the PI/Coordinator in conjunction with iPad-based informed consent?</td>
<td>How do you see the role of the PI/Coordinator in conjunction with iPad-based informed consent?</td>
</tr>
<tr>
<td>What are the issues/challenges you are interested in having the iPad-based informed consent address?</td>
<td>What are the issues/challenges you are interested in having the iPad-based informed consent address?</td>
</tr>
<tr>
<td></td>
<td>How do you think the iPad-based informed consent would influence the safety and security of patients?</td>
</tr>
<tr>
<td></td>
<td>How can we make iPad-based informed consent innovative but at the same time comply with IRB requirements?</td>
</tr>
<tr>
<td></td>
<td>How much time should be spent on the informed consent process?</td>
</tr>
<tr>
<td></td>
<td>Do you think we should keep both processes (paper and iPad-based informed consent)?</td>
</tr>
</tbody>
</table>

In general, all participants were open to the use of the iPad to present the informed consent. When asked about the level of comfort with the paper IC, participants expressed lack of comfort with the paper IC process. They felt that the IC was hard to read and understand, and full of many large words. They said that even though they read it, they did not really understand the content most of the time. Some said it is always the same thing and it is being repeated over
and over. Participants stressed the importance of having the IC read out aloud to them. Some even said, “If no one reads it for me, I will just sign it and roll the dice”. Most participants felt the process was too long and should be made simpler. When asked about what is missing from the paper IC process, they said: “it should include a summary, some videos to let me know what to expect.” There were also comments regarding the need for patients to be informed. In summary, although participants were concerned about the safety and security of their private health information, they were very excited about the HIT approach to deliver the IC process.

We conducted a summative analysis of participant phrasing and word use. We used information gathered from this focus group to create user cases and personas.

User Requirements Collection and Analysis.
Because patient input is vital to understanding the IC and how it is perceived from a patient prospective, we have identified two patients from the community to work directly on the project as co-investigators. We engaged the co-investigators in all aspects of the project including design, development, implementation and evaluation. The information gathered from the Front-End Focus Group was used to formulate the Initial User Requirements. Then, the research team then held several design sessions to carefully define and list all accepted technical and functional requirements of VIC.

Conceptual Model of the VIC mHealth App.
Based on the outcome of the Front-End Focus Group and the development of the conceptual model, our focus was mainly on developing VIC as a web-based App that would run on iPads to present IC content in a style that is patient-friendly and easy to understand. The initial model revolved around using a virtual coach or provider (see Figure 2), which guides the patient through the IC process. In VIC, the information and messages are displayed on the iPad screen and spoken through headphones for patient privacy. One of the innovative features of VIC is the ability to deliver health information through the use of multimedia and virtually guided interviews by a virtual provider (i.e. VIC), which explains the risks, benefits, and alternatives of the clinical procedure. Patients can view demos and presentations, and listen to comments and explanations. Animations and presentations will illustrate complicated content. The desired readability level for the text in VIC will be the 8th-grade level. This will enable us to safeguard patients with limited literacy. VIC is modeled around the idea that the content IC session should be brief and designed to quantify participant understanding, offering the participant a customized level of information. Additionally, the model included automated quizzes to assess patient comprehension. It also required providing Internet access to the consent, a retrievable electronic record of IC, electronic signature, and integration with the electronic health record (EHR). Moreover, it required adopting comprehensive security strategies that would guarantee the confidentiality and privacy of the patient and the clinical information.

App Wireframe and Screens Prototypes.
We used Balsamiq Mockups to design the wireframe and screen mockups. Balsamiq Mockups is a drag-and drop mockup and a rapid wire framing tool that helped us work more efficiently. It reproduces the experience of sketching on a whiteboard, using a computer for the prototyping purposes. The quick prototyping helped us simulate and test different versions of the tool’s User Interface. We continued to modify the mockups until we reached consensus on the final design. We used the content of the conceptual model and the system requirement to design the VIC mockups using the Balsamiq Mockups for the UI of the tablet (See Figure 4). Multiple sets and versions of VIC screens mockups were developed and tested and the modified and refined. During this process, we modified many features and properties of the end user interface. This included for example the process for testing user ability to use audio and the placement of navigation control and video control. The mockups provided us with means to test the flow of the screens and information and the amount of text that can be placed on single iPad screen. It also helped us decide on the most appropriate font size and color. We were able to evenly order and space the important sections and the main procedures. The Mockups expedited the development process and gave each research team member to visualize the software application and to be able to contribute to the design process in the joint design sessions. Using Balsamiq allowed us to produce a PDF prototype with full navigation capability comparable to the one of the final App.
System Architecture.
In this subsection, we include detailed descriptions of the VIC system architecture, methods, technologies, tools, and the support services that were used in the implementation of the mobile App. We designed VIC to be a Web-based mobile application that runs using traditional Internet browser interface, or on Apple iPads and other mobile devices. We developed it using Microsoft C# .NET and it runs on Microsoft Windows Server, Microsoft Internet Information Server (IIS), and Microsoft SQL Server 2012. Figure 3 demonstrates VIC’s three-tiered architecture. In addition to tier separation and firewall protection, the separate tiers provide a model for creating secure, flexible, and reusable mobile application components.

**Tier 1 – Presentation Logic Tier (i-DMZ1):** The Presentation Logic Tier is the IIS Web Application Server, and it consists of standard ASPX files and HTML pages. The function of this tier is to gather the requests from the presentation graphical user interface (GUI) and then forward the requests to the Business Tier once a user makes a request from a web browser. The Presentation Logic Tier will support different end-user devices (e.g. iPad, desktop, laptop, etc.). In return, the Presentation Logic Tier receives the results/output from the Business Tier, and transforms the results into readable formats and forwards data back to the Presentation GUI.

**Tier 2 – Business Tier (i-DMZ2):** The Business Tier is the brain of the VIC system. It contains the business logic, business rules, user security, validation rules and processing rules of the functions provided by the application. For example, once information is received from the Presentation Logic Tier, the information is validated against the predefined business and validation rules and business logic. If the information meets all the criteria, the information is forwarded to the Data Tier. The Business Tier also generates confirmation codes and forwards it back to the Presentation Logic Tier. The Business Tier will reside on the Business Server.

**Tier 3 – Data Tier (i-DMZ3 (SQL Server 2012):** The Data Tier handles the storage and retrieval of the information in the database. It can save and retrieve information from the database. *Microsoft SQL Server 2012 Enterprise Edition* will be the Database Management System (DBMS) for the proposed application.

Text-to-Speech.
We wanted to create a patient-centered easy-to-use IC tool that is acceptable by all users. A tablet with text-to-speech interfaces addresses literacy issues and makes the IC process an option for inexperienced computer users. Text-to-
speech translation is a key feature of VIC and is achieved by online and automated text-to-speech translation. One way to add the audio functionality to VIC is to record the audio associated with textual content of every screen in waveform and then play it back when the screen is displayed. However, this approach needed significant time and effort to record and edit the audio. In addition, we will require fresh recording every time the text on this screen is edited or changed. This would have slowed down the development and reduced the quality of the App. Our challenge was to find a low-cost, efficient solution to translate text in English into speech in high quality audio without compromising App performance. A more effective method was to use Text-To-Speech (TTS), a means of converting text to synthetic speech using a computer. In VIC, we used an online TTS solution (www.ispeech.org) to handle text to speech translation. For each text segment in the App that will be displayed on the screen, VIC requests (in real-time) an audio file translation associated with that text from ispeech.org through an API call. The API will then return the equivalent MP3 file VIC and VIC plays to the user.

**VIC App Usability Evaluation.**

After the VIC App was developed and tested, we wanted to ensure that the tool followed user interface standards-compliant coding practices for usability and acceptability. We performed a usability evaluation on VIC using a structured walk-through process with recognized usability principles. **Usability** refers to the ease in which the users can interact with tools. For the final design, we conducted the usability evaluation with nine representative asthma patients recruited from the YCAAD clinic. Patients were approached by the study coordinator and informed of this pilot study and offered the choice to participate voluntarily. If they chose to participate in the study, then they were scheduled for a study visit at a later convenient date and time in our Medical Simulation Center. Usability evaluation sessions were conducted in one-on-one settings to gather performance and satisfaction data. We evaluated the extent to which the application’s user interface is functioning according to the user requirements in the design specification to ensure that user needs are met. On the study visit, participants were reminded of the study protocol, study purpose, and patient’s role.

At the beginning of the session, a description of the study and the order of activities were read to participants. Each participant was asked to sign the consent form before participating in the study. In addition, they were asked to fill a demographic survey and a Computer Efficacy Scale (CES). Participants were told that the session will be audio and video recorded and the iPad must remain within the view of the above recording camera, for future analysis of potential on-screen inefficiencies. Participants were provided with a headphone in keeping with the intended future use of the VIC in the clinical setting. The participants were asked to review the VIC tool on the iPad in the presence of a researcher, while being videotaped, and share their perceptions of the application by speaking out loud. Participants were asked to complete a demographic questionnaire, CES, before they use VIC. For CES, participants were asked to rate their confidence about using the new technology on the scale of 1–10 (1= not at all confident and 10= completely confident) for each of the listed condition.

Additionally, they completed a system usability scale (SUS) and supplemental quantitative/qualitative data gathering, after using VIC. The SUS, created by John Brooke, is an industry standard, quick and reliable tool for measuring usability that can be used for small sample sizes. The validated SUS is a ten item Likert Scale that measured user attitude, effectiveness, efficiency, and satisfaction with the application. A short post-use interview was conducted to allow for additional qualitative feedback from the participants. In this phase, participants were asked open-ended questions and probed for recommendations to enhance VIC usability. At the end, participants were rewarded with a gift card for $15 as appreciation for participation. Our study was reviewed and approved by the Yale University Human Investigation Committee (IRB).

Quantitative data was analyzed as numerical indicators and summarized using common descriptive statistics appropriate for discrete and continuous data. Non-numerical indicators were analyzed using qualitative methods. Key results were used to 1. Modify the VIC system to make it more usable and acceptable in terms of system’s design; and 2. Normalize VIC by reducing process and structural problems. We conducted data analysis by using audio and video recording, user screen capture, note-taking, and participant survey. Key usability measures included both qualitative and quantitative outcome measures. These include effectiveness, efficiency, and satisfaction.

**Usability Evaluation Results.**

In this section, we will list our main results of the usability evaluation based on using UCD approach for the development and evaluation of the VIC mHealth App. We found that the UCD approach was appropriate for VIC development. It provided hybrid techniques for requirement gathering and analysis and to create the initial and conceptual design of the VIC App with the patient in mind. Moreover, UCD approach was instrumental in guiding us
through various stages of the development including conducting focus groups, design and development of mockup screens and prototypes, development of the VIC system, and usability evaluation of VIC App.

Table 2. Computer Efficacy Scale (CES): confidence about using the new technology on the scale of 1 – 10 (1= Not at all confident and 10= Completely confident)

<table>
<thead>
<tr>
<th>Question/Subject No.</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>If there was no one around to tell me what to do as I go</td>
<td>1</td>
<td>9</td>
<td>5</td>
<td>9</td>
<td>10</td>
<td>10</td>
<td>9</td>
<td>3</td>
<td>6</td>
<td>6.9</td>
</tr>
<tr>
<td>If I had never used a product like it before</td>
<td>1</td>
<td>9</td>
<td>6</td>
<td>8</td>
<td>10</td>
<td>10</td>
<td>7</td>
<td>3</td>
<td>4</td>
<td>6.4</td>
</tr>
<tr>
<td>If I had only the product manuals for reference</td>
<td>1</td>
<td>10</td>
<td>7</td>
<td>9</td>
<td>10</td>
<td>10</td>
<td>7</td>
<td>2</td>
<td>7</td>
<td>7.0</td>
</tr>
<tr>
<td>If I had seen someone else using it before trying it myself</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>8</td>
<td>6</td>
<td>9</td>
<td>7.6</td>
</tr>
<tr>
<td>If I could call someone for help if I got stuck</td>
<td>10</td>
<td>5</td>
<td>8</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>5</td>
<td>7</td>
<td>7</td>
<td>8.0</td>
</tr>
<tr>
<td>If someone else had helped me get started</td>
<td>7</td>
<td>5</td>
<td>7</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>8.1</td>
</tr>
<tr>
<td>If I had a lot of time to complete the job for which the product was provided</td>
<td>8</td>
<td>9</td>
<td>8</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>7</td>
<td>5</td>
<td>8</td>
<td>8.3</td>
</tr>
<tr>
<td>If I had just the built-in help facility for assistance</td>
<td>8</td>
<td>10</td>
<td>7</td>
<td>10</td>
<td>9</td>
<td>10</td>
<td>8</td>
<td>4</td>
<td>9</td>
<td>8.3</td>
</tr>
<tr>
<td>If someone showed me how to do it first</td>
<td>8</td>
<td>7</td>
<td>5</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>7</td>
<td>10</td>
<td>8</td>
<td>8.6</td>
</tr>
<tr>
<td>If I had used similar products before this one to do the same job</td>
<td>7</td>
<td>10</td>
<td>5</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>9</td>
<td>6</td>
<td>9</td>
<td>8.4</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>55</strong></td>
<td><strong>79</strong></td>
<td><strong>64</strong></td>
<td><strong>96</strong></td>
<td><strong>99</strong></td>
<td><strong>100</strong></td>
<td><strong>80</strong></td>
<td><strong>50</strong></td>
<td><strong>76</strong></td>
<td><strong>77.7</strong></td>
</tr>
</tbody>
</table>

Table 3. Sample Demographic

<table>
<thead>
<tr>
<th>Factor</th>
<th>% (n)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Race/ethnicity</td>
<td></td>
</tr>
<tr>
<td>Non-Hispanic white</td>
<td>66.7% (6)</td>
</tr>
<tr>
<td>Non-Hispanic black</td>
<td>11.1% (1)</td>
</tr>
<tr>
<td>Mexican American</td>
<td>11.1% (1)</td>
</tr>
<tr>
<td>Other</td>
<td>11.1% (1)</td>
</tr>
<tr>
<td>Age (years)</td>
<td></td>
</tr>
<tr>
<td>21-45</td>
<td>33.3% (3)</td>
</tr>
<tr>
<td>45-64</td>
<td>33.3% (3)</td>
</tr>
<tr>
<td>65-74</td>
<td>22.2% (2)</td>
</tr>
<tr>
<td>Sex</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>44.4% (4)</td>
</tr>
<tr>
<td>Female</td>
<td>55.6% (5)</td>
</tr>
<tr>
<td>Education</td>
<td></td>
</tr>
<tr>
<td>High school graduate</td>
<td>11.1% (1)</td>
</tr>
<tr>
<td>At least some college</td>
<td>11.1% (1)</td>
</tr>
<tr>
<td>At least Bachelor's degree</td>
<td>77.8% (7)</td>
</tr>
<tr>
<td>Marital status</td>
<td></td>
</tr>
<tr>
<td>Married</td>
<td>55.6% (5)</td>
</tr>
<tr>
<td>Separated or divorced</td>
<td>33.3% (3)</td>
</tr>
<tr>
<td>Never married</td>
<td>11.1% (1)</td>
</tr>
<tr>
<td>Household income</td>
<td></td>
</tr>
<tr>
<td>$30,000-$49,000/year</td>
<td>22.2% (2)</td>
</tr>
<tr>
<td>$50,000-$69,000/year</td>
<td>22.2% (2)</td>
</tr>
<tr>
<td>$70,000-$89,000/year</td>
<td>0.0% (0)</td>
</tr>
<tr>
<td>$90,000 or more/year</td>
<td>55.6% (5)</td>
</tr>
</tbody>
</table>

Table 4. SUS: System Usability Scale

<table>
<thead>
<tr>
<th>System Usability Scale (SUS)</th>
<th>Strongly Agree, % (n)</th>
<th>Strongly Disagree, % (n)</th>
<th>Agree, % (n)</th>
<th>Neutral, % (n)</th>
<th>Disagree, % (n)</th>
<th>Strongly Disagree, % (n)</th>
</tr>
</thead>
<tbody>
<tr>
<td>I think I would like to use this system frequently</td>
<td>0% (0)</td>
<td>0% (0)</td>
<td>22% (2)</td>
<td>0% (0)</td>
<td>77% (7)</td>
<td>0% (0)</td>
</tr>
<tr>
<td>I found the system unnecessarily complex</td>
<td>77% (7)</td>
<td>22% (2)</td>
<td>0% (0)</td>
<td>0% (0)</td>
<td>0% (0)</td>
<td>0% (0)</td>
</tr>
<tr>
<td>I thought the system was easy to use</td>
<td>0% (0)</td>
<td>0% (0)</td>
<td>55% (5)</td>
<td>0% (0)</td>
<td>0% (0)</td>
<td>0% (0)</td>
</tr>
<tr>
<td>I think I would need the support of a technical person to be able to use this system</td>
<td>66% (6)</td>
<td>33% (3)</td>
<td>0% (0)</td>
<td>0% (0)</td>
<td>0% (0)</td>
<td>0% (0)</td>
</tr>
<tr>
<td>I found the various functions in this system were well integrated</td>
<td>0% (0)</td>
<td>0% (0)</td>
<td>44% (4)</td>
<td>0% (0)</td>
<td>55% (5)</td>
<td>0% (0)</td>
</tr>
<tr>
<td>I thought there was too much inconsistency in this system</td>
<td>66% (6)</td>
<td>33% (3)</td>
<td>0% (0)</td>
<td>0% (0)</td>
<td>0% (0)</td>
<td>0% (0)</td>
</tr>
<tr>
<td>I would imagine that most people would learn to use this system very quickly</td>
<td>0% (0)</td>
<td>0% (0)</td>
<td>11% (1)</td>
<td>33% (3)</td>
<td>0% (0)</td>
<td>55% (5)</td>
</tr>
<tr>
<td>I found the system very awkward to use</td>
<td>66% (6)</td>
<td>11% (1)</td>
<td>0% (0)</td>
<td>0% (0)</td>
<td>0% (0)</td>
<td>0% (0)</td>
</tr>
<tr>
<td>I felt very confident using the system</td>
<td>0% (0)</td>
<td>0% (0)</td>
<td>22% (2)</td>
<td>0% (0)</td>
<td>77% (7)</td>
<td>0% (0)</td>
</tr>
<tr>
<td>I needed to learn a lot of things before I could get going with this system</td>
<td>55% (5)</td>
<td>22% (2)</td>
<td>0% (0)</td>
<td>11% (1)</td>
<td>11% (1)</td>
<td>0% (0)</td>
</tr>
</tbody>
</table>

All 9 participants (4 males and 5 females) completed the VIC usability study. All participants reported having at least a high school education (Table 3). All had access to both a desktop or laptop computer and reported having above 50% confidence in adapting to a new technology. Income brackets were largely similar with most sample participants identifying themselves as middle-class and most were Caucasian with only 1 African American and 1 Asian American.
The average CAS normalized score was nearly 78, indicating that the sample was relatively not anxious about using computers and adequately self-sufficient when it came to technology use (Table 2). The average SUS normalized score for all participants was 90th percentile, which is above the industry benchmark of 68th percentile (Table 4). All users reported that the user interface of the VIC mhealth tool was appropriate and easy to use, given that all subjects gave either a 4 or 5 when it came to ease of use and future implementation.

Overall, the users were successful in completing the tasks with a favorable impression of the VIC system. They had little difficulty understanding how to perform navigating the screens, playing and pausing videos and illustrations, showing and hiding the close caption text, playing or muting audio, answering the quiz questions, agreeing or disagreeing to participate, and signing the consent with their fingers on the iPad screen. However, they had some difficulty reading the text and suggested increasing the font size to make it more readable and to add audio to the quiz answer screen.

The most frequently suggested change to the mhealth tool was to change the Play button and make it indicate sound control. Also, to alter the interface such that hovering over the icon would show the icon’s meaning. All participants preferred the use of headphones with VIC, felt they had ample time to complete the program, and would strongly recommend its use in future clinical practice. Recommendations from the usability evaluation for improving the user experience included using larger font and control on-screen messaging to let users know how to proceed in the system; providing audio to the quiz screen and to read out load the quiz question and answer.

Conclusion

The use of mHealth can be successful in delivering health communications. More specifically, in developing a usable, acceptable, and efficient tool to deliver the IC process. Care must be taken to address the challenges that come with designing a new IC tool. One such challenge is ensuring that the new IC tool improves patient comprehension. This study focused on testing the feasibility of developing and evaluating such a tool.

Using the UCD approach we were able to design, develop, and evaluate a highly interactive mHealth App to deliver the IC process. The UCD approach facilitates working with multiple stakeholders and helped us plan and execute many tasks including requirement gathering and analysis, conceptual design, focus groups, design and development of mockups screen and prototype, usability evaluation of the prototype, development of the VIC system, rapid prototyping, and usability evaluation of VIC App.

Our Web-based mobile VIC App with virtual coaching and text-to-speech audio translation and iPad® interface offers some compelling ways to surmount existing implementation barriers. Our App architecture intentionally and readily addresses literacy issues and patients’ needs in the informed consent process. This has important implications for facilitating more meaningful and broader implementation of an IC tool, enhancing patient comprehension, and reducing provider burden engaging patients in an interactive IC process. Our user-focused development approach resulted in a product that is usable, acceptable, and efficient. Overall, the VIC mHealth tool had high usability as rated by the representative asthma patients. Successful research increasingly requires multi-disciplinary and inter-disciplinary teams. Our team has unique and complementary backgrounds encompassing HIT, human computer interaction, mHealth, UCD, UX evaluation, system design and architecture, human subjects protection, and clinical, and translational research. This study presents how our team is a success story of interdisciplinary collaboration between several entities.

Possible future directions include assessing the VIC tool capabilities for different chronic. Potential modifications include improving the reusable infrastructure of the VIC App by using a larger and more diverse sample size to create a more patient-centered mHealth tool. In the longer term, we are also interested in assessing whether the VIC predicts more positive health outcomes for patients.

Acknowledgements

Project described was made possible by grant from AHRQ R21 HS23987. We would like to acknowledge the two asthma patients who in the research team as co-researchers included Sheinea Doram (a finance associate with moderately severe asthma), and Judy Kaya (a former nurse with severe asthma). This publication was made possible by the Yale CTSA grant UL1TR000142 from the National Center for Advancing Translational Sci. (NCATS), NIH.
References

Big data in healthcare– the promises, challenges and opportunities from a research perspective: A case study with a model database

Mohammad Adibuzzaman, PhD¹, Poching DeLaurentis, PhD¹, Jennifer Hill, MSc¹, Brian D. Benneyworth, MD, MS²

¹Regenstrief Center for Healthcare Engineering, Purdue University, West Lafayette, Indiana, USA
²Children’s Health Services Research Group, Department of Pediatrics, Indiana University School of Medicine, Indianapolis, USA

Abstract

Recent advances in data collection during routine health care in the form of Electronic Health Records (EHR), medical device data (e.g., infusion pump informatics, physiological monitoring data, and insurance claims data, among others, as well as biological and experimental data, have created tremendous opportunities for biological discoveries for clinical application. However, even with all the advancement in technologies and their promises for discoveries, very few research findings have been translated to clinical knowledge, or more importantly, to clinical practice. In this paper, we identify and present the initial work addressing the relevant challenges in three broad categories: data, accessibility, and translation. These issues are discussed in the context of a widely used detailed database from an intensive care unit, Medical Information Mart for Intensive Care (MIMIC III) database.

1 Introduction

The promise of big data has brought great hope in health care research for drug discovery, treatment innovation, personalized medicine, and optimal patient care that can reduce cost and improve patient outcomes. Billions of dollars have been invested to capture large amounts of data outlined in big initiatives that are often isolated. The National Institutes of Health (NIH) recently announced the All of Us initiative, previously known as the Precision Medicine Cohort Program, which aims to collect one million or more patients’ data such as EHR, genomic, imaging, socio-behavioral, and environmental data over the next few years¹. The Continuously Learning Healthcare System is also being advocated by the Institute of Medicine to close the gap between scientific discovery, patient and clinician engagement, and clinical practice². However, the big data promise has not yet been realized to its potential as the mere availability of the data does not translate into knowledge or clinical practice. Moreover, due to the variation in data complexity and structures, unavailability of computational technologies, and concerns of sharing private patient data, few projects of large clinical data sets are made available to researchers in general. We have identified several key issues in facilitating and accelerating data driven translational clinical research and clinical practice. We will discuss in-depth in the domains of data quality, accessibility, and translation. Several use cases will be used to demonstrate the issues with the “Medical Information Mart for Intensive Care (MIMIC III)” database, one of the very few databases with granular and continuously monitored data of thousands of patients³.

2 Promises

In the era of genomics, the volume of data being captured from biological experiments and routine health care procedures is growing at an unprecedented pace⁴. This data trove has brought new promises for discovery in health care research and breakthrough treatments as well as new challenges in technology, management, and dissemination of knowledge. Multiple initiatives were taken to build specific systems in addressing the need for analysis of different types of data, e.g., integrated electronic health record (EHR)⁵, genomics-EHR⁶, genomics-connectomes⁷, insurance claims data, etc. These big data systems have shown potential for making fundamental changes in care delivery and discovery of treatments such as reducing health care costs, reducing number of hospital re-admissions, targeted interventions for reducing emergency department (ED) visits, triage of patients in ED, preventing adverse drug effects, and many more⁸. However, to realize these promises, the health care community must overcome some core technological and organizational challenges.
3 Challenges

3.1 Data

**Big data is not as big as it seems**

In the previous decade, federal funding agencies and private enterprises have taken initiatives for large scale data collection during routine health care and experimental research\(^5,9\). One prominent example of data collection during routine health care is the Medical Information Mart for Intensive Care (MIMIC III) which has collected data for more than fifty thousand patients from Beth Israel Deaconess Hospital dating back to 2001\(^3\). This is the largest publicly available patient care data set of an intensive care unit (ICU) and an important resource for clinical research. However, when it comes to identifying a cohort in the MIMIC data for answering a specific clinical question, it often results in a very small set of cases (small cohort) that makes it almost impossible to answer the question with a strong statistical confidence. For example, when studying the adverse effects of a drug-drug interaction, a researcher might be interested in looking at the vital signs and other patient characteristics during the time two different drugs were administered simultaneously, including a few days before the combination and a few days after the combination. Often this selection criteria results in a very small cohort of patients limiting the interpretation of the finding and with statistically inconclusive results. As an example, a researcher may want to investigate if any adverse effect exists when anti-depressants and anti-histamines are administered simultaneously. A query of simultaneous prescriptions of Amitriptyline HCl (anti-depressant) and Diphenhydramine HCl (anti-histamines) returned only 44 subjects in the MIMIC database (Figure 1). Furthermore, by filtering the data with another selection criterion (e.g., to identify the subjects for which at least one day’s worth of data exist during, before and after the overlap) the query returned a much smaller cohort with only four records.

**Data do not fully capture temporal and process information**

In most cases, clinical data are captured in various systems, even within an organization, each with a somewhat different intent and often not well integrated. For example, an EHR is primarily used for documenting patient care and was designed to facilitate insurance company billing\(^10\), and pharmacy records were designed for inventory management. These systems were not developed to capture the temporal and process information which is indispensable for understanding disease progression, therapeutic effectiveness and patient outcomes. In an attempt to study clinical process of vancomycin therapeutic drug monitoring based on ICU patient records in the MIMIC database, it was discovered that such process is not easy to reconstruct. Ideally, a complete therapeutic process with a particular drug contains the history of the drug’s prescription, each of its exact administration times, amount and rate, and the timing and measurements of the drug in the blood throughout the therapy. From the MIMIC III database we were able to find prescription

![Diagram](image-url)
information but it lacks the detailed dosing amount and prescription’s length of validity. The “inputevents” table contains drug administration information but does not include the exact time-stamp and drug amount which is critical for studying intravenous infused vancomycin in the ICU. It is also difficult to match drug prescription and administration records because their recording times in the clinical systems often are not the precise event times, and prescribed drugs are not always administered.

<table>
<thead>
<tr>
<th>Time</th>
<th>Prescriptions</th>
<th>Input_events</th>
<th>Lab_events</th>
</tr>
</thead>
<tbody>
<tr>
<td>10/13/2141 0:00</td>
<td>10/16/2141 0:00 Vancomycin 1g Frozen Bag</td>
<td>1000 mg</td>
<td>IV</td>
</tr>
<tr>
<td>10/13/2141 8:30</td>
<td>10/13/2141 8:31 Vancomycin 1</td>
<td>dose</td>
<td>Vancomycin 1 g/mL</td>
</tr>
<tr>
<td>10/14/2141 8:31</td>
<td>10/14/2141 9:50 Vancomycin 1</td>
<td>dose</td>
<td>Vancomycin 2.2 g/mL</td>
</tr>
<tr>
<td>10/14/2141 9:50</td>
<td>10/14/2141 22:00 Vancomycin 1</td>
<td>dose</td>
<td>Vancomycin 17.3 g/mL</td>
</tr>
<tr>
<td>10/15/2141 6:18</td>
<td>10/15/2141 19:50 Vancomycin 1</td>
<td>dose</td>
<td>Vancomycin 23.1 g/mL</td>
</tr>
<tr>
<td>10/15/2141 19:50</td>
<td>10/16/2141 21:10 Vancomycin 1</td>
<td>dose</td>
<td>Vancomycin 3.8 g/mL</td>
</tr>
</tbody>
</table>

Figure 2: An example of vancomycin therapeutic process reconstruction of one unique ICU stay using data from three different tables in the MIMIC III database.

Moreover, since the MIMIC III database does not contain detailed infusion event records which may be available from infusion pump software, one cannot know the precise drug infusion amount (and over what time) for any particular administration. The sparse and insufficient information on drug administration makes it almost impossible to associate available laboratory records and to reconstruct a therapeutic process for outcomes studies. Figure 2 is such an attempt of process reconstruction using data from the MIMIC III database including prescriptions, input events, and lab events for one patient during a unique ICU stay. The record only shows one valid prescription of vancomycin for this patient with start and end dates but does not indicate the administration frequency (e.g., every 12 hours) or method (e.g., continuous or bolus). The input events data (the second main column) came from the nursing records but it only shows one dose of vancomycin administration on each of the three-day ICU stay: one in the morning and two in the evening. Even though, as shown in the third main column, the “lab event” data contain the patient’s vancomycin concentration levels measured during this period, without the exact amount and duration of each vancomycin infusion, it is difficult to reconstruct this particular therapeutic process for the purposes of understanding its real effectiveness.

The problem of missing data remains relevant, even when the nursing workflow was designed to capture the data in the EHR. For example, as part of the nursing workflow, the information of drug administration should be documented in the medication administration records each time vancomycin was administered, and the MIMIC system was designed to capture all. But this was often not the case from our review of the database. Additionally, often times a patient’s diagnoses, co-morbidities, and complications are not fully captured nor available for reconstructing the complete clinical encounter. Those pieces of information are usually documented as free text not discrete data that can easily be extracted. Moreover, precise timings of the onset of an event and its resolution are rarely present. In the previous example of analyzing the effect of simultaneously administering Amitriptyline HCl and Diphenhydramine HCl, based on our selection criteria, we were able to find only one or two cases where such data were recorded (Figure 3). In the figure, each color represents one subject, and only one color (green, ID:13852) is consistently present in the time window for the selection criteria indicating missing systolic blood pressure measurements for the other subjects. This example is not an exception for cohort selection from data captured during care delivery, but a common occurrence due to the complex nature of the care delivery process and technological barriers in the various clinical systems developed in the past decade or so.

3.2 Access

Accessibility to patient data for scientific research and sharing of the scientific work as digital objects for validation and reproducibility is another challenging domain due to patient privacy concerns, technological issues such as interoperability, and data ownership confusion. This has been a widely discussed issue in recent years of the so-called patient or health data conundrum as individuals do not have easy access to their own data. We are discussing these challenges in the context of privacy, share-ability, and proprietary rights as follows.
Figure 3: Example of a cohort with missing systolic blood pressure data for three out of the four subjects meeting our clinical selection criteria. Day 0 (zero) is when drug overlap begins. This start of the overlap is aligned with multiple subjects and is denoted by the thick red line. Each data point represents one measurement from the “chartevents” table and each color indicates one subject and the black line indicates the average of the selected cohort.

Privacy
Access to health care data is plagued by vulnerability due to patient privacy considerations which are protected by federal and local laws of protected health information such as Health Insurance Portability and Accountability Act of 1996 (HIPAA)\textsuperscript{13}. The fear of litigation and breach of privacy discourages providers from sharing patient health data, even when they are de-identified. One reason is that current approaches to protect private information is limited to de-identification of an individual subject with an ID, which is vulnerable to twenty questions-like problems. For example, a query to find any patient who is of Indian origin and has some specific cancer diagnosis with a residential zip code 3-digit prefix ‘479’ may result in only one subject; thus exposing the identity of the individual.

Share-ability
Even after de-identification of patient data, the sharing of such data and research works based on the data is a complicated process. As an example, “Informatics for Integrating Biology and the Bedside (i2b2)\textsuperscript{5} is a system designed to capture data for scientific research during routine health care. i2b2 is a large initiative undertaken by Partners Healthcare System as an NIH-funded National Center for Biomedical Computing (NCBC). It contains a collection of data systems with over 100 hospitals that are using this software system on top of their clinical database. As a member of this project, each participating hospital system needed to transform their data into a SQL based star schema after de-identification. It required much effort for each institution to make the data available for scientific research as well as to develop the software in the first place. Although i2b2 was used exhaustively for research, sharing of data and research work as digital objects (i.e., the coding and the flow of the analysis) is not easily achieved. We argue that current EHR and other clinical systems do not empower the patients to take control of their data and engage in citizen science. The crowd sourcing approach might be one way to make a paradigm shift in this area which, unfortunately, is not yet possible with the current systems such as i2b2. A good example is the success in open source software technologies in other disciplines and applications (such as Linux, Git-hub, etc.) which rely on the engagement of many talented and passionate scientists and engineers all over the world to contribute their working products as digital objects\textsuperscript{14}.  

387
Figure 4: Sensitivity for the machine learning algorithms for different training sizes for prediction of Medical Emergency Team (MET) activation from the MIMIC database. The X-axis represents training size for different trials. For each training set, the results of a 10 fold cross validation are reported as box plots (the central red line is the median, the edges of the box are the 25\textsuperscript{th} and 75\textsuperscript{th} percentiles, the whiskers extend to the extreme data points the algorithm considers to be not outliers, and the red + sign denotes outliers). The blue asterisks represent the performance on the validation set of the algorithm that performs best on the test set. The blue dashed lines represent the performance of the National Early Warning Score (NEWS).

Proprietary rights
A relevant issue is the ongoing debate about the ownership of patient data among various stakeholders in the healthcare system including providers, patients, insurance companies and software vendors. In general, the current model is such that the patient owns his/her data, and the provider stores the data with proprietary software systems. The business models of most traditional EHR companies, such as Epic and Cerner, are based on building proprietary software systems to manage the data for insurance reimbursement and care delivery purposes. Such approach does not encourage or makes it difficult for individual patients to share data for scientific research, nor does it encourage patients to obtain their own health records that may help better manage their care and improve patient engagement.

3.3 Translation
Historically, a change in clinical practice is hard to achieve because of the sensitivity and risk aversion of care delivery. As an example, the use of beta blockers to prevent heart failure took 25 years to reach a widespread clinical adoption after the first research results were published. This problem is much bigger for big data driven research findings to be translated into clinical practice because of the poor understanding of the risks and benefits of data driven decision support systems. Many machine learning algorithms work as a "black box" with no provision of good interpretations and clinical context of the outcomes, even though they often perform with reasonable accuracy. Without proper understanding and translatable mechanisms, it is difficult to estimate the risk and benefit of such algorithms in the clinical setting and thus discourages the new methods and treatments from being adopted by clinicians or approved by the regulatory bodies such as the FDA.
For example, if a machine learning algorithm can predict circulatory shock from patient arterial blood pressure data, what would be the risk if the algorithm fails in a particular setting based on patient demographics or clinical history? What should be the sample size to achieve high confidence in the results generated by the algorithm? These are some critical questions that cannot be answered by those traditional “black box” algorithms, nor have they been well accepted by the medical community, which relies heavily upon rule based approaches.

As an example, a decision tree algorithm might perform very differently for prediction of Medical Emergency Team (MET) activation based on the training set or sample size from the MIMIC data. Furthermore, the prediction result can be very different when another machine learning algorithm, the support vector machine (SVM), was used (Figure 4).

3.4 Incentive

Yet another barrier in using big data for better health is the lack of incentive for organizations to take initiative to address the technological challenges. As mentioned earlier, EHRs are developed for purposes other than knowledge advancement or care quality improvement, and that has led to unorganized, missing, and inadequate data for clinical research. An individual health system does not usually have the incentive to make these data organized and available for research, unless they are big academic institutions. It would be easier for each individual health system to share data if they were organized and captured using standard nomenclature and with meaningful and useful detailed information with significant detail. A key question any health organization faces is: what is the return on investment for my hospital to organize all the clinical data it gathers? One model is the Health Information Technology for Economic and Clinical Health Act (HITECH) which promotes the adoption and meaningful use of health information technology. The act authorized incentive payments be made through Medicare and Medicaid to clinicians and hospitals that adopted and demonstrated meaningful use of EHRs, and the US government has committed payments up to $27 billion dollars over a ten year period. This incentive has paved the way for widespread adoption of EHRs since HITECH was enacted as part of the American Recovery and Reinvestment Act in 2009. However, for the purpose of using clinical data for scientific innovation and improving care delivery process, no apparent financial incentives currently exist for any organization to do so.

4 Opportunities

4.1 Data

For data driven research in health care, we propose to record the most granular data during any care delivery process so as to capture the temporal and process information for treatment and outcomes. For example, in an intensive care unit, the exact time of medication administrations need to be captured. This can be achieved in a number of ways. As a nurse bar code scans an oral medication into the electronic medication administration record (eMAR) the system also timestamps the action in the EHR. Detailed intravenous drug infusions can be linked to the patient clinical records by integrating the smart infusion pumps with the EHR systems. The Regenstrief National Center for Medical Device Informatics (REMEDI), formerly known as the Infusion Pump Informatics, has been capturing for capturing process and temporal infusion information. The planned expansion of such data set will allow linked patient outcomes and drug admin data forming a more complete treatment process for answering research and treatment effectiveness questions related to the administration of drugs such as drug-drug interaction, safe and effective dosage of drugs, etc., among others.

In order to achieve a statistically significant sample size after cohort selection, we promote breaking the silos of individual clinical data systems and making them interoperable across vendors, types and institutional boundaries with minimal effort. For the next generation of EHRs, these capabilities need to be considered.

4.2 Access

Patient/citizen powered research

To replicate the success in open source technologies in other disciplines by enabling citizen science, data and research analysis must be accessible to everyone. At the same time, patient privacy needs to be protected complying with the
privacy law and proprietary rights of the vendors, and researchers need to be protected. As an example, we have demonstrated such a system with the MIMIC database where interoperable and extensible database technologies have been used on de-identified patient data in a high performance computing environment19.

**Shareable digital objects**

For the next generation of EHRs and other big data systems such as REMEDi18 and i2b25, data must be findable, accessible, interoperable and reproducible (FAIR)20. For big data systems, a software-hardware ecosystem could work as a distribution platform with characteristics analogous to an Apple or Android “app store” where any qualified individual can access the de-identified data with proper authentication without the need for a high throughput infrastructure and the rigorous work, including pre-processing of the data needed to reproduce previous works. The proposed architecture is shown in Figure 519.

**4.3 Translation**

**Causal understanding**

Historically, clinical problems and treatment are studied and understood as “cause and effect”. For example, genetic disposition and lifestyle could lead to frequent urination, fatigue and hunger, and can be associated with diabetes. Based on this, the patient may be treated for this disease. However, most machine learning algorithms do not provide such a rule based approach; rather they predict the outcome of a given set of inputs, which may or may not be associated with known clinical understanding. Unlike other disciplines, clinical applications require a causal understanding of data driven research. Hence, most clinical studies start with some hypothesis, that ‘A’ causes ‘B’. The gold standard to identify this causation is randomized controlled trials (RCTs), which have also been the gold standard for regulatory approval of new drugs. Unfortunately, EHRs and the like data captured during routine healthcare has sampling selection bias and confounding variables and hence it is important to understand the limitation of such data sets. To answer the causal questions, a new generation of methods are necessary to understand the causal flow of treatment, outcome, and molecular properties of drugs by integrating big data systems for analysis and validation of hypothesis for transportability across studies with observational data21, 22. These methods would enable the regulators to understand the risk and benefit of data driven systems in clinical settings for new guidelines enabling the translation. Once those guidelines are established, technological solution must also be enabled at the point of care such that clinicians
can access for data driven queries as part of their clinical workflow.

5 Conclusion

“Big data” started with many believable promises in health care, but unfortunately, clinical science is different from other disciplines with additional constraints of data quality, privacy, and regulatory policies. We discussed these concepts in pursuit of a holistic solution that enables data driven findings to be translated in health care, from bench to bedside. We argue that the existing big data systems are still in their infancy, and without addressing these fundamental issues the health care big data may not achieve its full potential. We conclude that to make it to the next level, we need a larger cohort of institutions to share more complete, precise, and time stamped data as well as with greater willingness to invest in technologies for de-identifying private patient data for it to be shared broadly for scientific research. At the same time, as more and more “big data” systems are developed, the scientific and regulatory communities need to figure out new ways of understanding causal relationship from data captured during routine health care, that would complement current gold standard methods such as RCTs as well as identify the relationship between clinical practice and outcomes, as there is a wide disparity in the quality of care across the country².

References


Applying the Concept of Nutrient-Profiling to Promote Healthy Eating and Raise Individuals’ Awareness of the Nutritional Quality of their Food

Mayda A. Alrige, MSc¹, Samir Chatterjee, PhD¹, Ernie Medina, DrPH², Jeje Nuval, R.D.²
¹Claremont Graduate University, Claremont, California, USA; ²Loma Linda University, Loma Linda, California, USA

Abstract
Diet-related chronic diseases are on the rise. Current dietary management approaches are mostly calorie-counter tools that draw our attention away from the nutritional quality of our food choices. To improve consumers’ dietary behavior, we need a simple technique to educate them about nutrition and increase their understanding of the nutritional quality of their food. This study aims to design a dietary tool to promote a nutrient-dense diet. To this end, we applied the concept of Nutrient Profiling to classify food recipes based on their nutritional quality, by developing the Intelligent Nutrition Engine. This engine undergirds our mobile-based application, Easy Nutrition, which was designed to enable users to find food recipes and understand their nutritional quality. To evaluate the usability and understandability of our approach, we piloted the prototype of Easy Nutrition on 24 consumers. The results indicate that our approach provides a sustainable avenue to help consumers manage their diets.

Introduction
Diet-related chronic health conditions, such as diabetes, obesity, hypertension, and cardiovascular disease, present a major public health concern. As of 2014, more than one-third (36.5%) of U.S. adults are obese, according to the National Center for Health Statistics (2011-2014)¹. Two-thirds are overweight². These trends are spreading worldwide. According to a global WHO report from 2016, an estimated 422 million adults worldwide were living with diabetes, compared to 108 million in 1980. The global prevalence of diabetes has nearly doubled since 1980, as it had raised from 4.7% to 8.5%³. Furthermore, the American Diabetes Association reports that 29.1 million Americans (9.3% of the population) had been diagnosed with diabetes in 2012⁴. Another WHO report from 2004 on food and health in Europe states that diseases with major nutritional determinants account for 41% of disability-adjusted life years among all diagnosed diseases in Europe⁵. The root of the problem of all these conditions is a poor diet. Tackling this problem is by no mean easy and requires complex lifestyle changes. However, a healthy diet is a key component of a healthy lifestyle that can prevent the onset of these chronic diseases or mitigate their severity⁶.

Technology-tailored interventions to facilitate dietary management have been introduced in different formats, mainly through mobile-based applications (mHealth). Although evidence for the efficacy of mHealth is generally sparse⁷, research has shown that the use of the hand-held devices can improve the dietary intake of healthy food groups, such as whole grains and vegetables⁸. Also, the use of mHealth has the potential to reduce health costs, and to improve well-being in various ways, for instance, through the promotion of a healthy lifestyle by constant monitoring and self-management⁹,¹⁰. In their systematic review, Kroeze et al. conclude that there is sufficient evidence in favor of computer-tailored interventions for improving dietary behavior¹¹. These findings have also been supported by Long and colleagues in their review on the technology employed for dietary assessment¹². This suggests that mHealth may have a significant effect on individuals sustaining a healthy lifestyle.

Most of the current technology-tailored dietary tools deploy diet recalls and food records as the main dietary management approaches¹³,¹⁵. These approaches focus on the quantity of the food. That is, users are prompted to quantify the portion of the food that they consumed. One of the critical issues in this context is time. It is very tedious and time-consuming to track food intake. Another problem in addition to the time it takes to track food intake is the issue of recall⁶. Users have to sit down at least once a day, remember what they had eaten during the day in correct portions, and type in their food intake. Despite their effectiveness, these tools require performing some tasks that are unpractical to apply on a daily basis. Research suggests that too much detailed information on mobile-Phone may result in user being discouraged from using these tools¹⁶. The focus on food quantity with relation to the issues of time and recall has undermined the effectiveness of technology-tailored dietary tools.

The notion of nutrition profiling can alleviate these issues because it presents how healthy a food item is in a single holistic measure that is easy to follow and intuitive to understand. Nutrient profiling is defined as the science of
ranking foods according to their nutritional composition for reasons related to preventing disease and promoting health (WHO). By focusing on the food quality instead of the quantity, nutrient-profiling based systems aim to educate users about the overall nutritional quality that constitute a good or bad food choice. In light of this notion, systems such as Nuval (Nutritional Value) and ANDI (Aggregate Nutrition Density Index) have been developed to rank foods according to their nutritional quality. These systems have been tested, and widely accepted in the market landscape to rank food products based on their nutritional content.

This study aims to apply the concept of nutrient profiling in a mobile-based dietary application called Easy Nutrition. Easy Nutrition aims to increase consumers’ understanding about nutrition and raise their awareness about the nutritional value of the food recipes they choose. To rank different food recipes, we developed the Intelligent Nutrition Engine. This algorithm takes into account the three major macronutrients, two micronutrients and the number of calories the consumer needs daily. To present the nutritional quality of a particular food recipe in a simple, easy to understand manner, we adopted the notion of the traffic-light diet. The nutritional information is not given as a strict tri-color output. Rather, it is by analogy a color-coding food rating scale of eight values that rates the food recipe based on its nutrition from red (for less nutritious choices) to green (for nutritious choices).

While the goal of this study is to increase individuals’ understanding about nutrition and ultimately improve their dietary behavior by avoiding poor nutrition diets, the goal of this paper is to presents consumers’ feedback regarding our approach in presenting nutritional information. The rest of the paper is organized as follows. The literature will be reviewed from two perspectives. First we will investigate the state of dietary management and nutrition education to prevent the onset of chronic diseases. Second we will investigate the role of technology to facilitate this process. The method section will elaborate on the design science research, the approach we followed to develop the Intelligent Nutrition Engine and Easy Nutrition. This section will additionally outline the three-stage evaluation plan for my research, of which this paper represents the first stage. In the fourth section, the results of this stage (pilot study) will be illustrated and discussed in terms of the usability of Easy Nutrition and the understandability of its interfaces. Finally, we will conclude by arguing that our study contributes equally to both science and society. Our approach provides a new methodology (the Intelligent Nutrition Engine) to rank food and a dietary tool (Easy Nutrition) to help consumers better manage their diet by focusing the nutritional quality of their food choices.

Literature Review

Dietary Management

Managing diets is essential when it comes to treating diet-related chronic diseases. Furthermore, for healthy individuals, it is a preventative measure to maintain a healthy weight and overall wellbeing. Essentially, there are four different approaches to manage diets. These approaches encompass: dietary recall, food records, self-management and menu planning. In the first approach, the patient is asked over the phone about the amount of food, drinks consumed typically in a 24-hour period along with the method of preparation and the brand of the food items. Using food records, patients will have the chance do the same by writing this information down. In the self-management approach, personal digital assistant software is used. The patient will set some pre-defined goals on what quantities of food to consume. In this case, patients are in charge of how to keep their diets under control no matter what kind of food is consumed. The last approach is menu planning where patients get to have their meals planned based on previously identified food preferences. As the first three approaches are prone to the tediousness of counting and recording food intake multiple times a day, the focus in this study is to utilize the last approach, which is menu planning where users get to plan their meals after learning about their nutritional value.

Many studies have been conducted to develop and evaluate computerized dietary tools that are based on diet recall and food record. These two methods are designed as diet trackers and calories counter to assist individuals better manage their diet. However, the notion of calorie counting is very tedious and entails many issues from the patient’s perspective. First, the underlying method can suffer from recall issues, where users have to log their daily food intake. Examples of these food and calorie tracker apps include “MyFitnessPal”, “Lose It!”, or “Calorie Count” among many others. These apps allow users to log their food on a daily basis, define personal weight loss goals and review and analyze the gathered data against these goals. In this context, the second issue stems from a realistic limitation, which is the food database where users get to pick from. The quality of the food tracking/dairy app is tightly dependent on this underlying database. However, even the largest current food databases are still far from being complete and often contain only country-specific products. The notion of nutrition education or nutrition
profiling help mitigate these issues as it aims to educate patients about the underlying nutrients that constitute a good or a bad food choice.

**Nutrition Education**

To better manage diets and sustain a healthy lifestyle, one has to be aware and knowledgeable about the nutritional content in the food consumed. Being educated and aware of the macro nutrition and micro nutrition contributes to one’s overall health by minimizing poor nutrition diet. This leads to a healthy behavior that holds the promise of preventing the onset of chronic diseases and sustaining a healthy lifestyle. Diet education is effective to both understand diet requirements and control body weight and blood sugar levels. This is especially true when such educational material is delivered in an easy to follow, and convenient to understand manner. Indeed, knowledge-based nutrition education alone does not change dietary behavior. Behavioral nutrition education does a better job to change dietary behavior. The effectiveness of the behavioral nutrition approaches are investigated by Bader and her colleagues in. These approaches range from gaining familiarity with general nutrition principles, acquiring general planning frameworks, and finally planning tools to use meal replacements or prepared meals.

Bader and colleagues in their attempt to evaluate the effectiveness of nutritional education conducted a pilot study where they investigated one of the most preferred dietary management approaches: menu planning. The study targets type-2 diabetics. Authors conducted a single-arm clinical trial to evaluate one of the commercial internet-based menu planning tools. They examined pre- to post intervention changes in body weight, blood pressure, and glycaemia among overweight patients with T2DM (n=33). Nutritional recommendations were operationalized into weekly Internet-delivered menu plans. The findings indicate that there was 5% weight reeducation in 18% of the participants. The study highlights the effectiveness of the behavioral nutrition approaches that are less structured. These approaches range from gaining familiarity with general nutrition principles, acquiring general planning frameworks (eg, carbohydrate exchanges, “points”) and finally planning tools to use meal replacements or prepared meals.

**Nutrient Profiling**

Nutrient profiling is one of these behavioral nutrition approaches that do not dictate people what to eat or what not to eat. Rather, it aims at educating users about the overall nutritional quality of the food and leaving the choice of the meal to them. Nutritional profiling aims to rank food based on their nutritional content, as it is driven by the focus on food quality instead of food quantity. Individuals who follow high-scored food would most likely improve their dietary behavior. Simplicity is a key when it comes to presenting nutritional information. It has been suggested that nutritional information on mobile phones should be easy to read and understand. Approximate information is better than accurate facts that are harder to access but more precise. It was further suggested that the food information on phones should not be too fine-grained, as too much detailed information may result in user discouragement and little user friendliness. This stresses the idea that carb and calories intake counting used for dietary management has become less preferred in favor of a more generalized nutritional information about the quality of the food compositions. Nutrient profiling is the science of ranking foods according to their nutritional composition for reasons related to preventing disease and promoting health (WHO).

This idea has led to the creation of many nutritional rating systems. Examples of these systems include Nuval and ANDI. In addition, Leonard H. Epstein and his colleagues have developed the Traffic-Light Diet in the 1970’s. In this dietary approach, Epstein used a tri-color palette to create an easy-to-follow diet for overweight children. The notion of traffic-light diet had for two decades inspired new research due to its groundbreaking nature. The traffic-light diet is a structured eating plan that divides food by the color of the traffic signals. Green is for low-calorie food (go) that can be eaten at any time, orange (caution) is for moderate–calorie food that can be eaten occasionally, and red (stop) is for high-calorie food that should be eaten rarely. Since it was launched, the Traffic Light Diet has been used widely by pediatricians to encourage healthy eating habits among their patients. Many studies have been conducted utilizing the “Traffic-light” dietary approach and showed promising results. The Traffic Light Diet is used as a part of a comprehensive treatment, and the results show a significant decrease in obesity in preadolescents children. Significant changes in eating patterns have been reported when comprehensive obesity treatment has been combined with the Traffic-Light Diet. Reductions in “red foods” have been observed after treatment with significant associations between changes in intake of “red food” and weight loss or decrease in percent of
overweight. In our study, we adopted the dietary approach of Traffic Light Diet to present the nutritional information.

The Role of Technology

mHealth (short for mobile health) is a term for using mobile devices for health services and information. These devices can be mobile phones, patient monitoring devices, tablets, personal digital assistants, or other wireless devices. Although evidence for the efficacy of mHealth is generally sparse, research has shown that the use of the hand-held devices can improve the dietary intake of healthy food groups, such as whole grains and vegetables. Also, the use of mHealth has the potential to reduce health costs, and to improve well-being in various ways, for instance, through the promotion of a healthy lifestyle by constant monitoring and self-management. In their systematic review, Kroeze et al. conclude that there is sufficient evidence in favor of computer-tailored interventions for improving dietary behavior. These findings have also been supported by Long and colleagues in their review on technology employed for dietary assessment. This suggests that mHealth may have a significant effect on individuals sustaining a healthy lifestyle.

Method

Research Approach

This study follows the design science research approach, DSR, suggested by Hevner and Chatterjee. Alan and colleagues also present design science as a legitimate research paradigm to be employed in Information System research projects, where the goal is to solve practical problems. DSR aims to design IT-based artifacts to gain a better understanding of the problem in an iterative process. It involves the two main activities of building and evaluating. In this paper, we have developed two main artifacts. First is the Intelligent Nutrition Engine, the algorithm that is used to rank different food recipes based on their nutritional compositions. Second is Easy Nutrition, a mobile-based application that was designed to present how the algorithm performs to classify different food recipes. To ensure the utility and quality of these artifacts, we evaluated the prototype of Easy Nutrition for its usability and the understandability of its interfaces.

The Intelligent Nutrition Engine

We developed the Intelligent Nutrition Engine. This algorithm encompasses five different nutrients: the three major macronutrients that derive calories, which are fat, protein and carb and two micronutrients, which are dietary fibers and sodium. The algorithm checks the calories, the percentage of macronutrients, and also the amount of micronutrients in a selected recipe. These nutrients determine a healthful, or unhealthful choice. For people with diabetes, we consult the nutrition therapy recommendations by the American Diabetes Association to find the important nutrients that have to be considered, along with the appropriate limit for each nutrient. Some of these criteria have slight different specification for comorbid type-2 diabetes and hypertension. For example, the sodium recommendation for the general population is less than 2,300 mg/day. However for individuals with both diabetes and hypertension, further reduction in sodium intake should be individualized (no more than 1,500 mg/day). In addition, for comorbid T2D and hypertension, no more than 7% of fat should come from saturated fat. These criteria have been reviewed and validated by a registered dietician in Loma Linda university medical center. The list of these nutrients is outlined in Table 1.

<table>
<thead>
<tr>
<th>Macronutrients</th>
<th>Carbs</th>
<th>45 and 65 %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fats</td>
<td>25 and 35 %</td>
<td></td>
</tr>
<tr>
<td></td>
<td>For those with hypertension: No more 7% of this percentage should come from saturated fat</td>
<td></td>
</tr>
<tr>
<td>Protein</td>
<td>15 and 20 %</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Micronutrients</th>
<th>Dietary fibers</th>
<th>20-30 grams</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sodium</td>
<td>no more than 2300 mg</td>
<td></td>
</tr>
<tr>
<td></td>
<td>no more than 1500 mg daily for those with hypertension</td>
<td></td>
</tr>
</tbody>
</table>
Harris Benedict Equation

All these nutrition amounts/percentages are based off one’s daily calories intake. Because the recommended caloric intake differs based on age, gender, height and weight, we apply the Harris Benedict Equation to determine that basal metabolic rate (BMR) based on these factors (See Equation 1)

**Equation 1** Harris Benedict Equation

\[
\text{BMR for Men} = 66.47 + (13.75 \times \text{weight in kg}) + (5.003 \times \text{height in cm}) - (6.755 \times \text{age in years}) \\
\text{BMR for Women} = 655.1 + (9.563 \times \text{weight in kg}) + (1.850 \times \text{height in cm}) - (4.676 \times \text{age in years}).
\]

The final formula encompassing all these nutrients as well as the calories will produce one single holistic number that represent the overall nutritional quality of a particular recipe (Figure 1). For the sake of simplicity, we will present this number in a traffic-light scale that ranges from red to green through some intermediate colors. This is illustrated in Figure 1.

Figure 1. The nutritional score presented behind a traffic-light scale.

The cursor would start right in the middle of the traffic-light scale as an initial score for any given food recipe, as can be seen in Figure 1. The cursor would move to the right as a certain nutrient is within the recommended percentage/amount. On the other hand, the cursor would move to the left if a certain nutrient exceeds the maximum limit or fail to meet the minimum limit of the recommended range. This algorithm will be applied to Spoonacular, the largest online food API in order to rank different food recipes accordingly*

**Design of Easy Nutrition**

We developed an initial prototype of Easy Nutrition. Using Easy Nutrition, end-users will have the option to find online recipes that are tailored to their favorite cuisine and learn about their nutritional quality though our Intelligent Nutrition Engine. This algorithm will present the nutritional quality of the recipe in a traffic light scale as can be seen in Figure 2. This particular example shows that “Mahi Mahi Tacos” recipe is moderately unhealthy choice as the cursor is more toward the red indicator.

* Details of the algorithm and how it works can be requested from the first author (mayda.alrige@cgu.edu).
If the user is interested to know more about the nutrients that lower the overall nutritional quality, he/she can click on the “nutrition” tap to find out which nutrient is beyond the recommended range, as can be seen in Figure 3. Other pertinent information for each recipe is presented as well. These include the ingredients, instructions, and some healthy tips on how to maximize the nutritional value of the selected recipe.

Evaluation

Evaluation is a significant part of any design science research, to ensure the artifact(s) utility, efficacy and quality. To this end, we evaluated Easy Nutrition from a socio-technical perspective in a three-stage plan. The present study conducted the first stage of the evaluation plan, which is a pilot study to test the usability and understandability of our approach represented in Easy Nutrition. Put differently, in the first stage, the goal is to test whether or not presenting the nutritional quality in a traffic-light scale appeals to consumers and motivates them to select healthier, more nutrient-dense recipes. Since this stage is a pilot study, we intended to obtain participants’ feedback for improvement. Once we have established that the technology is sufficient to be employed, we will finalize the full version of Easy Nutrition and evaluate it in the second stage. The second stage aims to test Easy Nutrition as an enabling tool to help users increase their knowledge and understanding of the nutritional quality of their food and hence improve their dietary behavior. Based on participants’ answers to the questions related to their dietary behavior questions, we will generate a healthy behavior score. This score will be used as a baseline for post comparison. An increase in this score suggests improving in the dietary behavior in a pre-post intervention study. In the third stage, we will test Easy Nutrition effect in managing diabetes. In this paper, we will present the results of the first stage.

Participants

The first stage targets adults both with and without diet-related chronic conditions to evaluate the usability and understandability of our approach. The inclusion criteria are basic familiarity with smartphones so they can navigate through the app. This study is conducted in compliance with the ethical principles of Institutional Board or Review (IRB), at Claremont Graduate University.

Intervention: Procedures and Measures

Users were asked to perform basic tasks on the prototype of Easy Nutrition so they grasp an idea of what it is like to present the nutritional information in a traffic-light scale. After navigating through the app, participants were asked to answer a set of questions about Easy Nutrition usability. The System Usability Scale is utilized for this purpose.
SUS is a reliable and valid instrument used for usability assessment. It is a 10-item questionnaire that has five response options (strongly agree to strongly disagree). Prior research suggests that a SUS score above 68 is considered above average and that the website or mobile-app under investigation is usable. On the other hand, anything below 68 has to go under a lot of improvement to increase its usability. To obtain the SUS score, we followed the instructions of interpreting a participant’s score for each question that ranges from 1 (for strongly disagree) to 5 (for strongly agree). We added all the converted scores for the 10 questions for all the participants. Then, we divided this number by 24 (number of participants) and finally multiplied it by 2.5 to convert the original scores to 100-point scale.

In addition, in order to assess user satisfaction with the way we are presenting the nutritional quality of a recipe, we adopted the QUIS instruments. QUIS is the Questionnaire for User Interaction Satisfaction. QUIS aims to gauge users’ subjective satisfaction with specific aspects of the human-computer interface, such as screen visibility, terminology, system feedback and learning factors. Each area measures the users’ overall satisfaction with that facet of the interface on a 9-point scale. The questionnaire is designed to be configured according to the needs of each interface analysis by including only the sections that are of interest to the researcher.

Our goal is to capture users’ reaction mainly to the way the nutritional information is presented. We utilized the latest version of QUIS, in particular version 7.0 (the short form). Specifically, we used three measures out of nine: screen, terminology and users’ overall reaction to the application. As this is a pilot study, we concluded our survey with four open-ended questions to obtain users feedback regarding Easy Nutrition and the way we present nutritional information in a traffic-light scale.

**Results**

We utilized the Email channel to invite individuals to participate. Out of 50 recipients, about 6 participants have partially completed the survey, and 24 have fully completed the survey. 5 out of 24 participants have been diagnosed with diet-related condition, 2 are pre-diabetics, 2 are diabetics, and one had previously been diagnosed with Cancer (Figure 4).

![Figure 4](image)

**Figure 4** The number of participants with diet-related health condition

**Usability**

After navigating through the prototype of Easy Nutrition, participants were directed to an online survey, which asked them about the usability of Easy Nutrition. As mentioned above, we utilized the System Usability Scale (SUS). The usability score of Easy Nutrition came to 71.13. This is a good indicator that Easy Nutrition fits the context of dietary management and is appropriate to use in this context.

---

1 For a full list of the questions, please contact the first author at (mayda.alrige@cgu.edu).
Users’ Satisfaction

To capture participants’ subjective satisfaction, we asked them questions about their experience interacting with Easy Nutrition. For this purpose, we utilized the QUIS instrument, which asks questions about the app screens, terminologies and users’ overall reaction to the interfaces. As opposed to SUS, QUIS is a diagnostic tool, which means it provides a basis for improvement. This is because it asks very detailed questions about all the elements that make up the app interface, such as messages, characters, highlights, screen layouts, and sequence of screens. Participants in this study were totally satisfied with all interfaces relevant aspects in Easy Nutrition including the way we present the nutritional quality of the recipe in a traffic-light scale. Figure 5 and Figure 6 illustrate users’ satisfaction on 9-point scale.

One of the useful feedback comments we have obtained is to use some labels (e.g. “nutrient-dense” vs. “poor nutrition”) in the traffic-light scale that is used to represent the nutritional quality of the food recipe. The participant’s comment states that, “in order for Easy Nutrition to reach its full potential, I recommend adding some brief labels in conjunction with the scale”. We will consider such feedback when developing the full-version of Easy Nutrition for the second stage of this research. We will address this by labeling the traffic-light scale to range from poor nutrition (by the red side) to excellent nutrition (by the green side).

Discussion and Conclusion

The current study applies the concept of nutrient profiling on the domain of dietary management to help consumers gain a better understanding of the nutrition in their food and maximize the nutritional value of their choices. To this end, we have developed an algorithm, the Intelligent Nutrition Engine, to produce a single nutritional indicator of the food recipe. This algorithm takes into account the amount of calories, carbs, fat, protein, sodium and fibers. The Intelligent Nutrition Engine works by comparing the amount/percentages of these nutrients in a recipe against the recommended range. We applied this algorithm to rate some food recipes and present their nutritional quality in a traffic-light scale. To test the usability and understandability of this approach, we developed and piloted a mobile-based application called Easy Nutrition. We evaluated Easy Nutrition for its usability and understandability of its interfaces in a small sample size of 24 participants. The results of this pilot study show that Easy Nutrition is usable,
and that all of its interfaces are easy to understand. In addition, the feedback we obtained from these consumers gave us some basis for improvement.

This research contributes to the body of knowledge in two different levels: to the society and to the science. The first level of contribution is presented to the society as a nutrition-based dietary tool, called Easy Nutrition. Easy Nutrition presents the nutritional content of different food recipes in a traffic-light scale. The nutritional information is better absorbed if they are presented in a practical, easy to follow and easy to understand manner. The second and more important level of contribution is presented to the science as a new method for nutrition education. This is the Intelligent Nutrition Engine that is used to rank food recipes based on their nutritional quality. Nutrition education is a key component when it comes to dietary management. Thus, it is very important to find a simple, intuitive mechanism to present nutritional information. Our algorithm is based on the notion of behavioral nutrition approaches stressed by Bader and her colleagues23. The notion behind behavioral nutrition approaches is to increase the likelihood that individuals will implement the nutritional strategy learned. With Easy Nutrition, we don’t offer general nutritional information. Rather, the nutritional information is presented through individual recipes in a practical19 manner. This study can add to the evidence base that nutritional behavior strategies may be a modern adjunct to diabetes (or any other diet-related chronic condition) dietary management. Our study suggests that Easy Nutrition may have some beneficial effects to improve the dietary behavior of consumers.

In addition, the steps of this algorithm can be viewed as a set of design principles. The algorithm can be tailored to tackle different health conditions. Both the nutrients and the criteria for each nutrient can be tailored according to the health condition being treated. For example, cardiovascular diseases have certain nutrition therapy recommendations that are slightly different than those for diabetics. While diabetes management give priority to carbs consumption, CVD gives a special attention to fat consumption. The selection of the nutrients and the criteria for each nutrient can be adjusted to target different diet-related health conditions.

References
35. Aaron Bangor, PhD, CHFP, James Miller, PhD, Philip Kortum, PhD. Determining What Individual SUS Scores Mean: Adding an Adjective Rating ScaleJUS. JUS - J Usability Stud. :114–23.
Open Globe Injury Patient Identification in Warfare Clinical Notes

Emilia Apostolova, PhD1,3, Helen A. White, MBA2, Patty A. Morris2, David A. Eliason, MD2, Tom Velez, PhD3

1Language.ai, Chicago, IL; 2DoD/VA Vision Center of Excellence, Falls Church, VA; 3Computer Technology Associates, Cardiff, CA

Abstract

The aim of this study is to utilize the Defense and Veterans Eye Injury and Vision Registry clinical data derived from DoD and VA medical systems which include documentation of care while in combat, and develop methods for comprehensive and reliable Open Globe Injury (OGI) patient identification. In particular, we focus on the use of free-form clinical notes, since structured data, such as diagnoses or procedure codes, as found in early post-trauma clinical records, may not be a comprehensive and reliable indicator of OGIs. The challenges of the task include low incidence rate (few positive examples), idiosyncratic military ophthalmology vocabulary, extreme brevity of notes, specialized abbreviations, typos and misspellings. We modeled the problem as a text classification task and utilized a combination of supervised learning (SVMs) and word embeddings learnt in an unsupervised manner, achieving a precision of 92.50% and a recall of 89.83%. The described techniques are applicable to patient cohort identification with limited training data and low incidence rate.

Introduction

Open globe injury (OGI) refers to full thickness wound of the eyeball. OGIs range from small and self-sealing penetrations or lacerations to globe rupture with prolapse of intraocular contents or obliteration of the whole globe. The Birmingham Eye Trauma Terminology (BETT)1 defines and classifies OGIs as shown in Figure 1.

OGIs are considered a major cause of significant loss of vision, blindness and/or total loss of the eye. Given the possibility of such catastrophic outcomes, OGI treatment, outcomes, and risk factors are an active field of research, for example2–9. OGI clinical research poses challenges, as such injuries are typically rare. It has been reported that in the US, penetrating eye injury accounts for 3.81 injuries per 100 000 injuries annually10. In the context of military operations, however, the occurrence of OGI is more prevailing. Eyes injuries make up a significant portion of all casualties experienced in battle by our service members and veterans. OGIs are a significant number of these. Furthermore, unlike in peacetime where unilateral injuries are the rule, ocular war injuries are bilateral in 15 to 25% of cases11.

In this study, we utilize the Defense and Veterans Eye Injury and Vision Registry (DVEIVR), a web-based application, to develop methods for comprehensive and reliable OGI patient cohort identification. Congressionally mandated, DVEIVR was established to longitudinally track diagnosis, surgical intervention or other procedures, other treatments and follow up of active Service members and Veterans with significant eye and vision injuries from point of injury through rehabilitation starting 9/11/2001 and forward12. The DVEIVR registry includes abstracted data derived from initial battlefield acute care and subsequent definitive care encounters. The goal of DVEIVR is to aid researchers and providers in discovery of post-injury care pathways designed to prevent vision related injuries, protect and preserve the visual system, and restore vision. The DVEIVR data provided for the study is de-identified and complies with the DoD Privacy Office requirements, therefore exempted from IRB review.

In the context of DVEIVR, structured patient data can be used to somewhat reliably identify OGI patients. For example, ICD9 Diagnosis Codes 871.1 (Ocular laceration with prolapse or exposure of intraocular tissue), 360.61 (Foreign body in anterior chamber), etc. or Exam Types Globe - Anophthalmic, Globe - IOFB are reliable indicators of OGI. However, structured data is not always a comprehensive indicator of OGIs3. Diagnosis Codes can be ambiguous.
Figure 1: The BETT Classification of Open Globe Injuries.

and used to describe both OGI and closed globe injuries. Furthermore, codes and structured data in early post-trauma clinical records could be inaccurate for a number of practical reasons. We have observed a number of OGI incidents identified only in free-form text (clinical notes). The following brief notes are all clear indicators of the presence of an OGI:

**HISTORY OF PRESENT ILLNESS:** ruptured right globe in 2004 with a metal FB going through nasal right cornea through lens and into the vitreous. Pt underwent vitrectomy and phaco iol surgery right eye in 2006

**CHIEF COMPLAINT TEXT:** EYE TRAUMA INTRAOCULAR FOREIGN BODY

**ENCOUNTER COMMENT:** SPECIFY: Other (GLOBE EXPLORATION ); FINDINGS: EBL ¡0.1 CCNO GLOBE WALL PENETRATION AND DISRUPTION NOTED WITH THROUGH EXPOSURE AND EXPLORATION

In such cases, structured data could be missing (e.g. no ICD9 Diagnosis Code), the ICD9 Code used could include eye injuries referring to both OGI and non-OGI instances, e.g. 871.4 (Unspecified laceration of eye), 369.9 (Unspecified visual loss), or the code may have been incorrectly selected (e.g. a foreign body coded as *intraocular* with an associated free text treatment comment that the foreign body was removed using a *q-tip*, inferring the wound was actually superficial).

Given the importance of identifying all available OGI patients, we explored the use of Natural Language Processing and Machine Learning (ML) to identify OGI patients from clinical notes. The nature of the notes poses a number of challenges: the notes are typically very short, abundant in domain-specific concepts (ophthalmology in the context of warfare and veteran care), with a high rate of abbreviations and typos. The described techniques are applicable, in general, to patient cohort identification with challenging datasets, limited training data, and low incidence rates.

**Dataset**

The dataset consists of de-identified patient data from the DVEIVR database exported in May 2016. It comprises of both structured and unstructured fields describing diagnoses, treatments, procedures, visits, and exams. The data contains 26,131 unique patients and 76,809 unique encounters.

We utilized the available structured OGI information to semi-automatically generate a training data set. Tables 1 and 2 list structured information values found to be reliable indicators of OGI patients. All free-form text fields associated with structured OGI information were identified and labeled as positive examples. A randomized sample of the remaining records was used as negative examples. As descriptive free-form text associated with OGI structured data does not always indicate OGI by itself, the positive dataset was manually reviewed and false positive examples were discarded. This resulted in a training dataset consisting of 1,645 free-form text notes indicating OGI (positive examples) and 10,732 negative examples.

The characteristics of the training dataset poses a number of challenges. First, the number of positive examples is small for the purposes of text classifications. In addition, individual examples are typically very short. The median
**Table 1:** List of ICD9 Diagnosis Codes used as reliable indicators of Open Globe Injuries.

<table>
<thead>
<tr>
<th>ICD9 Diagnosis Code</th>
<th>ICD9 Diagnosis Code Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>360.5</td>
<td>Retained (old) intraocular foreign body magnetic</td>
</tr>
<tr>
<td>360.50</td>
<td>Foreign body, magnetic, intraocular, unspecified</td>
</tr>
<tr>
<td>360.51</td>
<td>Foreign body, magnetic, in anterior chamber of eye</td>
</tr>
<tr>
<td>360.52</td>
<td>Foreign body, magnetic, in iris or ciliary body</td>
</tr>
<tr>
<td>360.53</td>
<td>Foreign body, magnetic, in lens</td>
</tr>
<tr>
<td>360.54</td>
<td>Foreign body, magnetic, in vitreous</td>
</tr>
<tr>
<td>360.55</td>
<td>Foreign body, magnetic, in posterior wall</td>
</tr>
<tr>
<td>360.59</td>
<td>Intraocular foreign body, magnetic, in other or multiple sites</td>
</tr>
<tr>
<td>360.6</td>
<td>Retained (old) intraocular foreign body nonmagnetic</td>
</tr>
<tr>
<td>360.60</td>
<td>Foreign body, intraocular, unspecified</td>
</tr>
<tr>
<td>360.61</td>
<td>Foreign body in anterior chamber</td>
</tr>
<tr>
<td>360.62</td>
<td>Foreign body in iris or ciliary body</td>
</tr>
<tr>
<td>360.63</td>
<td>Foreign body in lens</td>
</tr>
<tr>
<td>360.64</td>
<td>Foreign body in vitreous</td>
</tr>
<tr>
<td>360.65</td>
<td>Foreign body in posterior wall of eye</td>
</tr>
<tr>
<td>360.69</td>
<td>Intraocular foreign body in other or multiple sites</td>
</tr>
<tr>
<td>871</td>
<td>Open wound of eyeball</td>
</tr>
<tr>
<td>871.0</td>
<td>Ocular laceration without prolapse of intraocular tissue</td>
</tr>
<tr>
<td>871.1</td>
<td>Ocular laceration with prolapse or exposure of intraocular tissue</td>
</tr>
<tr>
<td>871.2</td>
<td>Rupture of eye with partial loss of intraocular tissue</td>
</tr>
<tr>
<td>871.3</td>
<td>Avulsion of eye</td>
</tr>
<tr>
<td>871.5</td>
<td>Penetration of eyeball with magnetic foreign body</td>
</tr>
<tr>
<td>871.6</td>
<td>Penetration of eyeball with (nonmagnetic) foreign body</td>
</tr>
<tr>
<td>871.7</td>
<td>Unspecified ocular penetration</td>
</tr>
<tr>
<td>871.9</td>
<td>Unspecified open wound of eyeball</td>
</tr>
<tr>
<td>940.5</td>
<td>Burn with resulting rupture and destruction of eyeball</td>
</tr>
</tbody>
</table>

The number of words per example is only 7. Furthermore, as in most clinical texts, abbreviations, spelling variations, typos, and misspellings are quite common. All these factors contribute to a relatively large vocabulary size, while at the same time both the number of positive examples and the number of words per example are relatively small. Lastly, the vocabulary of the military ophthalmology domain is distinctive from general-domain clinical notes and doesn’t quite fit general purpose medical ontologies and terminologies.

**Method and Results**

We modeled the task as a binary classification of free-form clinical notes. It has been shown that Support Vector Machines\(^\text{14}\) achieve superior results in most text classifications tasks, including short texts (e.g.\(^\text{15,16}\) ). SVMs were selected as a sensible first choice. The individual clinical notes were represented as a bag of words (1-grams)\(^2\). The tokens were all converted to lower case and non-alphabetic characters were discarded. In addition, tokens that are present in more than 50% of all documents or less than 3 times across all documents were also discarded. This resulted in a dictionary of size 4,032 unique tokens. Tokens were weighted using the term-frequency / inverse-document-frequency scheme (tf-idf). We performed 10-fold cross-validation using linear kernel SVMs. We increased the positive class weight to address the unbalanced dataset (fewer positive examples). Best results achieved with positive class weight of 2. This resulted in a precision of 90.25 and a recall of 87.78 as shown in Table 3.

While the available positive training data is quite limited, it seemed sensible to attempt to utilize all available free-form text notes from all 76,809 patient encounters. Word embeddings\(^\text{18,19}\) have gained popularity as a way of capturing semantic knowledge of words without relying on available training data. Unsupervised algorithms have also been

\(^2\) Including bi-grams slightly deteriorated the model performance, a behaviour not atypical for text classification tasks\(^\text{17}\).
Table 2: List of Exam Detail Types used as reliable indicators of Open Globe Injuries.

| Exam Detail Types                                      | Used to represent variable pieces of texts such as sentences and paragraphs as fixed-length feature representations (Paragraph Vectors) \textsuperscript{20}. Studies have shown that Paragraph Vectors outperform bag-of-words models on a number of text classification tasks \textsuperscript{20, 21}. We used the free-text notes from all patient encounters to train Paragraph Vectors. The paragraph vectors the training notes were then fed to a logistic regression classifier and a neural network. In both cases results significantly under-performed the bag-of-words model. The underperformance of Paragraph Vectors is possibly due to the idiosyncrasies of the data: short text snippets, occasionally providing information on several, not necessarily related topics.

However, word embeddings revealed to be a powerful source of identifying semantically related abnormalities and procedures, variant spellings, abbreviations, and typos. Given the rare occurrence of OGIIs, the intuition was that identifying all possible tokens expressing key OGI concepts would improve our model performance. Using all available free-form text, we generated word vectors for all words in the vocabulary \textsuperscript{3}. The words in the training set were then clustered using Agglomerative Clustering. The average cosine distance between word vectors was used as a measure of similarity between clusters. The automatically derived clusters revealed to accurately capture related vocabulary words. Table 4 shows samples of word clusters derived with the described approach. The samples show that word embeddings accurately captured semantically related concepts (e.g. Cluster 7 lists various antibiotics); abbreviations

\textsuperscript{3}We used word skip-gram model, vectors of size 400, and window of size 4.

<table>
<thead>
<tr>
<th>Eye Laceration</th>
<th>Globe - Anophthalmic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Globe - IOFB</td>
<td>Globe - IOFB - Earth/Mud</td>
</tr>
<tr>
<td>Globe - IOFB - Glass</td>
<td>Globe - IOFB - Metallic</td>
</tr>
<tr>
<td>Globe - IOFB - NFS</td>
<td>Globe - IOFB - Other</td>
</tr>
<tr>
<td>Globe - IOFB - Stone/Sand</td>
<td>Globe - IOFB - Wood</td>
</tr>
<tr>
<td>Globe - Intraocular Foreign Body (IOFB)</td>
<td>Globe - Intraocular Foreign Body (IOFB) - Metallic</td>
</tr>
<tr>
<td>Globe - Open, Blunt Trauma, Ruptured</td>
<td>Globe - Open, Penetrating</td>
</tr>
<tr>
<td>Globe - Open, Penetrating - Cornea Only</td>
<td>Globe - Open, Penetrating - Corneascleral</td>
</tr>
<tr>
<td>Globe - Open, Penetrating - NFS</td>
<td>Globe - Open, Penetrating - Other</td>
</tr>
<tr>
<td>Globe - Open, Penetrating - Post Equatorial</td>
<td>Globe - Open, Penetrating - Sclera Only</td>
</tr>
<tr>
<td>Globe - Open, Perforating - With Uveal Prolapse</td>
<td>Globe - Open, Perforating</td>
</tr>
<tr>
<td>Globe - Open, Perforating - NFS</td>
<td>Globe - Open, Perforating - Other</td>
</tr>
<tr>
<td>Globe - Open, Perforating - With Uveal Prolapse</td>
<td>Globe - Ruptured</td>
</tr>
<tr>
<td>Retina - Macular Foreign Body - Glass</td>
<td>Retina - Macular Foreign Body - Other</td>
</tr>
<tr>
<td>Retina - Macular Foreign Body - Stone/Sand</td>
<td>Vitreous - Foreign Body</td>
</tr>
<tr>
<td>Vitreous - Foreign Body - Metallic</td>
<td>Vitreous - Foreign Body - Metallic-Magnetic</td>
</tr>
<tr>
<td>Vitreous - Foreign Body - Metallic-Non Magnetic</td>
<td>Vitreous - Foreign Body - NFS</td>
</tr>
<tr>
<td>Vitreous - Foreign Body - Other</td>
<td>Vitreous - Foreign Body - Stone/Sand</td>
</tr>
<tr>
<td>Socket - Status of Globe - Enucleation</td>
<td>Socket - Prosthetic Fit - NFS</td>
</tr>
<tr>
<td>Socket - Status of Globe - Exenteration</td>
<td>Anterior Chamber - Foreign Body</td>
</tr>
<tr>
<td>Anterior Chamber - Foreign Body - Metallic</td>
<td>Cornea - Laceration Full Thickness (Perforating)</td>
</tr>
<tr>
<td>Cornea - Laceration Full Thickness (Perforating) - Linear, Corneal Scleral</td>
<td>Cornea - Laceration Full Thickness (Perforating) - NFS</td>
</tr>
<tr>
<td>Cornea - Laceration Full Thickness (Perforating) - Stellate, Corneal Scleral - with Tissue Loss</td>
<td>Cornea - Laceration Partial Thickness (Penetrating)</td>
</tr>
<tr>
<td>Iris - Iridectomy</td>
<td>Iris - Iridotomy - Other</td>
</tr>
<tr>
<td>Macula - Foreign Body</td>
<td>Retina - Choroidal Rapture</td>
</tr>
<tr>
<td>Retina - Foreign Body</td>
<td>Sclera - Laceration</td>
</tr>
<tr>
<td>Sclera - Perforation Posterior - Superior Nasal</td>
<td>Socket - Prosthesis</td>
</tr>
<tr>
<td>Socket - Prosthesis - Conformer</td>
<td>Socket - Prosthesis - Prosthetic</td>
</tr>
<tr>
<td>Socket - Prosthesis - Scleral Shelf</td>
<td>Socket - Prosthetic Fit - Good</td>
</tr>
<tr>
<td>Socket - Prosthetic Motility - Fair</td>
<td>Socket - Prosthetic Motility - Good</td>
</tr>
<tr>
<td>Socket - Prosthetic Motility - Poor</td>
<td>Socket - Status of Globe - Enucleation</td>
</tr>
</tbody>
</table>
Table 3: Classification Results. SVM\textsuperscript{b}=baseline, SVM\textsuperscript{wc+ne}=word vector clusters + negation

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>F1-score</th>
<th>Vocabulary Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM\textsuperscript{b}</td>
<td>90.25</td>
<td>87.78</td>
<td>88.99</td>
<td>4,032</td>
</tr>
<tr>
<td>SVM\textsuperscript{wc+ne}</td>
<td>92.50</td>
<td>89.83</td>
<td>91.14</td>
<td>2,706</td>
</tr>
</tbody>
</table>

and typos (e.g. Cluster 6 lists 7 different spellings of \textit{fluorescein}); and many domain-specific semantic relations (e.g. Cluster 10 lists various military vehicles, while Cluster 3 lists procedures and conditions relevant to the concepts of \textit{enucleation}). We utilized the generated clusters to reduce the vocabulary size by treating all words in a cluster as equivalent.

Table 4: Sample word clusters generated using agglomerative clustering and the cosine distance of word embeddings.

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>detonated, exploded, attack</td>
</tr>
<tr>
<td>2</td>
<td>eccymosis, bruising, ecchymoses, echymosis, bruising, ecchymosis</td>
</tr>
<tr>
<td>3</td>
<td>enucleated, exenteration, anophthalmic, enucleation, orblototomy, enculeation, evisceration, nlp, enuc, phthisis</td>
</tr>
<tr>
<td>4</td>
<td>eyelids, lids, eyelid, palpebral, lid</td>
</tr>
<tr>
<td>5</td>
<td>flipped, everted, inversion, flipping, inverted, eversion</td>
</tr>
<tr>
<td>6</td>
<td>fluorescein, fluoro, fluorescein, florescein, flourescein, fluoro</td>
</tr>
<tr>
<td>7</td>
<td>gentamicin, mycin, emycin, erythromycin, antibiotic, abx, bacitracin, antibiotics, polytrim, tobradex, vigamoc, ilotycin</td>
</tr>
<tr>
<td>8</td>
<td>glasses, spec, specs, gls, srx, eyeglasses, eyeglass</td>
</tr>
<tr>
<td>9</td>
<td>ha, confusion, lightheadedness, vomiting, headache, dizziness, headaches, nausea, phonophobia</td>
</tr>
<tr>
<td>10</td>
<td>humvee, vehicle, truck, mrap</td>
</tr>
<tr>
<td>11</td>
<td>lasix, lasic, lasik, prk</td>
</tr>
<tr>
<td>12</td>
<td>schrapnel, gunshot, scrapnel, gsw, shrapnel</td>
</tr>
<tr>
<td>13</td>
<td>swab, tip, cotton, cta</td>
</tr>
<tr>
<td>14</td>
<td>washed, wash, irrigated, flushed, irrigation, lavage, rinse, flushing, flush</td>
</tr>
</tbody>
</table>

Lastly, we addressed the issue of negation (e.g. \textit{no globe rupture}) as negated conditions and procedures are quite common in the dataset. We ran a customized version of the NegEx\textsuperscript{22} algorithm to identify negated words and phrases. Encoding negated tokens in a bag-of-words model as separate token (e.g. \textit{rupture} vs \textit{no rupture}) is a common approach to treating negation. However, given the limited dataset and the large vocabulary size, we instead simply discarded the negated tokens.

We again performed 10-fold cross-validation using linear kernel SVM and positive class weight set to 2. Both precision and recall increased from 90.25 to 92.50 and from 87.78 to 89.83, respectively, while the vocabulary size was reduced significantly.

Related Work

There is a vast amount of literature on the use of free-form text and clinical notes for the purposes of patient cohort identification. A review of approaches to identifying patient phenotype cohorts using electronic health records\textsuperscript{23} describes a number of studies using clinical notes, most often in combination with additional structured information, such as diagnosis codes. The study asserts that clinical notes are often the only source of information from which to infer important phenotypic characteristics.

Most studies map textual elements to standard vocabularies, such as the Unified Medical Language System (UMLS). For example, Bejan et al.\textsuperscript{24} describe a system for pneumonia identification from narrative reports using n-grams, UMLS concepts, and assertion status. Similarly, Elkin et al.\textsuperscript{25} encoded radiology reports using the SNOMED CT Ontology and developed a set of rules to identify pneumonia cases. Carroll et al.\textsuperscript{26} used as features a combination of ICD-9 codes, UMLS concepts, and medication names and build an SVM classifier for the identification of rheumatoid arthritis patients.
A number of studies use rules, such as keywords and regular expressions, to identify the phenotype of interest. Friedlin et al. compare methods for identifying pancreatic cancer patients and utilize a tool that uses a combination of regular expressions and algorithms to detect keywords and their context. Xu et al. develop a system to detect patients with colorectal cancer that uses a combination of heuristic rule-based approach and the MedLEE system to extract relevant concepts. Sohn and Savova improve the Mayo Clinic Smoking Status Classification System by introducing a rule-based component for patient-level smoking status assignments.

A growing number of studies utilize the use of Machine Learning (ML) on the task of patient cohort identification from clinical narratives. The i2b2 (Informatics for Integrating Biology to the Bedside) project organized a challenge with the task of determining the patient smoking status from free-form discharge notes. The 11 participating teams utilized a variety of rule-based and ML approaches, including Naive Bayes, Decision Trees, AdaBoost classifiers, logistic regression, and neural networks. The most common ML algorithm was SVMs, as reported by 5 participating teams. The ShARe/CLEF eHealth 2013 task was organized to evaluate the state of the art in disorder recognition and normalization of the clinical narrative. Most of the participating systems employed hybrid approaches by supplementing features to a machine-learning algorithm and the best performing system utilized Conditional Random Fields and Structured SVMs. Lin et al. develop a system for automatic prediction of rheumatoid arthritis disease activity from the Electronic Medical Records (clinical narratives and lab values). They report that the best performing combination was linear kernel SVMs with UMLS concepts, feature selection, and lab values. Rumshisky et al. utilize Latent Dirichlet Allocation and SVMs to predict early psychiatric readmission from narrative discharge summaries.

Conclusion

We demonstrated an approach to identifying rare patient phenotypes (OGIs) utilizing free-form text in a challenging domain: military ophthalmology. The characteristics of the clinical narratives include extreme brevity, idiosyncratic terms and abbreviations, typos and misspellings. Using available structured data, we were able to construct a free-text training dataset with minimal manual review. We then employed supervised Machine Learning methods for classification of clinical notes. Best results were achieved utilizing clustering of the vocabulary (unigrams) using word embeddings and performing SVM classification by increasing the weight of the positive class to account for the imbalanced dataset. We were able to achieve an overall F1-score of 91.14. The described methods are applicable to short clinical text classification in narrowly specialized domains.

Acknowledgements

Research reported in this Open Globe Injury Patient Identification in Warfare Clinical Notes was supported by the DoD/VA Vision Center of Excellence under award number GS00Q09BGD0031. The views expressed in this article are those of the authors and do not necessarily reflect the official policy or position of the Department of the Navy, Department of Defense, or the U.S. Government.

References


Intelligent Word Embeddings of Free-Text Radiology Reports

Imon Banerjee, Ph.D\textsuperscript{1}, Sriraman Madhavan, B.E.\textsuperscript{1}, Roger Eric Goldman, M.D., Ph.D.\textsuperscript{1}, Daniel L. Rubin, M.D.\textsuperscript{1}
\textsuperscript{1}Department of Radiology, Stanford University School of Medicine, Stanford, USA

Abstract

Radiology reports are a rich resource for advancing deep learning applications in medicine by leveraging the large volume of data continuously being updated, integrated, and shared. However, there are significant challenges as well, largely due to the ambiguity and subtlety of natural language. We propose a hybrid strategy that combines semantic-dictionary mapping and word2vec modeling for creating dense vector embeddings of free-text radiology reports. Our method leverages the benefits of both semantic-dictionary mapping as well as unsupervised learning. Using the vector representation, we automatically classify the radiology reports into three classes denoting confidence in the diagnosis of intracranial hemorrhage by the interpreting radiologist. We performed experiments with varying hyperparameter settings of the word embeddings and a range of different classifiers. Best performance achieved was a weighted precision of 88% and weighted recall of 90%. Our work offers the potential to leverage unstructured electronic health record data by allowing direct analysis of narrative clinical notes.

1 Introduction

The Picture Archiving and Communication Systems (PACS) stores a wealth of unrealized potential data for the application of deep learning algorithms that require a substantial amount of data to reduce the risk of overfitting. Semantic labeling of data becomes a prerequisite to such applications. Each PACS database serving a major medical center contains millions of imaging studies “labeled” in the form of unstructured free text of the radiology report by the radiologists, physicians trained in medical image interpretation. However, the unstructured free text cannot be directly interpreted by a machine due to the ambiguity and subtlety of natural language and variations among different radiologists and healthcare organizations. Lack of labeled data creates data bottleneck for the application of deep learning methods to medical imaging\textsuperscript{1}.

In recent years, there is movement towards structured reporting in radiology with the use of standardized terminology\textsuperscript{2}. Yet, the majority of radiology reports remain unstructured and use free-form language. To effectively “mine” these large free-text data sets for hypotheses testing, a robust strategy for extracting the necessary information is needed. Methods for structuring and labeling the radiology reports in the PACS may serve to unlock this rich source of medical data.

Extracting insights from free-text radiology reports has been explored in numerous ways. Nguyen et al.\textsuperscript{3} combined traditional supervised learning methods with Active Learning for classification of imaging examinations into reportable and non-reportable cancer cases. Dublin et al.\textsuperscript{4} and Elkin et al.\textsuperscript{5} explored sentence-level medical language analyzers and SNOMED CT-based semantic rules respectively, to identify pneumonia cases from free-text radiological reports. Huang et al.\textsuperscript{6} introduced a hybrid approach that combines semantic parsing and regular expression matching for automated negation detection in clinical radiology reports.

In recent years, the word2vec model introduced by Mikolov et al.\textsuperscript{7,8} has gained interest in providing semantic word embeddings. One of the biggest problems with word2vec is the inability to handle unknown or out-of-vocabulary (OOV) words and morphologically similar words. The challenge is exacerbated in domains, such as radiology, where synonyms and related words can be used depending on the preferred style of radiologist, and a word may only have been used infrequently in a large corpus. If the word2vec model has not encountered a particular word before, it will be forced to use a random vector, which is generally far from its ideal representation. Thus, we explore how the word2vec model can be combined with the radiology domain-specific semantic mappings in order to create a legitimate vector representation of free-text radiology reports. The application we have explored is the classification of reports by confidence in the diagnosis of intracranial hemorrhage by the interpreting radiologist.
Our two core contributions are:

1. We proposed a hybrid technique for a dense vector representation of individual words of the radiology reports by analyzing 10,000 radiology reports associated with computed tomography (CT) Head imaging studies. (Word embeddings are publicly released in: https://github.com/imonban/RadiologyReportEmbedding)

2. Using our methods, we automatically categorized radiology reports according to the likelihood of intracranial hemorrhage.

We derived the word embeddings from large unannotated corpora that were retrieved from PACS (10,000 reports), and the classifiers were trained on a small subset of annotated reports (1,188). The proposed embedding produced high accuracy (88% weighted precision and 90% recall) for automatic multi-class (low, intermediate, high) categorization of free-text radiology reports despite the fact that the reports were generated by numerous radiologists of differing clinical training and experience. We also explored the visualization of vectors in low dimensional space while retaining the local structure of the high-dimensional vectors, to investigate the legitimacy of the semantic and syntactic information of words and documents. In the following sections, we detail the methodology (Sec. 2), present the results (Sec. 3) and finally conclude by mentioning future directions (Sec. 4).

2 Methodology

Figure 1 shows the proposed research framework that comprises five components: Dataset retrieval from PACS, Data Cleaning & Preprocessing, Semantic-dictionary mapping, Word and Report Embedding, and Classification. In the following subsections, we describe each component.

![Figure 1: Components of the proposed framework](image)

### 2.1 Dataset

The dataset consists of the radiology reports associated with all computed tomography (CT) studies of the head located in the PACS database serving of our adult and pediatric hospitals and all affiliated outpatient centers for the year of 2015. Through an internal custom search engine, candidate studies were identified on the PACS server based on imaging exam code. The included study codes captured all CT Head, CT Angiogram Head, and CT Head Perfusion studies. A total of 10,000 radiology reports were identified for this study. In order to provide a gold standard reference for the vector-space embedding algorithm, a subset of 1,188 of the radiologic reports were labeled independently by two radiologists. For each report, the radiologists read the previous interpretation and then graded the confidence of the interpreting physician with respect to the diagnosis of intracranial hemorrhage. For each study, a numeric label was provided on a scale ranging from 1 to 5 with labels as follows: 1) No intracranial hemorrhage; 2) Diagnosis of intracranial hemorrhage unlikely, though cannot be completely excluded; 3) Diagnosis of intracranial hemorrhage possible; 4) Diagnosis of intracranial hemorrhage probable, but not definitive; 5) Definite intracranial hemorrhage. These labels were chosen to reflect heuristics employed by radiologists and treating physicians to interpret the spectrum of information produced by the imaging study.
2.2 Data Cleaning & Preprocessing

All 10,000 radiology reports were transformed through a series of pre-processing steps to truncate the free-text radiology reports and to focus only on the significant concepts, which would enhance the semantic quality of the resulting word embeddings. We developed a python-based text processor - Report Condenser, that executes the pre-processing steps sequentially. First, it extracted the ‘Findings’ and ‘Impressions’ sections from each report that summarizes the CT image interpretation outcome, since our final objective was to classify the reports based on radiological findings.

In the next pre-processing stage, the Report Condenser cleansed the texts by normalizing the texts to lowercase letters and removing words of following types: general stop words, words with very low frequency (<50), unwanted terms and phrases (e.g. medicolegal phrases - “I have personally reviewed the images for this examination and agreed with the report transcribed above.”, headers - ‘FINDINGS’, ‘IMPRESSION’, ‘Additional comment’). These words usually appear either in all the reports or in a very few reports, thus of little or no value in document classification. We used the NLTK library for determining a stop-word list and discarded them during indexing. Examples of the stop-words are: a, an, are,...,be, by,...,has, he,...,etc. The Report Condenser also discarded datestamps, timestamps, the radiologist details (e.g. names, contacts) and other recurring phrases in reports. Removal of these terms significantly reduced the number of words that the system had to handle.

Following the removal steps, Report Condenser searched the updated corpus to identify frequently appearing pairs of words based on pre-defined threshold value of occurrence (> 500) and concatenated them into a single word to preserve useful semantic units for further processing. Some examples of the concatenated words are: ‘midline shift’ → ‘midline shift’, ‘mass effect’ → ‘mass effect’, ‘focal abnormality’ → ‘focal abnormality’.

In the next step, Report Condenser identified and encoded negation dependencies that appear in the radiology reports via simple string pattern matching. For example, in the phrase ‘No acute hemorrhage, infarction, or mass’, negation is applied to ‘acute hemorrhage’, ‘infarction’ as well as ‘mass’. Therefore, the Report Condenser encodes the negation dependency as: ‘No acute hemorrhage’, ‘No infarction’, ‘No mass’. Such phrases were identified automatically by analyzing the whole corpus and transformed accordingly.

2.3 Semantic-dictionary mapping

The main idea of the Semantic-dictionary mapping is to use a lexical scanner that recognizes corpus terms which share a common root or stem with pre-defined terminology, and map them to controlled terms. In contrast with traditional NLP approaches, this step does not need any sentence parsing, noun-phrase identification, or co-reference resolution. We used dictionary style string matching where we directly search and replace terms, by referring to the dictionary. We implemented a lexical scanner in python which can handle 1 kilobyte of text per millisecond. On average, the size of each radiology report after cleaning was 1 kilobyte and our scanner took less than 10 seconds to complete the whole mapping process for 10,000 radiology reports. We applied the following two-stage process.

1. Common terms mapping: First, we used the more general publicly available CLEVER terminology to replace common analogies/synonyms for creating more semantically structured texts. We focused on the terms that describe family, progress, risk, negation, and punctuations, and normalized them using the formal terms derived from the terminology.


2. Domain-specific dictionary mapping: For this case-study, we used the domain-specific RadLex ontology for mapping the variations of radiological terms that are related to hemorrhage, to a controlled terminology. We created an ontology crawler using SPARQL that grabs the sub-classes and synonyms of the domain-specific terms from Radlex, and creates a focused dictionary for “Intracranial hemorrhage” radiology reports. Using the dictionary all the equivalent terms of hemorrhage are formalized in the corpus as: {‘apoplexy’, ‘contusion’, ‘hematoma’, .. } → ‘hemorrhage’.

413
In Figure 2, we present the outcome of preprocessing and dictionary mapping by showing free-text reports and the corresponding processed texts side-by-side. In our corpus, average word count of original free-text reports is 285 and the average word count of processed reports is 98, which is approximately 3x reduction in size.

### 2.4 Word and Report Embedding

After pre-processing and dictionary mapping, the corpus of 10,000 processed reports (see examples in Figure 2) was used to create vector embeddings for words in a completely unsupervised manner using the word2vec model that can be trained on a large text corpus to produce dense word vectors. Two unsupervised algorithms were introduced to obtain word to vector representation: Continuous Bag of Words (CBOW) and Skip-gram. Those algorithms learn word representations that maximize the probabilities of a word given other contextual words (CBOW) and of a word occurring in the context of a target word (Skip-gram).

Our semantic dictionary mapping step considerably reduced the size of our vocabulary by mapping the words in corpus to their root terms, thereby making the words in the vocabulary more frequent. CBOW is several times faster to train than the Skip-gram, with slightly better accuracy for frequent words. The CBOW architecture also captures the semantic regularities of words. Thus, CBOW approach appeared to be more suitable to be integrated into our framework, and, as expected, results of preliminary experiments with Skip-gram and CBOW showed CBOW to be the better performing model.
We first constructed a vocabulary from our pre-processed tokenized corpus that contains 10,000 free-text radiology reports, and then learned vector representations of words in the vocabulary. We build our predictive model using the Gensim 2.1.0 library\(^\text{12}\). The CBOW word2vec model predicts a word given a context where context is defined by the window size. The loss function of CBOW is: \( E = -v_{w_o} \cdot h + \log \sum_{j=1}^{V} \exp(v_{w_j} \cdot h) \), where \( w_o \) is the output word, \( v_{w_o} \) is its output vector, \( h \) is the average of vectors of the context words, and \( V \) is the entire vocabulary. Once the model constructs the vectors, we can use the cosine distance of vectors to denote similarity, thereby deriving analogies. The resulting word vectors can be used as features in many natural language processing and machine learning applications.

As the training algorithm, we used both Hierarchical Softmax as well as Negative Sampling. Based on preliminary results, we found Negative Sampling to be better training algorithm. Mikolov et al.\(^\text{8}\) also described Negative Sampling as the method that results in faster training and better vector representations for frequent words, compared to more complex hierarchical softmax. The cost function of Negative Sampling is: \( E = -\log \sigma(v_{w_o} \cdot h) - \sum_{w_j \in \omega_{neg}} \log \sigma(-v_{w_j} \cdot h) \), where \( \omega_{neg} \) is the set of negative samples, \( w_o \) is the output word, \( v_{w_o} \) is its output vector and \( h \) is the average of vectors of the context words.

Finally, the document vectors were created by simply averaging the word vectors created through the trained model. According to Kenter et al.\(^\text{13}\), averaging the embeddings of words in a sentence has proven to be a successful and efficient way of obtaining sentence embeddings. Each document vector was computed as: \( v_{doc} = \frac{1}{||V_{doc}||} \sum_{w \in V_{doc}} v_{w} \), where \( V_{doc} \) is the set of words in the report and \( v_w \) refers to the word vector of word \( w \).

### 2.5 Visualization of the embeddings

Our idea is to visualize the vector representation of words and documents to validate the semantic quality of the embeddings in two different levels. In the first level, the visualization of the trained individual word embeddings can verify the positioning of synonyms (and related words), antonyms and other word-to-word relations, and can show at the very low scale that if our vector embedding is able to preserve legitimate semantics of the natural words and clinical terms. Second, the visualization of the document vectors can fulfill the purpose of analyzing the proximity of documents that have different levels of likelihood of intracranial hemorrhage. If the documents corresponding to the same class (risk) appear close to each other and form clusters, we can infer that our embedding can be useful to boost the performance of any standard classifier.

Our trained embeddings are expected to be high dimensional and may lie near a low-dimensional, non-linear manifold. Therefore, standard linear dimensionality reduction techniques (e.g. Principal Component Analysis) are not well-suited for preserving the distance between similar data points in low-dimensional representation of the vector space. We adopted t-Distributed Stochastic Neighbor Embedding (t-SNE) technique\(^\text{14}\) to visualize the trained embeddings using sklearn python library. t-SNE is a technique for dimensionality reduction that is particularly well suited to serve our application since it is capable of capturing much of the local structure of the high-dimensional data very well, while also revealing global structure such as the presence of clusters at several scales. It employs Gaussian kernel in the high-dimensional space and defines a soft border between the local and global structure of the data. For pairs of data points that are close to each other relative to the standard deviation of the Gaussian, t-SNE determines the local neighborhood size for each data point separately based on the local density of the data. We describe the results of t-SNE visualization of word and document vectors in the following section (Sec. 3.2).

### 2.6 Classification

In this study, the resulting document vectors were used as features to develop a computerized hemorrhage likelihood assessment system that aims to assign a ‘risk’ label to the free-text radiology reports while being trained on the subset of reports with the ground truth labels created by the experts (see Sec. 2.1). We observed that our dataset had imbalanced distribution of training data, i.e. class 2, 3, and 4 had fewer instances than class 1 and 5. Thus, we grouped classes 2-4, and re-defined the class labels to ensure variation of the likelihood of intracranial hemorrhage as: (1) ‘no risk’ - no intracranial hemorrhage; 2) ‘medium risk’ - probability of having intracranial hemorrhage; (3) ‘high risk’- definite diagnosis of intracranial hemorrhage. The re-definition of the class labels were validated by forming a mutual agreement between the two expert radiologists. In Table. 1, we show the number of examples per class for
the final three categories. To quantify the performance of the classifier, the 1,188 annotated reports were randomly divided into 80% training set (950 reports) and 20% test set (238 reports). To demonstrate the true power of our vector embedding, we performed experiments using three classifiers - Random Forests, Support Vector Machines, K-Nearest Neighbors (KNN) in their default configurations.

2.7 Evaluation

We experimented with different types of kernels in SVM classifier (Radial kernel & Polynomial kernel), and different values of ‘k’ in kNN (k= 5,10) classifiers. To investigate the benefits of the proposed hybrid framework, we also tested each classifier’s performance by creating vector embeddings of the radiology reports without the domain-specific semantic mapping (Sec. 2.3) where we skipped replacing the radiology terms and their synonyms using RadLex. However, we still substituted the common terms using the CLEVER base terminology for preserving the semantic structure of the radiology reports. In the Result section (see Sec. 3.3), we describe the performance of each classifier on the hold-out test set (238 reports) in a tabular format. Standard precision, recall and F1 score were used as metrics to quantify the classification performance.

3 Results

3.1 Word analogies

On feeding the entire corpus to the system, the final size of the resulting vocabulary was 4,442 words. We created word embeddings or semantic vector representations of words appearing in the corpus, from which several kinds of analogies could be derived by computing the similarity. The similarity score between the word vectors was computed as cosine similarity which is inner product on the normalized space that measures the cosine of the angle between two words:

$$\text{Similarity} = \frac{A \cdot B}{\|A\| \|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}.$$  

Table 2 shows some synonyms/closely associated words and the cosine similarity scores of their respective word embeddings. Table 3 shows some antonyms and the cosine similarity scores of their respective word embeddings. The data demonstrate that the system has formed embeddings such that pairs of synonyms have high similarity scores while antonyms have negative similarity scores.

Table 2: Similarity scores of word embeddings of synonyms/closely associated words

<table>
<thead>
<tr>
<th>Word 1</th>
<th>Word 2</th>
<th>Similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>new</td>
<td>recent</td>
<td>0.941</td>
</tr>
<tr>
<td>overinflated</td>
<td>balloon.appears</td>
<td>0.999</td>
</tr>
<tr>
<td>infarction</td>
<td>evidence_hemorrhagic_conversion</td>
<td>0.910</td>
</tr>
<tr>
<td>infarction</td>
<td>acute_infarction</td>
<td>0.928</td>
</tr>
<tr>
<td>hemorrhage</td>
<td>rightward_midline_shift</td>
<td>0.958</td>
</tr>
<tr>
<td>hemorrhage</td>
<td>subdural_hemorrhage</td>
<td>0.964</td>
</tr>
<tr>
<td>hemorrhage</td>
<td>intraventricular_hemorrhage</td>
<td>0.959</td>
</tr>
<tr>
<td>hemorrhage</td>
<td>subarachnoid_hemorrhage</td>
<td>0.968</td>
</tr>
</tbody>
</table>

3.2 Vector Visualization

Figure 3 shows the 2D visualization of word vector embedding constructed using the t-SNE approach (Sec. 2.5) where each data point represents a word. A total of 4,442 words are visualized in the figure. As seen from the figure, similar words reside fairly close together and form a cluster in the map without even inclusion of any prior knowledge. This map illustrates that our word embedding can preserve semantics of the terms.
Table 3: Similarity scores of word embeddings of antonyms, NEGEX represents negation and QUAL represents severe terms (see Sec. 2.3)

<table>
<thead>
<tr>
<th>Word 1</th>
<th>Word 2</th>
<th>Similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>large</td>
<td>NEGEX_enlarged</td>
<td>-0.245</td>
</tr>
<tr>
<td>hemorrhage</td>
<td>NEGEX_QUAL_hemorrhage</td>
<td>-0.074</td>
</tr>
<tr>
<td>hemorrhage</td>
<td>NEGEX_QUAL_intracranial_hemorrhage</td>
<td>-0.245</td>
</tr>
<tr>
<td>infarction</td>
<td>NEGEX_QUAL_infarction</td>
<td>-0.070</td>
</tr>
<tr>
<td>large_territory_infarct</td>
<td>NEGEX_QUAL_large_territory_infarct</td>
<td>-0.157</td>
</tr>
<tr>
<td>midline_shift</td>
<td>NEGEX_QUAL_midline_shift</td>
<td>-0.206</td>
</tr>
<tr>
<td>abnormalities</td>
<td>NEGEX_QUAL_abnormalities</td>
<td>-0.283</td>
</tr>
<tr>
<td>mass_effect</td>
<td>NEGEX_QUAL_mass_effect</td>
<td>-0.170</td>
</tr>
</tbody>
</table>

Figure 3: All word embeddings (4,442 words) - visualized in two dimensions using t-SNE

In Figure 4, we also highlight a group of clinical terms particularly relevant for this case-study and their negations using the same t-SNE visualization technique. The figure illustrates ability of the embedding to automatically organize concepts and implicitly learn the relationships between them. To show the word-to-word relations, we visualize only a few significant terms and their negations, but same technique can be used to infer other analogies among the terms present in our vocabulary (e.g. synonyms, antonyms, finding-finding, finding-diagnosis).

We also visualize the subsequent vectors of complete reports projected in two dimensions using the t-SNE technique (Figure 5). This visualization has been created only for the 1,188 annotated reports since the main idea is to see if our proposed embedding can be useful to compute clusters with varying risk factors. From the Figure 5, we can see that the reports denoting high risk of intracranial hemorrhage cluster together, and the reports with intermediate risk are mostly residing close to high risk reports. Though this is a two dimensional projection of the original high dimensional document vector, the result clearly shows that the embeddings carry signals that could be very informative to automatically annotate the reports using state-of-the-art classifiers.
3.3 Classification performance

We used the document vectors to classify each report into one of three classes denoting varying likelihood of intracranial hemorrhage (see Sec. 2.6). As mentioned earlier in the paper, our radiology report embedding is flexible enough to be combined with both parametric and non-parametric classifiers. We experimented with three state-of-the-art classifiers - Random Forests, Support Vector Machines and K-Nearest Neighbors (KNN).

To give more insight into the quality of the learned vectors, we used the grid search approach to tune the two main hyperparameters of our embedding for the targeted annotation, i.e. Window Size and Vector Dimension. The hyperparameter search was done individually for each classifier using cross-validation on the training data set. The effects of the hyperparameters on the resulting classifier performance are shown in Figure 6 where the optimal points of the classifier’s performance are highlighted. Based on the optimal points, we selected the hyperparameters and evaluated the classifiers’ performance on the test set. For instance, Random Forest was evaluated with the word embeddings that
Table 4: Performance of different classifiers with and without semantic mapping, and with unigrams features.

<table>
<thead>
<tr>
<th>Classifier</th>
<th>With Domain-specific dictionary</th>
<th>Without Domain-specific dictionary</th>
<th>Baseline with unigrams feature</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision</td>
<td>Recall</td>
<td>F1 score</td>
</tr>
<tr>
<td>Random Forests</td>
<td>88.64%</td>
<td>90.42%</td>
<td>89.08%</td>
</tr>
<tr>
<td>KNN (n = 10)</td>
<td>88.60%</td>
<td>89.91%</td>
<td>88.88%</td>
</tr>
<tr>
<td>KNN (n = 5)</td>
<td>88.54%</td>
<td>89.62%</td>
<td>88.76%</td>
</tr>
<tr>
<td>SVM (Radial kernel)</td>
<td>64.19%</td>
<td>80.09%</td>
<td>71.25%</td>
</tr>
<tr>
<td>SVM (Polynomial kernel)</td>
<td>63.25%</td>
<td>79.49%</td>
<td>70.43%</td>
</tr>
</tbody>
</table>

The classifiers’ performance on the test set is reported in Table 4 with optimal hyperparameters. We also present performance of the classifiers only using unigrams as features which can be considered as the baseline performance to be compared with word embedding. While the reported performance accuracy (F1-score) of baseline with unigrams is on average 71%, the word embedding resulted F1 score over 80% for most cases which demonstrates that our vector representation was able to capture the significant facets of the radiology reports. The Random Forest classifier yielded a weighted precision of 88.64% and weighted recall of 90.42% with 730 dimensional word vectors, and closely outperforms all the other classifiers used in this study. However, KNN \((n = 10)\) produces a weighted precision of 88.60% and weighted recall of 89.91% that is close to the Random Forest’s performance, employing a reduced optimal word vector dimension (130).

In Table 4, we present the classifiers’ performance with and without dictionary mapping as well as with unigrams as feature. In general, the word embedding improves the performance of the baseline classifiers and every classifier’s performance is consistently better with the proposed hybrid technique. However, performance difference is incremental for the particular case study which is hypothesized to be due to the choice of dataset in which all the reports are associated to a very narrow domain and from the same institution, i.e. CT Head reports, and thus the variation in the vocabulary is relatively small. We expect that superiority in the performance of the proposed hybrid method may be more significant when multi-topic and multi-institutional free-text reports will be considered where the semantic and syntactic variations are more prominent.

4 Conclusion

In this study, we have shown how to efficiently learn dense vector representations of individual words as well as entire radiology reports by using a hybrid technique that combines word2vec and semantic dictionary mapping. Our experimental results show that our proposed embeddings were able to learn the actual semantics of the ra-
diological terms from free-text reports. Thanks to the embeddings, we successfully annotated the radiology reports according to the likelihood of intracranial hemorrhage with 89.08% F1 score. We have publicly released (https://github.com/imonban/RadiologyReportEmbedding) our trained embeddings that have been used to test the classifiers performance (Table 4), which can be directly reused to support similar radiological applications, e.g. inferring relations between clinical terms, annotation of radiology reports, etc. The techniques introduced in this paper can be used also for creating vector representation from clinical notes of different domains (e.g. oncology) given a domain-specific ontology that can be used to reduce underlying term variations in the corpus.

In the prospective future studies, we will compare alternative neural word embedding methods (e.g. GloVe) since we believe that the performance of any such method will be boosted by the semantic mapping, as the models are initialize with random vector for out-of-vocabulary words which is far for reality. In the future version of the pipeline, we will incorporate log-likelihood ratio and mutual information for identify frequently appearing pairs, and will consider different linear functions (max pool, average pool, min pool etc.) to create document embedding from word vectors.

Acknowledgement

This work was supported in part by grants from the National Cancer Institute, National Institutes of Health, U01CA142555, 1U01CA190214, and 1U01CA187947

References

Gap Analysis and Refinement Recommendations of Skin Alteration and Pressure Ulcer Enterprise Reference Models against Nursing Flowsheet Data Elements

Karen M. Bavuso, RN, MSNa, Perry L. Mar, PhDa,b,c, Roberto A. Rocha, MD, PhDa,b,c, Sarah A. Collins, RN, PhDa,b,c

aClinical Informatics, Partners eCare, Partners Healthcare Systems, Boston, MA; bDivision of General Internal Medicine and Primary Care, Brigham and Women’s Hospital, Boston, MA; cHarvard Medical School, Boston, MA;

ABSTRACT

Reference models are an essential instrument to provide structure and guidance in the creation and use of data elements within an organizations’ electronic health record (EHR). Standardization of data elements is imperative to ensure clinical data is consistently and reliably captured for use in clinical documentation, care communication, and a variety of downstream data uses. Ongoing assessment and refinement of reference models and data elements are necessary to ascertain clinical data capture is applicable and inclusive across a variety of caregivers and domains. We performed a gap analysis on current state nursing data elements against two validated interprofessional reference models: skin alteration and pressure ulcer assessments. We present our findings along with recommendations for reference model refinements. We also highlight additional findings of inconsistencies and redundancies within data elements used for nursing documentation and highlight recommendations for improvement.

Introduction

An important aspect of an Electronic Health Record (EHR) is to capture and apply clinical data as a means to support patient care and improve outcomes.1,2 Utilizing data collected within an EHR is foundational to the National Academies vision of a ‘learning health system’, which supports that a health system can effectively use, learn from, and share health data to guide and improve the overall quality of healthcare.3,4 Standardization of clinical data is essential to effectually capture, interpret, share and reuse data in a consistent and meaningful way.5 The integrity and representation of data collected and documented within an EHR is vital for the reliability and accuracy of downstream usage.6,7 EHR data are used to support activities at the point of care, such as communicating clinical information and coordinating care within and across organizations. Additionally, secondary use of data (downstream data reuse) includes analytics, safety and quality reporting, regulatory compliance, and clinical decision support algorithms. Conversely, a lack of reliable and complete data elements results in semantic inconsistencies and poor data integrity that may increase the potential for errors when data are used to inform patient care decisions and for secondary uses.5,8,9,10

Detailed clinical data element reference models, referred to here simply as reference models, support the consistency of data capture by standardizing structured clinical data elements (SCDEs) necessary to represent a specific clinical topic.5,11 The resultant standardization of data capture ultimately promotes increased data integrity and reliability in downstream data uses. For a reference model to be most effective a detailed gap analyses should ensure that it is clinically relevant across interprofessional clinical end-users and domains and applicable to a variety of EHR workflows and functionality. Importantly, these detailed gap analyses should include refinement on an ongoing basis to incorporate new clinical knowledge or EHR functionality, even after initial validation by clinical subject matter experts.

Skin alteration assessment documentation, including assessment of wounds and pressure ulcers, is an important multi-disciplinary clinical topic that has clinical, regulatory compliance, and financial impacts to organizations. Pressure ulcers are one of the “Present on Admission (POA)” indicators hospitals are required to document for each inpatient admission.12 Moreover, based on the Hospital-Acquired Condition (HAC) Reduction Program, Stage III and IV pressure ulcers acquired during a patient’s hospitalization are not reimbursable.13 Accurate documentation of pressure ulcers upon admission, pressure ulcer risk identification, managing pressure ulcer prevention measures and appropriate documentation of pressure ulcer staging throughout a hospital stay therefore have become increasingly important to prevent erroneous incidence of non-payment of services. Furthermore, documenting wound assessment...
in a structured and consistent manner is paramount to proper assessment, treatment, and optimal wound management and outcomes.13,14

Background

Partners HealthCare System, a large Boston-based non-profit hospital and physician’s network, implemented an enterprise-wide Electronic Health Record (EHR) in May 2015. An enterprise Structured Clinical Data Element (SCDE) workgroup was formed and tasked with the principle responsibility of identifying high priority clinical topics and cataloging, evaluating, and standardizing enterprise SCDEs for those clinical topics. The group is comprised of clinical, informatics, and EHR analyst experts from across the organization.

Previous publications outlining work done within our organization describe the process used to guide the governance of reference models for use in our EHR, including development, validation, implementation, and evaluation. In one publication, a 10-step approach for reference model governance is outlined: 1) identify clinical topics, 2) create draft reference models for clinical topics, 3) identify downstream data needs for clinical topics, 4) prioritize clinical topics, 5) validate reference models for clinical topics, 6) perform gap analysis of EHR SCDEs compared against reference model, 7) communicate validated reference models across project members, 8) request revisions to EHR SCDEs based on gap analysis, 9) evaluate usage of reference models across project, and 10) monitor for new evidence requiring revisions to reference model.15 A subsequent publication specifies detailed methods used to identify priorities for the definition of clinical data reference models and to identify and resolve gaps between existing EHR data collection tools and validated reference models.11 Based on that prior work, skin alteration and pressure ulcer were identified as high priority clinical topics.11

The skin and pressure ulcer reference models were created with an interprofessional group of SMEs across our organization’s various sites and with diverse stakeholder participation. The resultant skin reference models include 48 data elements within the 10 distinct categories: 1) skin Inspection, 2) skin alteration type, location and condition, 3) skin alteration size, 4) wound bed, 5) wound tunneling, 6) wound undermining, 7) wound exudate, 8) wound drain 9) wound dressing, and 10) Skin Alteration Sub-Types. The pressure ulcer reference model includes a total of 22 data elements without additional categorization. The intent of the models is to represent basic skin and pressure ulcer documentation and does not include pediatrics or clinical specialties. The reference model was then compared to and validated against existing “form-based” enterprise data elements being used in the Though the content from legacy systems may have been leveraged in the creation of the reference models, the gap analysis was based solely on the live enterprise EHR. The form-based format of capturing clinical documentation is typically used by physicians and other care team members in the inpatient and outpatient setting. However, nursing documentation in the inpatient setting requires a somewhat different approach due to differences in workflow and clinical documentation requirements. Instead of a form-based format, a flowsheet format is used for most inpatient nursing data collection. Though nursing representation was included in the reference model creation and validation, the uniqueness of the nursing workflow, the differences in data capture needs, and the EHR format limitations required an additional detailed gap analysis (comparison of skin alteration and pressure ulcer reference models with current state flowsheet data elements used to capture skin and pressure ulcer documentation) to ensure reference model comprehensiveness and to guide refinement recommendations for implementation in nursing flowsheet documentation.

Methods

The analysis was conducted over a 10 week period of time totaling approximately 100 hours of work. An initial preparatory step in this analysis was to identify all current state flowsheet data elements used to capture skin and pressure ulcer documentation across the enterprise. Our SCDE workgroup identified all unique skin and pressure ulcer flowsheet data element sets used in our EHR for analysis. An EHR data element extraction was completed based on the identified data element sets and queried for all associated data elements and the respective value-sets. Flowsheet data elements are nested within data element sets as structures that organize small groups (e.g., typically less than 10) of clinically-related data elements. A data element may belong to multiple sets. These sets, and the data elements that comprise them, are further stratified based on their intended use and EHR functionality. In the use case of skin alteration and pressure ulcer assessments the stratifications are: 1) documentation of simple skin alteration assessment at one point in time (e.g., skin color), 2) documentation of static data for a complex skin alteration that allows for comparative assessments overtime, such as the location of a pressure ulcer, and 3) documentation of dynamic data for a complex skin alteration that allows for comparative assessments overtime, such as pressure ulcer drainage. The different types of data element sets described above assist in organizing the
flowsheet structure and functionality; however, the gap-analysis performed in this study was completed at the data element level since the value-set list associated to that data element was a significant feature of the comparison and any required alignment. The same process for comparison was used across all data elements.

For each unique flowsheet data element a gap analysis was performed using predefined metrics (See Table 1) against both the skin alteration and the pressure ulcer reference models using the following fields: 1) data element name, 2) data element set associations, 3) data element type (i.e. boolean, value list, numeric, and string), 4) value-set (where applicable). Value-sets for each category list data element were considered when determining whether a similar documentation artifact was an exact match or a partial match. This was an important designation since the aim of the comparison was to align both the data elements and any associated value-sets.

Our team has previously reported on methods used for gap analyses of EHR data elements compared against reference models. Here we use an adaptation of the MUC-5 (Fifth Message Understanding Conference) Evaluation Metrics. We used the MUC-5 Evaluation metrics to compare the 2 reference models: skin alteration and pressure ulcer against current state flowsheet data elements. These metrics identify if flowsheet data elements are a match, a partial match, extra, missing, or out of scope. Data elements used to assess skin concepts unrelated to a skin alteration were out of scope for the reference models. Additionally, because the reference models were not initially validated against the pediatric or newborn population, data elements used exclusively in pediatric or newborn sets were deemed out of scope for reference model inclusion, pending further expansion of the reference model for those populations.

Table 1. Gap Analysis Mapping Codes and Descriptors

<table>
<thead>
<tr>
<th>Mapping Code*</th>
<th>Descriptor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exact Match</td>
<td>Validated reference models and current state flowsheet data elements and values match exactly</td>
</tr>
<tr>
<td>Partial Match</td>
<td>Validated reference model and current state flowsheet data elements and value-sets partially match</td>
</tr>
<tr>
<td>Extra</td>
<td>Validated reference models do not include data elements from current state skin and pressure ulcer flowsheets</td>
</tr>
<tr>
<td>Missing</td>
<td>Validated reference model data element not included in current state skin and pressure ulcer flowsheets</td>
</tr>
<tr>
<td>Not in Scope: Pediatric/Newborn</td>
<td>Out of scope for analysis due to Pediatric/Newborn based sets only</td>
</tr>
<tr>
<td>Not in scope: Not Skin/Wound</td>
<td>Out of scope for analysis due to data element not a true skin or pressure ulcer based data element</td>
</tr>
<tr>
<td>Adaptation of MUC-5 Evaluation Metrics</td>
<td></td>
</tr>
</tbody>
</table>

Once the initial data element level mappings were complete, the extra and partial match data elements underwent a further analysis to determine recommended inclusion in the reference model using pre-determined recommendation metrics (See table 2). The value sets of the partial match category list data elements were also cross mapped against the reference model value-sets and evaluated for value-set alignment. Recommendations were considered based on scope and breadth of the reference model and clinical alignment to skin alteration and pressure ulcer concepts and integrity of clinical data capture (See table 3). Once the author performed the analysis, the methods and results were checked by a second author to ensure accuracy and completeness. Additionally, the analysis and recommendations for inclusion were reviewed with the SCDE workgroup for feedback and consensus.
Table 2. Data Element Recommendation Metrics

<table>
<thead>
<tr>
<th>Recommendation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Add to Reference Model</td>
<td>Recommendation that this data element be added to the enterprise skin or pressure ulcer reference model</td>
</tr>
<tr>
<td>2. Do Not Add-No evidence</td>
<td>Recommendation that this data element NOT be added to the enterprise skin or pressure ulcer reference model due to absent or limited evidence</td>
</tr>
<tr>
<td>3. Do Not Add-Not in Scope</td>
<td>Recommendation that this data element NOT be added to the enterprise skin or pressure ulcer reference model due to being out of scope</td>
</tr>
<tr>
<td>4. Do Not Add-Inherent in existing data element</td>
<td>Recommendation that this data element NOT be added to the enterprise skin or pressure ulcer reference model due to being included within another data element or value set of a data element</td>
</tr>
<tr>
<td>5. Align EHR to Reference Model</td>
<td>Recommendation that this data element be aligned with enterprise skin or pressure ulcer reference model</td>
</tr>
</tbody>
</table>

Table 3. Data Element Value-set Recommendation Metrics

<table>
<thead>
<tr>
<th>Recommendation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Recommend Value Inclusion in Reference Model</td>
<td>Recommendation that a flowsheet value be added to the enterprise skin or pressure ulcer reference model</td>
</tr>
<tr>
<td>2. Do Not Include-No evidence</td>
<td>Recommendation that this value-set NOT be added to the enterprise skin or pressure ulcer reference model due to absent or limited evidence</td>
</tr>
<tr>
<td>3. Do Not Include-Not in Scope</td>
<td>Recommendation that this value-set NOT be added to the enterprise skin or pressure ulcer reference model due to being out of scope</td>
</tr>
<tr>
<td>4. Do Not Include-Inherent in existing data element</td>
<td>Recommendation that this value-set NOT be added to the enterprise skin or pressure ulcer reference model due to being included within another data element or value set</td>
</tr>
<tr>
<td>5. Recommend Build Alignment to Reference Model</td>
<td>Recommendation that this data element type or value-set be aligned with enterprise skin or pressure ulcer reference model</td>
</tr>
</tbody>
</table>

Results

Data extraction results

We identified and extracted 73 unique skin and pressure ulcer flowsheet data element sets used in our EHR. 465 unique flowsheet data elements were nested within the 73 identified sets. The data elements and data element sets were broken out based on their intended functionality for documentation: 1) simple skin alteration, 2) static data for a complex skin alteration, and 3) dynamic data for a complex skin alteration (See Table 3).

Table 3. Count of Skin/Pressure Ulcer Flowsheet Data Elements per Type

<table>
<thead>
<tr>
<th>Flowsheet Set Type</th>
<th>Set Counts</th>
<th>Data Element Counts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Set A: Simple Skin Alterations</td>
<td>29</td>
<td>253</td>
</tr>
<tr>
<td>Set B: Static Data for Complex Skin Alteration</td>
<td>16</td>
<td>54</td>
</tr>
<tr>
<td>Set C: Dynamic Data for Complex Skin Alteration</td>
<td>28</td>
<td>158</td>
</tr>
<tr>
<td>Total</td>
<td>73</td>
<td>465</td>
</tr>
</tbody>
</table>
Data element gap analysis results

Figure 1 shows the results of the 465 unique current state flowsheet data elements against both the skin and the pressure ulcer reference models based on the predefined metrics. Of particular importance to the intent of this analysis are the 46% of current state flowsheet data elements mapped as “extra” - not included in the reference model, and the 29% qualified as a “partial match” - indicating that the general theme of the data elements matched, although may contain a value-set disparity or the concept was defined at a broader or narrower level in another data element. The 15 data elements (3%) that qualified as an exact match were, not surprisingly, those used within validated assessment tools (e.g. Braden scale assessment concepts) or those with a concrete definition such as with the measurement concepts (e.g. length and width), or a Boolean concept with a very specific and unambiguous meaning (e.g. “Is wound present on admission”). Interestingly, we identified 3 data elements that exist in the reference models though are missing from current state flowsheet data elements.

![Frequency of Mapping Code for EHR Data Elements Compared to Reference Model Frequencies](image)

**Figure 1. Frequency of Mapping Code for EHR data elements Compared to Reference Model Frequencies**

Out of scope data elements were included in the list of 465 extracted data elements due to their inclusion in some of the 73 sets (i.e. neurovascular wound assessment set) and general proximity in documentation due to the sequence of clinical assessments. In this example, though capillary refill is an important clinical assessment of overall skin health and potential risk of skin impairment, the intent of the reference model does not capture that breadth or specificity. These combined data elements that were not in scope totaled 21% (n=98), with 5% (n=21) defined as pediatric/newborn related and reserved for future reference model scope extension.

**Recommendation results for data element inclusion in reference model**

Of the 215 flowsheet data elements identified as extra (i.e., included in the flowsheet data element sets though not in the skin alteration or pressure ulcer reference models) 1 data element, “wound shape”, was determined to be a candidate for inclusion (See figure 2). Wound shape is considered an important part of the nursing wound assessment documentation and may be a clinically essential indicator when considering the wound location, appropriate wound measurement, and as a factor for overall wound management. The recommendation to include the “wound shape” data element was presented at the SCDE workgroup and attained group agreement with the caveat that future work include an analysis of data usage for this data element and validation with SMEs to confirm clinical relevance.

The reference models were created to comprise a broad level of skin alteration assessment and pressure ulcer assessment documentation. The scope did not include specialty areas such as NICU, pediatrics, or dermatology. Whereas the vast majority of “extra” flowsheet data elements (99%) were specialty focused, this content was not
relevant for inclusion in the models. Future work could expand the scope of our reference models to include these specialty domains.

No flowsheet data elements were coded as either: 1) Do Not Include-No evidence, or 2) Do Not Include-Inherent in existing data element. Whereas the majority of the extra flowsheet data elements were more specific than the reference models there was not a recommendation of any extra flowsheet data elements to align with the reference models (See figure 2).

![Figure 2 Data Element Recommendations for Inclusion in Reference Models](image1)

**Recommendation results for “partial match” value-set reference models/flowsheet alignment**

Of the 138 flowsheet data elements identified as a partial match (i.e. flowsheet data element similar to reference model data element), 33 flowsheet data elements had value-sets that were recommended for extension to existing reference models value-sets. Those 33 flowsheet data elements were then mapped to 9 reference model data elements (See Table 4). The skin alteration reference model includes 7 data element value-set adjustments whereas pressure ulcer includes 5 data element value-set adjustments. Three data elements overlap across both models.

Twenty-two partial match flowsheet data elements were not recommended for inclusion because the concept was inherent at a different level of specificity within another data element or value set. Two of the partial match flowsheet data elements were determined to be out of scope and no flowsheet data elements were excluded due to lack of evidence. Furthermore, it was recommended that all of those same 33 data elements and an additional 81 flowsheet data element value sets (a total of 114) be aligned with the reference models data element value-sets (See Figure 3).

![Figure 3 Data Element Value-set Recommendations for Inclusion in Reference Models](image2)
Table 4. Data Elements Value-set Inclusion to Reference Models

<table>
<thead>
<tr>
<th>Reference Model</th>
<th>Reference Model Data Element</th>
<th>Value-set Inclusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Skin alteration</td>
<td>Skin alteration type</td>
<td>Abscess, Skin Graft Donor Site, Extravasation, Fasciotomy, Flap, Frostbite, Excoriation, Hives, Lesion, Tear</td>
</tr>
<tr>
<td>Skin alteration</td>
<td>Wound bed appearance</td>
<td>(Appearance) Erythema, Hematoma, Moist (Color) Black, Yellow, Red</td>
</tr>
<tr>
<td>Skin alteration</td>
<td>Primary wound dressing types</td>
<td>Button, Eye shield, Gauze- Iodoform, Hydrofiber w silver, Hydrogel- impregnated gauze, Impregnated foam, Lap Pad, Negative Pressure Wound Therapy, Packing strip, Packing strip/iodine compound, Petroleum- impregnated gauze, Silicone, Silver, Tourney, Towel-Radiopaque, Vaginal Packing, Vessel Loop, Xeroform</td>
</tr>
<tr>
<td>Skin alteration</td>
<td>Secondary wound dressing types</td>
<td>Button, Eye shield, Gauze- Iodoform, Hydrofiber w silver, Hydrogel- impregnated gauze, Impregnated foam, Lap Pad, Negative Pressure Wound Therapy, Packing strip, Packing strip/iodine compound, Petroleum- impregnated gauze, Silicone, Silver, Tourney, Towel-Radiopaque, Vaginal Packing, Vessel Loop, Xeroform</td>
</tr>
<tr>
<td>Skin alteration</td>
<td>Associated device type</td>
<td>Prophylactic device (e.g., intermittent pneumatic compression)</td>
</tr>
<tr>
<td>Skin alteration</td>
<td>Orthopedic device type</td>
<td>Cervical collar, Prosthetic, Splint</td>
</tr>
<tr>
<td>Skin alteration and</td>
<td>Periwound condition</td>
<td>Intact skin, Crystals, Denuded, Eroded, Hematoma, Hypergranulation Hyperplasia, Ulcerated</td>
</tr>
<tr>
<td>pressure ulcer</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Skin alteration and</td>
<td>Skin/pressure ulcer alteration anatomical location</td>
<td>Cheek, Spine, Generalized</td>
</tr>
<tr>
<td>pressure ulcer</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Skin alteration and</td>
<td>Skin/pressure ulcer alteration anatomical location</td>
<td>Bilateral, Dorsal, Palmar, Plantar</td>
</tr>
<tr>
<td>pressure ulcer</td>
<td>qualifier</td>
<td></td>
</tr>
</tbody>
</table>

**Additional findings**

Throughout the gap analysis there were many instances identified of inconsistency within and across flowsheet data elements. Wound drainage color, wound description, and wound odor for example are discrete clinical wound assessment concepts, however, in some instances the concepts were grouped together into one data element with a combined value-set. Additionally there were multiple data elements describing the same concepts in differing ways. There were also unique data elements describing a concept with different value-sets.

The concept of “wound location” demonstrated a key challenge of EHR flowsheet implementation. In our reference models only two clinical concepts are needed to capture a generic wound location: 1) anatomical location (i.e. abdomen, foot) and 2) anatomical location qualifier (e.g. right, left, and medial, proximal). However, EHR flowsheets constrain the ability to reuse generic location data elements due to ambiguity in associating the correct location data with the correct skin alteration for a patient with multiple skin alterations. There was also notable variation in the value-sets for these anatomical location data elements that was not related to flowsheet constraints. In some instances, the value-sets included both the location and the location qualifier (i.e. abdomen, foot, right, left, medial, proximal) (See Table 5). In addition to the redundancy and overlap amongst wound location data elements, there were yet another group of data elements for location that were free text and were used for simple skin alteration assessments.
Table 5. Inconsistencies of Value-sets Within and Across Data Elements

<table>
<thead>
<tr>
<th>Data Element</th>
<th>Value-sets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Site Location</td>
<td>Right, Left, Lateral, Midline, Upper, Lower, Abdomen, Ankle, Arm, Chest, Elbow, Flank, Hand, Hip, Knee, Leg, Shoulder, Umbilicus, Wrist, Other (comment)</td>
</tr>
<tr>
<td>Location</td>
<td>Abdomen, Ankle, Anus, Arm, Axilla, Back, Breast, Bridge of nose, Buttocks, Chest, Coccyx, Ear, Elbow, Eye, Face, Finger (Comment which digit), Foot, Forehead, Generalized, Groin, Hand, Head, Heel, Hip, Ischial tuberosity, Jaw, Knee, Labia, Leg, Lip, Malleolus, Meatus, urinary, Mouth, Nares, Nasal Septum, Neck, Nose, Pelvis, Perineum, Rib Cage, Sacrum, Scapula, Sciera, Scrotum, Shoulder, Spine, Sternum, Thigh, Throat, Tibia, Toe (Comment which digit), Trach site, Trocanter, Umbilicus, Vagina, Wrist, Other (Comment)</td>
</tr>
</tbody>
</table>

Discussion

Based on the gap analysis it is our recommendation that the skin alteration and pressure ulcer reference models are refined to include 1 additional data element, “wound shape”, and 9 value-set adjustments. We also recommend that 114 of the current state flowsheets data elements be modified to align with the validated reference models to minimize redundancy and inconsistency across data elements and value-sets. Moreover, data type alignment to minimize free text documentation is an important part of flowsheet adjustments. We also recommend future analysis to identify additional inconsistencies in flowsheet data elements and value-sets and usage data analysis to help determine clinical need and guide further refinement.

Overall, the results of the analysis provide evidence that the validated skin alteration and pressure ulcer reference models sufficiently comprise a broad level of skin and wound clinical concepts required to accurately capture nursing domain clinical concepts within flowsheet documentation. The findings also provided useful insight into inconsistent and overlapping data elements, as well as free text data elements, that require adjustment to align to the reference models and the need for future analysis to identify inconsistencies.

Pragmatic reference models are an important mechanism to guide the end user in collecting clinical data in a consistent and meaningful manner. Without a means for relevant and complete data capture, downstream data usage and patient care are negatively impacted. Continuous refinement of reference models used within an EHR is an important activity to ensure relevant and complete clinical data capture. Similarly, continuous refinement of the EHR over time is also an ongoing process that requires appropriate resources. Though the skin alteration and pressure ulcer reference models were initially validated with SMEs and for implementation in the form-based documentation in our EHR, the nursing flowsheet data elements required additional review and validation.

Importantly, reference models should remain pragmatic, reflecting real use needs and any constraints imposed by EHR systems.

Next steps

The data element usage data for the recommended additions to the reference models will provide insight into how many times the data element “wound shape” is being used as an indicator of clinical relevance and rationale for exclusion. Future work evaluating skin alteration and pressure ulcer assessment concepts, and other clinical reference model topics, should include evaluation of usage data as a tool to identify relevant and important documentation fields. Similarly, usage data can provide information about nurses’ interactions with the data element value-sets for optimization of values. Additionally, validation of the skin and pressure ulcer reference models by comparing against the EHR content of other organizations may help to reduce any unintended institutional bias and to assure external validity of our findings.

Limitations

Data was limited to 1 EHR implementation and 2 reference models developed and validated at our organization. Process should be reusable at other size and types of healthcare organizations, even if models and flowsheets are different. Findings may vary however at other sites and gap analyses should be conducted at other sites to identify generalizability of findings for reference model refinement recommendations.
Conclusion

Refinement of existing reference models that guide SCDE build and resultant data collection tools across an enterprise EHR system is an important activity. Skin alteration and pressure ulcer documentation is considered a high priority model due to regulatory, billing, and downstream CDS impact. Through a gap analysis of current state data elements against our skin alteration and pressure ulcer enterprise reference models we confirmed the comprehensiveness and relevance of the models to accurately capture nursing domain clinical concepts for flowsheet documentation. Based on the analysis, one data element, “wound shape”, was identified as a potential addition to the pressure ulcer model. Additionally, nine data elements are recommended for value-set expansion. We also identify many instances of current state data elements and associated value-sets that are not aligned with the models and have made recommendations for EHR modifications.

Acknowledgements

We would like to acknowledge the Structured Clinical Data Element workgroup and members of the development teams from the Partners eCare project and members of the subject matter expert panels that contributed to this work.

References


Eliciting Values of Patients with Multiple Chronic Conditions: Evaluation of a Patient-centered Framework

Andrew B.L. Berry, MS1, Catherine Lim, MDes2, Andrea L. Hartzler, PhD1,2, Tad Hirsch, PhD1, Evette Ludman, PhD1,2, Edward H. Wagner, MD, MPH2, James D. Ralston, MD, MPH1,2

1University of Washington, Seattle, WA, USA; 2Kaiser Permanente Washington Health Research Institute, Seattle, WA, USA

Abstract

Patients with multiple chronic conditions often face competing demands for care, and they often do not agree with physicians on priorities for care. Patients’ values shape their healthcare priorities, but existing methods for eliciting values do not necessarily meet patients’ care planning needs. We developed a patient-centered values framework based on a field study with patients and caregivers. In this paper we report on a survey to evaluate how the framework generalizes beyond field study participants, and how well the framework supports values elicitation. We found that respondents frame values in a way that is consistent with the framework, and that domains of the framework can be used to elicit a breadth of potential values individuals with MCC express. These findings demonstrate how a patient-centered perspective on values can expand on the domains considered in values clarification methods and facilitate patient-provider communication in establishing shared care priorities.

Introduction

In the United States, two-thirds of adults over 65 have multiple chronic conditions (MCC), and this population continues to grow.1-2 Compared to individuals with single chronic conditions, individuals with MCC have poorer quality of life, more physical disabilities, more frequent adverse drug events, and higher mortality.3 Individuals with MCC face competing care demands, especially when self-care for one condition conflicts with self-care for another condition. For example, a patient may want to exercise to improve cardiovascular health and control their diabetes, but if they also have osteoarthritis, exercise could be very painful. Conflicting self-care demands are further compounded by the everyday complexities of living with chronic illness, including individual preferences, resources, and diverse lifestyles.4 Choosing the right care priorities for individuals with MCC is important since the ability to self-manage competing demands can affect the quality and length of life.5

To help individuals with MCC and their health care providers establish shared care priorities, first we must understand patients’ values. Patients’ personal values—what they “consider important in life”6—underlie the decisions they make about their health and health care. Health services research has emphasized the importance of understanding and honoring patients’ values, particularly for establishing shared care priorities with MCC patients.7 Although the importance of orienting care around patient-important outcomes is well documented,8 few clinical guidelines include ways to incorporate patients’ personal values in care planning.9,10

Many research studies have designed and evaluated methods for providers to elicit and clarify patients’ values, particularly for shared decision making.11-14 While diverse, most values clarification methods (VCMs) focus on specified treatment preferences that are pre-defined without the ability for patients to introduce their own concerns or explore the underlying values that shape their care preferences.11 Since patients with MCC mostly manage health outside of clinical contexts, it is important for patients and providers to communicate about what patients value in their daily lives when establishing shared care priorities. Our prior work underscores this need by revealing that some individuals with MCC filter what they share with healthcare providers because they do not perceive personal values from daily life as pertinent to their healthcare.15 There is a need to better understand personal values from the perspective of patients to inform the design of tools that can improve patient-provider communication about care planning.

Based on prior field work15-17 with 31 individuals with MCC and 17 family caregivers, we formulated a patient-centered values framework that characterizes the types of personal values that patients described as most important to their well-being and health. We conducted home visits that included photo elicitation18 as part of a semi-structured interview. Prior to the visit we mailed the patient a camera that produced instant photographs and asked them to take pictures to show what was important to their well-being and health in daily life. When we visited the
patient’s home, we began the semi-structured interview by asking them to explain what was important in each photograph. Next we discussed topics including daily activities, self-management, tradeoffs (e.g., times when the patient found it difficult to balance what was important to them), how their values related to their health, and how they communicated with their healthcare providers. Interviews were recorded and transcribed verbatim. Analysis of the transcripts was guided by grounded theory methods, and included open coding, focused coding, and writing memos. Two authors (AB, CL) open coded transcripts independently, wrote code memos to define and clarify codes, and met regularly to coordinate, refine, and consolidate codes. Wherever possible, we labeled and defined codes using the language of participants. As codes stabilized, we conducted focused coding on the remaining transcripts, and recoded the initial transcripts to ensure consistency. As we coded, we wrote memos to group codes into themes. These themes are the domains that comprise the patient-centered values framework. Throughout the analysis process, all authors participated in regular meetings to discuss and clarify codes and themes. A paper reporting this field study and the resulting patient-centered values framework is currently under review.

The framework expresses the breadth of patients’ personal values across six domains: principles, relationships, emotions, activities, abilities, and possessions. Principles are beliefs and standards to live by, such as honesty or independence. Relationships are connections with others, such as family or friends. Emotions are feelings, affect, or mood, such as joy or serenity. Activities are pursuits such as working, reading, gardening, or volunteering. Abilities are physical or mental capacities or skills, such as mental sharpness or mobility. Possessions are tangible objects or spaces, such as photographs, a car, or a woodshop.

The values framework expands the breadth of domains considered in values clarification methods and could be used to inform patient-provider discussions about shared care priorities. Yet, it is important to evaluate how well this framework translates beyond individuals in our prior field work to other individuals with MCC. The aim of the study we report here was to evaluate the values framework through a survey with a larger sample of individuals with MCC and inform the design of tools that facilitate communication between patients with MCC and their healthcare providers about personal values in the context of care planning. We investigated two research questions:

RQ1. How similar to the framework do respondents frame values? (i.e., Do they generate similar kinds of values that participants in the prior field study described? Do they assign values to similar categories?)

RQ2. Is there a difference in how respondents rate the importance of values from different domains in the framework?

Related Work

We adopt the definition of values from Friedman et al.: “what a person or group of people consider important in life.” Although values are central to the patient-provider relationship, there remain critical gaps in how to elicit patients’ values and incorporate them into care planning discussions. We investigate how evaluation of our patient-centered values framework could build on and overcome limitations of existing values clarification methods by considering values from the perspective of patients.

There has been significant research into values clarification methods (VCMs), which are techniques to help patients clarify what matters to them in the context of specific health-related decisions, such as decision aids that utilize pros and cons or ratings. Examples of VCMs include decision aids to help patients choose the best treatment for type 2 diabetes and decision aids to help patients select lifestyle changes for improving cardiac health. Witteman et al. reviewed 98 VCMs, developed a taxonomy of VCM design features, and discussed limitations of VCMs. Most methods reviewed were designed for patients to complete as an independent activity (59%) rather than with a provider, used closed-ended and pre-set options (61%), and few encouraged patients to explore values as part of an iterative discovery process (9%). These key limitations could increase the chance that providers miss opportunities to identify what is important to patients broadly, and fail to incorporate those values into care planning that is so important for MCC. Furthermore, care planning for individuals with MCC involves juggling multiple concerns and decisions, but existing VCMs generate values about single health conditions. Out of 98 VCMs reviewed, 23 related to chronic illness, but none explicitly addressed MCC. Of the 23 VCMs related to chronic illness, nine focused on cardiovascular health, eight focused on other chronic conditions, and six focused on advance care planning. This suggests a lack of VCMs for care planning for individuals with MCC who face conflicting self-care demands.

The first limitation identified by Witteman et al. that is relevant to our goal of informing the design of VCMs for individuals with MCCs is that most VCMs are completed by patients independently. In our work, we conceive of the patient-centered values framework as a conversational tool to stimulate patient-provider discussions about values.
The second limitation of VCMs relevant to our study is that most VCMs are closed-ended and do not enable individuals to generate and incorporate their own attributes or elements for consideration. For example, the VCM designed by Breslin et al. helped individuals decide which medication to use to treat type 2 diabetes based on attributes such as how each medication affected blood sugar and weight, but these attributes were predefined, not patient-generated. Of the VCMs Witteman et al. reviewed, 60 out of 98 did not allow the individual to add elements or attributes to a decision. In this study, we sought to explore the nature of open-ended “free response” elicitation of values compared to closed-ended elicitation of predefined “domain-driven” values. There may be benefits and drawbacks to each approach. Understanding how responses to these types of elicitation methods could vary will inform the design of VCMs to support individuals with MCC.

The third key limitation of VCMs identified by Witteman et al. is that VCMs often offer little support for iterative exploration of values. Out of 98 reviewed, 21 VCMs explicitly did not allow users to explore and revise their values, but rather required they identify and express values in a single attempt. Another 65 technically allowed revision (e.g., to page back and change response) but did not explicitly encourage it. It is important for patients to be able to revise their values as their understanding of a decision develops. In this study, our survey structure enables us to develop a better understanding of how elicitation of multiple values in series might affect the kinds of values participants shared. In addition, we can make some judgments about how different kinds of elicitation methods in series (e.g., initial free-response followed by domain-driven followed by a second round of “informed free-response”) might elicit values of different types, specificity, or importance to the patient.

Given these limitations of existing VCMs, we sought to evaluate how our values framework might be used to inform the design of methods for eliciting patient values that are applicable beyond singular health-related decisions to meet the needs of individuals with MCC. Specifically, we explored the utility of “free-response” versus “domain-driven” elicitation of values, and we explored the relative importance individuals place on domains within the framework.

**Methods**

To evaluate our values framework, we conducted a phone survey to ask patients with MCC to generate and categorize values into domains (RQ1) and then rate the importance of those domains (RQ2). After completing the survey, participants received $30 by mail. Survey procedures were approved by the Institutional Review Board at Kaiser Permanente Washington Health Research Institute (KPWHRI).

**Recruitment**

Participants were recruited from Kaiser Permanente Washington (KPW), an integrated healthcare system in Washington State. Participants were required to have type 1 or type 2 diabetes and at least two of the following common conditions: depression, osteoarthritis, or coronary artery disease. Eligibility was determined based on diagnosis codes stored in the electronic health record. We chose these illnesses because all require a high degree of self-management, and self-management tasks for these conditions may either overlap or compete with one another. For example, recommendations to improve outcomes for diabetes and coronary artery disease encourage physical activity, but arthritis might limit significant physical activity. In other combinations, such as diabetes and depression, treating both together is optimal.

Participants had to have a primary care provider in a KPW clinic, had to be enrolled in a KPW plan at the time of recruitment, and had to have had at least two visits with a primary care physician in the preceding 12 months. Participants were excluded if they had been diagnosed with HIV, AIDS, dementia, or other major psychiatric diagnoses for the preceding 10 years, and were excluded if in the preceding six months they had two or more in-person visits to medical oncology or radiation oncology or if they were enrolled in hospice or palliative care. We purposively sampled for educational status, aiming for a distribution similar to the U.S. population, which is approximately 50% with educational level high school or less, and oversampled minority racial and ethnic populations. Eligible patients who met the sample criteria received a letter explaining the purpose of the survey and the survey procedures, then received a phone call from a member of the KPWHRI Survey program to administer the survey by phone. In total, 54 respondents completed the survey. Of those, 17 had participated in our prior field study and 37 had not.

**Procedures and Analysis**

The phone survey had three parts: (1) generating and categorizing values; (2) rating the importance of values; and (3) demographics.
Generating and categorizing values: We approached the generation of values in three ways: free-response, domain-driven, and informed free-response (Table 1). The survey first asked respondents to name two things that are important to their well-being and health through free response (values 1-2). Respondents generated these values freely without knowledge of the domains in the values framework (i.e., “free-response”). Second, the survey asked respondents to generate values for each of the six domains from our patient-centered values framework after being read a short definition for each (values 3-8). During this “domain-driven” approach, respondents were given the choice to not provide an example if they felt the domain was not important to them. Third, respondents were given the option to provide up to five additional values of their choosing (values 9-13). These optional additions reflect values elicited through free-response after respondents were exposed to domains of the values framework (i.e., “informed free-response”).

Table 1. Elicitation methods to generate patient values

<table>
<thead>
<tr>
<th>Method</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Free-response</td>
<td>1-2</td>
</tr>
<tr>
<td>“Name two things that you feel are important to your well-being and health.”</td>
<td></td>
</tr>
<tr>
<td>Domain-driven</td>
<td>3-8</td>
</tr>
<tr>
<td>“I will read a list of categories that some people report as important to their well-being and health. I will ask you for an example of something within that category that is important to your well-being and health.”</td>
<td></td>
</tr>
<tr>
<td>Informed free-response</td>
<td>9-13</td>
</tr>
<tr>
<td>“Can you think of anything else important to your well-being and health?”</td>
<td></td>
</tr>
</tbody>
</table>

For categorizing values, the survey asked respondents to assign the free response values (values 1-2, and 9-13) to the six domains (i.e., principles, relationships, emotions, activities, abilities, and possessions). For each value (e.g., “value x”), participants responded to: “Earlier you told me that ‘value x’ is important to you. Of the categories we just discussed, where does ‘value x’ belong: activities, possessions, relationships, emotions, principles, or abilities?” Respondents were given the option to assign a value to multiple domains if they wished. If the respondent did not believe the value fit in any of the domains, the respondent was given the option to name a new domain where the value fit, and to provide a description.

To answer RQ1, we first examined the coverage of values generated through free-response, domain-driven, and informed free-response elicitation methods by counting the number of values resulting from each of the three methods. We then examined how respondents categorized those values by assigning them to domains of our values framework. We inspected values that respondents chose not to assign to one of the six domains and any new domains they suggested. Finally, we compared how respondents categorized values with how the research team categorized those same values. Team-based categorization was reached through consensus and was done in a blinded manner in which team members were unaware of how respondents categorized values. Three members of the research team (“coders”) assigned by consensus each respondent-generated value to one of the six domains from the values framework. Coders assigned values to domains based on domain definitions formulated from our prior field study based on home interviews. Values that did not fit any domain, that lacked sufficient context to assign to a domain, or for which consensus could not be reached were distributed to two additional team members to assign to domains independently. The three coders used the independent assignments to reach consensus on categorization of the remaining values. We assessed agreement between respondent-categorized and team-categorized values with Cohen’s Kappa (K). Throughout the coding process, the three coders and two additional team members maintained a record of responses for which reaching a consensus categorization required discussion. This record included notes clarifying the definitions of domains from the values framework and notes about potential extensions to the framework.

Rating the importance of values: The survey asked respondents to rate how important each of the values they generated (i.e., values 1-13) are to their well-being and health on a 5-point Likert scale (anchors were 1 = “important”, 3 = “very important”, and 5 = “the most important”). To answer RQ2, we compared respondents’ importance ratings among value domains with a Friedman test ($\chi^2$) and conducted post hoc pairwise comparisons using Mann-Whitney U tests with a Bonferroni correction. We chose Friedman and Mann-Whitney U as non-parametric tests because the data are Likert ratings and not normally distributed. We applied the Bonferroni correction to reduce the chance of type I error with multiple pairwise comparisons.
Demographics: The survey asked respondents to provide their age, sex, race, ethnicity, and education level, which we summarized with descriptive statistics in Microsoft Excel.

Results

Table 2 shows demographics of the survey respondents.

Table 2. Respondent demographics (n=54)

| Age mean (SD) | 65.4 (10.7) |
| Age range     | 45-87       |
| Sex n (%)     | 31 (57.4%)  |
| Race n (%)    |             |
| American Indian/Alaska Native | 1 (2%) |
| Asian         | 4 (7%)      |
| Black/African American | 6 (11%) |
| Native Hawaiian/Pacific Islander | 1 (2%) |
| White         | 31 (57%)    |
| More than one race | 8 (15%) |
| Other         | 3 (6%)      |
| Hispanic/Latino n (%) | 5 (9%) |
| Education n (%) |     |
| Some high school, but not a graduate | 4 (7%) |
| High school graduate or GED | 10 (19%) |
| Some college or 2 year degree | 24 (44%) |
| 4-year college degree | 5 (9%) |
| More than 4-year college degree | 11 (20%) |

Generating and categorizing values

Free-response: The 54 respondents generated a total of 107 values through free-response (i.e., values 1-2). Only one respondent did not generate both values. Free-response values were diverse—some were medically oriented (e.g., “good medical care”, “my heart”, “keeping my cholesterol down”, “good communication with doctors”) whereas others reflected everyday priorities that were personal in nature, such as “family”, “happiness”, and “proper frame of mind”. Respondents categorized 103/107 (96%) of these values into one or more of the six domains of the values framework. In fact, respondents applied 69 of the 103 values (67%) to 2 or more of the 6 domains. The distribution of assignment of the 103 values to domains was: relationships (56%), abilities (55%), activities (54%), emotions (47%), principles (33%), and possessions (22%). Because respondents could assign each value to more than one domain, the total of this distribution is greater than 100%. Respondents suggested new domains for 3 of the 4 uncategorized values, including the domain “self-discipline” to assign the value “get out of pain with my back”, the domain “health provider and patient relationships” to assign the value “concern”, and the domain “diet and exercise” to assign the value “keeping my blood sugar low”. A new category was not suggested by the respondent for the remaining uncategorized value “being independent”.

The research team categorized 88/107 (86%) values generated from respondents through free-response. The 19 uncategorized values lacked sufficient detail for the team to assign to any of the six domains, such as “my weight”, “cost”, “having my health care records be accurate and up to date”, “prompt appointments with my primary doctor”, “getting my medication when I need them”, and “health issues are seen in the context of my life”. Agreement between respondent-categorization and team-categorization of the 88 values was moderate, $K = 0.47$. Because we allowed respondents to assign values to more than one domain (i.e., multiple categorization), this estimate may be higher than if we required each value to be assigned to only one domain; the research team’s categorization could match with any of the domains to which respondents assigned a given value.
Domain-driven: Respondents generated a total of 318 values that were specific examples across the six domains (i.e., values 3-8). Nearly all respondents generated a value for all six domains. Six respondents (11%) were able to generate examples for only five domains. One respondent could not generate an activity and two respondents reported that possessions were not important to their well-being and health. One respondent reported “I don’t know” when asked to provide an example for the “relationships” domain. Two respondents reported “I don’t know” when asked to provide an example for the “emotions” domain.

By the nature of the task, the domain-driven values were categorized by respondents. The research team categorized 310/318 (98%) of these values. The 8 uncategorized values lacked sufficient detail for the team to assign to any of the six domains. Examples included respondent-generated principles “listening” and “caring” and respondent-generated abilities “communication” and “my health.” Agreement between respondent-categorization and team-categorization of the 310 values into the six domains was high, $K = 0.85$. The proportion of respondent-categorized and team-categorized values that agreed included: 96% relationships (n=53), 94% possessions (n=51), 92% activities (n=53), 88% principles (n=52), 86% emotions (n=51), and 68% abilities (n=50).

Informed free-response: After eliciting values through free-response and domain-driven methods, we asked respondents if there was anything else important to their wellbeing and health. Respondents could generate up to 5 additions (i.e., values 9-13) and resulted in a total of 78 values. Thirty-three respondents (61%) added one more value, 19 (35%) added 2 more values, 13 (24%) added 3 more values, 10 (19%) added 4 more values, and 3 (6%) added 5 more values. Examples tended to reflect everyday priorities that were personal in nature (e.g., “family”, “hobbies”, “feel safe”, “being able to go out and do things”) and few medically oriented values (“following closely the directives of my doctor”). Respondents readily categorized all but 2 of these 78 additional values (97%) into one or more of the six domains. The distribution of assignment of the 76 values to domain was: emotions (66%), activities (66%), relationships (63%), abilities (61%), principles (42%), and possessions (34%). This distribution is similar to initial free-response, but the percentage assignment increased for each category. The 2 uncategorized values were “watching my security” and “ability to take off time for myself to not worry about health”. Respondents applied 60 of the 78 values (77%) to multiple domains.

The research team categorized 74/78 (95%) values that respondents optionally added. The 4 uncategorized values lacked sufficient detail for the team to assign to any of the six domains, which included “my health”, “being healthy”, “having challenges”, and also included “watching my security” which was uncategorized by respondents. Agreement between respondent-categorization and team-categorization of the 73 values into the six domains was high, $K = 0.78$. Similar to free-response values, we suspect agreement was high due to multiple categorization.

Relative importance of value domains
Table 3 shows importance rating for values generated through free-response, domain-driven, and informed free-response methods. For domain-driven values, relationships were rated as most important on average whereas possessions were rated as least important on average. There was a significant difference in importance ratings among the 6 domains ($\chi^2 = 44.84, p <0.001$). This analysis excludes ratings for the six respondents who did not generate examples for all six domain-driven values. Post hoc comparisons resulted in 4 significant differences in importance ratings between: relationships and possessions ($U = 639, p<0.001$), relationships and activities ($U = 599, p<0.001$), principles and activities ($U = 671, p=0.002$), and principles and possessions ($U = 707, p=0.004$). These findings follow from the most important (i.e., relationships and principles) and least important (i.e., activities and possessions) according to ratings.

### Table 3. Importance ratings for values assigned across methods

<table>
<thead>
<tr>
<th>Elicitation method</th>
<th>Mean (SD)</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Free-response (n=107)</td>
<td>4.5 (0.8)</td>
<td>1-5</td>
</tr>
<tr>
<td>2. Domain driven (n=318)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Relationships (n=53)</td>
<td>4.7 (0.7)</td>
<td>1-5</td>
</tr>
<tr>
<td>Principles (n=54)</td>
<td>4.6 (0.7)</td>
<td>1-5</td>
</tr>
<tr>
<td>Emotions (n=51)</td>
<td>4.4 (0.9)</td>
<td>1-5</td>
</tr>
<tr>
<td>Abilities (n=54)</td>
<td>4.2 (1.0)</td>
<td>1-5</td>
</tr>
</tbody>
</table>
Discussion

Principal findings from this survey are that respondents frame personal values in a way that is consistent with our patient-centered values framework—they assigned nearly all values to at least one domain of the framework, and agreement of that assignment with the research team was high (RQ1). Respondents rated the importance of domains of values significantly different—namely, they rated relationships and principles significantly more important than activities and possessions (RQ2). In addition, we gathered insights that open-ended methods may encourage elicitation of both medically oriented and personal values connected with everyday lives outside the clinic, especially when offered the opportunity to express free-response values after learning the domains of the values framework.

**Precision and Clarity of Values Domains**

The level of agreement between respondents’ and team members’ categorization of personal values tells us about the precision with which value domains were defined, and the clarity with which respondents perceived the domains. There was moderate to high agreement for each elicitation method, with the highest agreement for domain-driven elicitation. Participants could generate personal values in response to the domains, and the research team consistently placed those values into the same domains from which they were elicited. This finding provides evidence that the definitions of the domains are conceptually precise, and that the wording we used to describe the domains to patients is clear. Table 4 includes each value domain, its definition, and representative examples. We updated the definitions of principles, emotions, and possessions to clarify their meaning. These updates reflect discussions we had during consensus categorization of survey respondents’ values. For principles, many respondents shared standards and virtues, which were reflected in the original definition, but respondents also shared aspirations that guided their behavior much like standards or virtues. For emotions, respondents often expressed values that were embodied and experiential, such as comfort, or relief from pain. Finally, respondents highlighted the nature of their relationship to possessions, such as objects being owned or cherished. These updated definitions improve the clarity and accuracy of the patient-centered values framework and boost its utility for use in values elicitation and care planning.

**Table 4. Patient-centered Values Framework: Six domains of patient values**

<table>
<thead>
<tr>
<th>Domain</th>
<th>Definition</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Principles</td>
<td>Standards or virtues to live by, including aspirations</td>
<td>Spirituality, independence, truth</td>
</tr>
<tr>
<td>Relationships</td>
<td>Connections with others</td>
<td>Family, friends, community</td>
</tr>
<tr>
<td>Emotions</td>
<td>Feelings or mood; states of being that are personal, embodied, and experiential</td>
<td>Accomplishment, comfort, serenity</td>
</tr>
<tr>
<td>Activities</td>
<td>Pursuits, things people do for work or leisure</td>
<td>Reading, gardening, self-care</td>
</tr>
<tr>
<td>Abilities</td>
<td>Physical or mental capacities or skills</td>
<td>Mental sharpness, mobility, vision</td>
</tr>
<tr>
<td>Possessions</td>
<td>Tangible things kept, owned, or cherished, including spaces</td>
<td>Computer, ’55 Chevy, home, woodshop</td>
</tr>
</tbody>
</table>

**Generalizability of Framework**

Because respondents could categorize nearly all values from free-response and informed free-response, this provides evidence that the values framework can account for a breadth of potential values patients with MCC express. More work is needed to further substantiate the generalizability of the framework, since it is possible that individuals with characteristics different than our sample might have values that do not fit the framework. Our sample is limited to
individuals in Washington State who have insurance, receive care within an integrated health system, and have a specific spectrum of chronic conditions. Also, 17 of the 54 respondents participated in a previous study in which we asked them to photograph things that were important to their well-being and health, and then interviewed them about those values. Those interviews were included in the 31 used to develop the patient-centered values framework.15–17 We cannot rule out that this overlap in samples limits the generalizability of the findings.

Utility of Priming for Free-Response Elicitation
Placement of domain-driven elicitation in the middle of the survey allowed for two types of free-recall elicitation: elicitation that occurred before participants learned of the values framework (i.e., “initial free-response”), and elicitation after the framework was introduced (i.e., “informed free-response”). Comparing agreement across initial free-response, domain-driven elicitation, and informed free-response, both the type of elicitation (free-response versus domain-driven) and the participant’s familiarity with the values framework (initial versus informed free-response) may be associated with different levels of agreement on categorization between respondents and the research team. Higher agreement for domain-driven versus initial free-response elicitation may suggest that domain-driven elicitation generates values that align more closely with domains of the framework. Thus, the type of elicitation may influence the level of agreement. However, higher agreement for informed free-response than initial free-response suggests that familiarity with the framework may also influence the level of agreement. Future studies could further explore reasons for these differences, and assess how such insights might inform the design of approaches for eliciting patient values.

Review of survey responses suggests qualitative differences in the content and categorization of values elicited through initial free-response versus informed free-response. Respondents’ values generated from initial free-response tended to be more medically oriented, focused on the healthcare system (e.g., “time at doctor visits”), than values generated through informed free-response elicitation. Our prior work15–17 suggests that the field study methods used to formulate the values framework (i.e., photo elicitation and extended, semi-structured interviews in participants’ homes) were effective for encouraging participants to think about well-being and health in their daily lives, outside of healthcare contexts. This kind of in-depth values elicitation may not be feasible in clinical settings due to time and resource constraints. We have also found that patients tend to withhold personal values in clinical contexts when they do not perceive those values to be pertinent.15 Given the time and resource intensiveness of field study methods, it is promising that introducing the domains of the values framework has potential to broaden the values patients share. Domain-driven and informed free-response queries appear to be efficient approaches to eliciting values that might otherwise require in-depth examination. Together, our prior work and the current findings support the idea that priming patients with definitions of values domains should be a key consideration when eliciting values of patients with MCC. Future work could explore how different methods of priming participants yield different kinds of values, as well as the clinical utility of the types of values elicited through different methods.

Relative Importance of Domains
We did not observe a difference in the importance respondents ascribed to values elicited by free-response and informed free-response elicitation. Respondents rated the importance of free-response values very high and these ratings were close to the highest-rated domain-driven values. Further, we observed a significant difference in the high rating of importance for relationship and principles domains compared with the lower rating of importance for possessions and activities domains. More work is needed to understand the care and design implications based on differences in ratings for different domains. A rating of 3.8 or 3.9 still falls between “very important” and “most important,” so our findings support including these domains when eliciting the values of patients with MCC for overall well-being and health. More research is needed to understand what these ratings mean in the context of care planning and decision-making.

Multiple Categorization
When given the choice, respondents often placed free-response values into multiple categories. Multiple categorization probably inflated agreement substantially, since our team’s categorization of a value could match on any domain to which the respondent assigned the value. This choice is both a limitation and a strength of our study design. The fact that respondents employed multiple categorization is interesting and insightful. From a conceptual standpoint, the prevalent use of multiple categorization among respondents supports the idea that the value domains might not be mutually exclusive. This is compatible with findings from our prior work that individuals with MCC often perceive values to be interrelated.16 For instance, if an individual values “gardening with her granddaughter”, this same value may be categorized under activities (i.e., gardening) and relationships (i.e., with granddaughter). Looking toward the design of tools for eliciting values for planning care, the prevalence of multiple categorization
means it may be more likely that domain-driven elicitation generates values that cross domains than values that do not. As long as individuals are given the opportunity to include more than one value per domain, this approach may also provide a more comprehensive set of values compared to free-response alone. Understanding how values cross domains and relate to other values has implications for care planning. For example, when facing declining health, patients and their health care providers might consider how this could affect a web of related values and values domains rather than focusing solely on medically-oriented effects, such as physical function. Methods and tools could be designed to facilitate discussions that link health concerns with patients’ values, and support care planning to mitigate these concerns.

Implications for Informatics

These findings have practical applications for health information technology to support conversations between patients and providers about patients’ personal values. A patient-facing questionnaire deployed in a personal health record (PHR) could capture personal values from patients and display them to the provider. This would enable the patient to respond while at home, which could facilitate reflection on personal values in daily life. The questionnaire could be formatted to mimic the elicitation methods tested in our survey. Such a questionnaire would be more likely to elicit a breadth of patient values if it used domain-driven or informed free-response methods. As we know from limitations in VCMs, patients would need to be able to revise and elaborate on personal values over time. Patients’ personal values could then be displayed in existing locations in clinical information systems, such as alongside patient demographics or social history, or could be incorporated into care planning or decision support tools.

Conclusion

We evaluated how a patient-centered values framework generalizes to a larger sample of patients with MCC. Through a series of free-response and domain-driven queries, we elicited respondents’ values, asked them to place those values into the domains of the framework, and asked them to rate the importance of values generated. We found that respondents framed values similarly to the framework, suggesting that the domains of the framework are clear and conceptually precise, and that the framework provides a promising tool for eliciting patient-centered values. These findings provide insight into the potential utility of the values framework for supporting care planning conversations between patients with MCC and their health care providers. Future work should build on these findings by exploring ways to link patient values with health concerns in the context of care planning, what patients’ ratings of the importance of values mean in the context of care planning, and the clinical utility of values generated through different elicitation types.

Acknowledgments

We thank Linda Kiel, Zoe Bermet, Luesa Jordan, and Deryn Haug for their valuable contributions to this project. We are grateful for support from the Agency for Healthcare Research and Quality (R01 HS022364). The opinions expressed here are ours and do not reflect the position of AHRQ or the U.S. Department of Health and Human Services.

References


16. Lim C, Berry AB, Hirsch T, Hartzler AL, Wagner EH, Ludman E, et al. Understanding the Values of Patients with Multiple Chronic Conditions. Under review;


Improving the ‘Fitness for Purpose’ of Common Data Models through Realism Based Ontology

Jonathan C. Blaisure, BS, MS, PhD student1,2, Werner M. Ceusters, MD1,2
1Institute for Healthcare Informatics, University at Buffalo, Buffalo, New York, USA;
2Department of Biomedical Informatics, University at Buffalo, Buffalo, New York, USA;

Abstract
Common data models are designed and built based on requirements that are aimed towards fitness for purpose. But when common data models are used as lenses through which reality is observed from the perspective according to which they are built, then they exhibit restrictions that distort such view. Realism-based ontology design, when done properly, does not have these limitations as its fitness for purpose is only determined by the degree to which reality is represented the way it is. Therefore, we can use the principles that realism-based ontologies adhere to, not only to design application ontologies serving some specific purpose, but also to assess whether and where common data models fall short in their representational adequacy and how they can be corrected. If a realism based ontological perspective on the portion of reality the some common data model is trying to represent is compared with the perspective of the common data model itself, it is possible to determine how the latter deviates from the former and to suggest solutions to correct the misrepresentations found. Applying this method to the common data model of the Observational Medical Outcomes Partnership, revealed two major categories of errors: one where relationships are restricted based on the constraints of the data model, and one where the representation of reality is oversimplified.

Introduction
The University at Buffalo’s Institute for Healthcare Informatics’ (IHI) has as primary function the aggregation of disparate distinct data sources such as in-patient, out-patient, claims and clinical study datasets into a centralized integrated data repository (CIDR). It is a design criterion of the IHI to create this resource for healthcare data in maximal compliance with the principles of Ontological Realism1, 2 insofar doing so does not interfere with (1) the need for timely access and (2) the availability of resources. This is done in the spirit of not letting the perfect come in the way of the good. The IHI realizes this function by providing a secure environment where data analysis, cohort discovery and other secondary data use requirements can be evaluated and accomplished. Although the production of a CIDR is the primary principal mission of the IHI, the researchers at the University at Buffalo (UB) have a need for advanced analytics in the interim. One pathway forward is to adopt a Common Data Model (CDM).

Traditional CDMs are designed around a ‘fitness for purpose’ paradigm according to which the data are organized in a way that solves specific organizational necessities thereby allowing portability and integration or federation of other datasets. Some of these requirements can dictate the use of specific implementation designs. For example, Informatics for Integrating Biology and the Bedside (i2b2)3 proposes a CDM that is designed specifically for cohort discovery. i2b2 achieves this by using a single fact table to represent observations about patients. The fact table is dimensionally described by what are called ‘dimension tables’ – patient dimensions, provider dimensions, and encounter dimensions are examples. This data model implements a paradigm which is known as a star schema database. The i2b2 star schema database as applied to integrated data coming from, for example, electronic healthcare records (EHR) is highly efficient in identifying patient cohorts using inclusion and exclusion criteria in queries which run over observations in the fact table. Data coming from EHRs undergo a complex mapping process that makes use of standardized terminologies (or in-house grown terminologies) and which applies to each row that is loaded into the observation_fact table. In that way, the assertion in the EHR becomes transformed into what is called a ‘fact’ as perceived through the data model although, of course, what is the case in reality might be different from what is stated to be a ‘fact’. Users can browse these terminologies and use them to identify in the data collection patient cohorts composed out of patients about whom a specific observation which maps to the specific terms of interest is asserted (or not). The determination of ‘fitness for purpose’ of i2b2 is to provide a high-speed query system to recognize patient cohorts for (mostly) clinical trials. Although i2b2 solves this particular obstacle elegantly, it falls short, in our opinion, when ‘fitting’ it to other evaluation criteria specifically those imposed by realism based ontology (RBO).
generally considered necessary for active safety analysis. Further examples of CDMs include the Observational Medical Outcomes Partnership (OMOP), the Patient-Centered Outcomes Research Network (PCORnet), the healthcare management organizations’ research network (HOMOR) virtual data warehouse, and the Study Data Tabulation Model (SDTM) of the Clinical Data Interchange Standards Consortium (CDISC). Several of these CDMs have been subjected to studies for their ‘fitness for purpose’ for storing data extracted from electronic medical records (EMRs) specifically for the purpose of secondary data use in research. The OMOP pilot program, for example, terminated June 2013, but development still proceeds at the Observational Health Data Sciences and Informatics (OHDSI) collaborative. OHDSI has released version 5.0 of the CDM and has helped develop a software tooling chain to facilitate extract, transform and load (ETL) processes from a diversity of source systems. These tools are designed to aid mapping of data sources to terminologies, loading source data into CDM-compatible data repositories, and data analysis on these repositories. The intent of the CDM is thus to provide a ‘common model’ for data coming from all healthcare information systems to be transformed and loaded into CDM-compatible data stores for the purpose of research, analytics, and data integration. It is intended to do this with minimal transformations and data loss. We chose the OMOP CDM as our intermediary data model for (1) the way it can deal with what is generally called ‘findings,’ (2) the variety of open source analytical tools available, and (3) its wider purpose. It has also been qualified in some studies as the ‘least lossy approach’ among several CDMs tested.

The work described here is the result of scrutinizing the OMOP CDM from a RBO perspective specifically on the ways the data model presents a distorted view on the reality of the world it is referencing. The primary principle of comparison is restrictions based on the ‘fitness for purpose’ of the CDM and how those restrictions inhibit referencing reality adequately. Several publicly available ontologies, and, perhaps more importantly, the principles that they adhere to, were used as references including the Basic Formal Ontology version 2 (BFO2) 10, the Information Artifact Ontology (IAO) 11 and the Ontology for General Medical Sciences (OGMS) 12.

Methods

We followed the approach outlined in ontological realism 13 which takes very serious the distinction between data and data models on the one hand versus what the data and data models are about on the other hand. This allows us then to determine the differences between the CDM’s ‘fitness for purpose’ versus the ‘fitness to reality’. One can for example, metaphorically, view the data model as a container with a defined shape and size and built out of a certain material that restricts in certain ways what it can be filled with. In the case of the CDM, the restrictions are brought about by the structure of the tables, the cardinality of relationships, and the constraints implemented in the model. The ‘shape’ and ‘size’ of the container determine the qualities of that model – how well the model represents reality and the technical requirements or the ‘fitness for purpose’. The structure of a data model, specifically in this case, a relational data model, is a lens through which one can view certain portions of reality (PoR) – sometimes exactly the way they are, sometimes, as we will demonstrate, not without distortions – while others are shielded off. According to the RBO-perspective, PoRs are composed of types (PERSON, ROLE, PROCESS – types are standardly written in SMALL-CAPS, while particulars are written in italics) and particulars – instances of types that carry identity (the two authors of this paper are both particulars which instantiate the type PERSON; with respect to the work presented in this paper, they each had particular roles each one of which was an instance of ROLE, and so forth). Types relate to other types by virtue of the way all particulars of these distinct types relate to each other. Such relationships between types can be expressed by axioms in a variety of formal ways. For example, in the Basic Formal Ontology (BFO) 10 an axiom stating that an EXTENDED ORGANISM 14 Instance-Of EXTENDED ORGANISM then John Doe is also Instance-Of MATERIAL ENTITY. Also John Doe’s Height is a particular which carries identity and which Inheres-In John Doe. This John Doe’s Height is a particular quality and is Instance-Of the universal QUALITY. It is not an Instance-Of the universal EXTENDED ORGANISM.

The OMOP CDM’s ‘fitness for purpose’ is ‘to accommodate data from the observational medical databases that are generally considered necessary for active safety analysis’ thereby being ‘analyist-friendly’ to meet the requirement to ‘allow the analytic methods to execute quickly enough to be practical’ 15, p55. On the other hand, the RBO-perspective is purpose independent, with the exception of reflecting the structure of reality 13. Therefore, it is hypothesized that the OMOP CDM’s ‘fitness for purpose’ limits the OMOP-perspective to a reductionist representation of reality and thus that the fields used in the tables deviate from realist types. Or in other words: comparing the two perspectives might lead to a conclusion that the OMOP view is an oversimplification (reductionist view) of or an unfaithful (deviant) view to reality.
The PoR which is represented by means of the OMOP perspective can also be represented by means of an RBO view by using RBO compliant ontologies such as the Basic Formal Ontology (BFO)\(^{19}\), the Information Artifact Ontology (IAO)\(^{14}\), the Ontology for Biomedical Investigations (OBI), and the Ontology for General Medical Science (OGMS)\(^{14}\). These ontologies, and the principles that they adhere to, can be used to provide context in the comparison of the RBO-perspective vs the OMOP-perspective. By comparing these two perspectives we can qualify the accuracy of representation to reality and understand why and in what way certain design restrictions distort the representation of reality. The OMOP-perspective should primarily reference types and relationships amongst types so that the data which are stored in OMOP-compatible data repositories represent relationships between particulars in exactly the same way the RBO-perspective would reference particulars, relations between those particulars, and relations between these particulars and types.

From the RBO-perspective, relational data models, including those used in EMRs, practice management systems, and CDMs are composed of individual parts – tables, relationships, columns and so forth. Under an RBO-perspective, these components are INFORMATION CONTENT ENTITIES (ICE)\(^{17}\). A particular ICE Is-About some entity in reality. For example, a patient medical record number (MRN) is an ICE which Is-About some person with a patient role that Inheres-In that person. A PatientID column in a relational database based on the OMOP CDM represents the type PERSON whereby each particular cell of that column represents an Instance-Of PERSON. On the other hand, a diagnosis is an ICE that references some output of a clinical diagnosis process that Is-About a disease, while a disease Inheres-In some person.\(^{14, 18}\) This may seem trivial – or perhaps overly complicated – to a clinician at the point of care, but the distinction is important to accurately represent reality: the diagnosis and the disease are separate and distinct entities but are often represented in CDMs as the same entity.

With this in mind, we explored whether there are design principles that are standardly used in information modeling approaches that may have had a negative impact on the implementation and development of the CDM. Our process took into consideration the ‘fitness for purpose’ as well as various projects’ conformity to the CDM as described in the literature. Looking with the eye of a realist ontologist to the ‘fitness for purpose’, requirements and design goals of the OMOP CDM may provide insight to where the model references reality objectively and adequately, and where it falls short. For example, a data model that has a requirement to conform to a certain structure to allow business intelligence (BI) tools to be able to ingest and query the data can be a limiting design restriction in itself. BI tools are designed in certain ways for maximum efficiency and to answer queries of a specific kind. A component-based\(^{19}\) ecosystem has been developed around specific versions of the OMOP CDM and the desire to take maximally advantage of the BI tools it brings with it provides an incentive to not stray away from the original design guidelines. The OMOP CDM has had noteworthy improvements from release to release but drastic changes to the CDM would cause decisions to be made about either, the immediate redesign of software – to bring it up to date with the changes in the current version of the CDM – or the acceptance of being out of version compliance. In fact, one of the design principles of the OHDSI consortium is ‘Backwards Compatibility’\(^{20}\). Additional issues to contemplate when designing a relational data model are the normalization of data – data duplication, restricting the number of joins required to traverse the data. Joins are expensive, and building constraints that provide data consistency but does not constrain the model in a way that clashes with the ‘fitness for purpose’ is a challenging problem.

We began this comparison by downloading the data definition language (DDL) for the OMOP CDM version (v5) from the GitHub repository maintained by the Observational Health Data Sciences and Informatics (OHDSI) collaborative\(^{21}\). The DDL scripts are instructions on how to create the tables, relationships, indices and constraints that makeup the OMOP CDM and were loaded into a locally installed PostgreSQL server. After these scripts finished running, we ran additional scripts on the database for the purpose of creating documentation such as an entity relationship diagram (ERD) that visually represents a relational database schema. ERDs allow the visual inspection of tables, their types, and the relationships to other tables (constraints) by means of tables and field descriptions, cardinality of relationships, etc. Afterwards, we downloaded the documentation supplied by the CDM v5\(^{22}\) and used it to compare the ERD with the purpose to derive the informal semantics of the OMOP CDM, specifically the lens representing OMOP CDMs PoR used to view entities and their relationships – the ‘OMOP-perspective’. We then queried PubMed to identify information pertaining to source data conversion into the OMOP CDM to proactively avoid problems reported in the literature. We performed an analysis to decisively compare common downsides and gained knowledge from other organizations experiences and perspectives to improve our approach.
Results and discussion

We have identified thus far two different ways in which the OMOP CDM design goals conflict with an RBO-perspective on the data represented in OMOP CDM-compatible data stores (Table 1). One design principle of the CDM is reported as follows: ‘The CDM aims at providing data organized in a way optimal for analysis, rather than for the purpose of operational needs of health care providers or payers’\(^\text{20}\). Operational needs and data analysis needs differ indeed. Data normalization in operational systems is for instance focused on making transaction speeds satisfy the requirements of the operational environment, which in the case of EMRs comes down to the ability of quickly entering and retrieving data about a single patient. This constitutes a relatively small amount of data with respect to the totality of data held over all patients in the EMR system so that the search space is quite small as well. Analyst queries on the other hand have to run over very large amounts of data while also returning large result sets. Although speed is an issue, it is not as severe as in EMR systems. For example, waiting hours for an analytics question to be answered using the data in a secondary use data store is for sure annoying, but does not need to disrupt the workflow of the data analyst. But it would be unacceptable to have the query run for days and weeks, what would be the case if a typical analytics question would be run over the back-end of the EMR system itself, rather than over the secondary use data store of which the model is optimized to handle these kinds of queries. By examining the CDM, specifically the Person-table, Observation-table, Provider-table and Location-table we start to see where some conflicts with the RBO-perspective as brought about by this sort of optimization arise from.

Table 1 – Identified problems.

<table>
<thead>
<tr>
<th>Type of Problem</th>
<th>Description</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cardinality</td>
<td>A problem where relationships are (incorrectly)</td>
<td>A person’s address can change over time.</td>
</tr>
<tr>
<td></td>
<td>restricted based on the constraints of the data</td>
<td></td>
</tr>
<tr>
<td></td>
<td>model.</td>
<td></td>
</tr>
<tr>
<td>Reductionist</td>
<td>A problem where the representation of reality is</td>
<td>A provider is a role that a person bears.</td>
</tr>
<tr>
<td></td>
<td>over simplified.</td>
<td></td>
</tr>
</tbody>
</table>

The Person-table is composed of many fields which hold data with specific data types, as for instance in the person_id-column. The OMOP specification documents this field as being ‘A unique identifier for each person’\(^\text{23}\). From the RBO-perspective we would phrase this that in relation to what occupies a specific cell in the person_id-column there exist a particular person_id which is an Instance-Of PERSON_ID, this type being itself a subtype of ICE. That particular person_id Is-About (at some point in time) a unique particular that is an Instance-Of PERSON. What is stored in the person_id-column of an OMOP-compatible data store is then a Concretization-Of that particular person_id. That concretization has ‘in the database on disk’ probably the form of a specific pattern of magnetization points. When that person_id is concretized on paper, it is most likely in the form of a (alpha-)numerical string. The paper can be destroyed separately from the database, or even both can be destroyed. But that would not result in the person_id to be destroyed, and for sure not the person about which this is the ID. There is in the RBO-perspective no ‘death through nullification’ for which the now abandoned HL7 RIM was once critiqued\(^\text{24}\).

Figure 1. RBO-perspective Person vs Provider
On the other hand, examination of the Provider-table documentation reveals that the column provider_id-column is defined as ‘A unique identifier for each Provider’ 25. Under the RBO-perspective, a PERSON is a subtype of MATERIAL ENTITY which itself is a subtype of INDEPENDENT CONTINUANT while a PROVIDER is a subtype of ROLE which itself is a subtype of REALIZABLE ENTITY that Inheres-In a PERSON. PROVIDER is trivially not the same type as PERSON nor is it subsumed by PERSON. Figure 1 illustrates these relationships from the RBO-perspective (rectangles with rounded corners represent instances while rectangles with square corners represent types). The question now is what exactly is meant in the OMOP CDM with ‘provider’: should this be interpreted as meaning ‘provider role’ or ‘person which has a provider role’? And it should make us wonder whether with ‘person’, OMOP really means what we typically understand under that term, or whether they mean ‘patient’. Under the interpretation that the provider table is a lens that captures references to instances of PROVIDER ROLE but not to the particular person that bears the PROVIDER ROLE, one would expect the persons that bears that role to be represented in the person table as well. Although a valid distinction, the goal of this comparison is to compare the implementation of the CDM to its ‘fitness for purpose’. Arguably a simple data analysis question: ‘How many unique entities of type PERSON are referenced in the CDM?’ would return too small a number if PERSONS who bear a PROVIDER ROLE are not included in the person table and only entries in the person table would be counted. On the other hand, if the same question would be answered by a query that returns the total number of both entries in the person table and entries in the provider table, then this would result in a number that is too high if there are persons represented in the dataset that have both a PATIENT ROLE and a PROVIDER ROLE. This is deviant of the RBO-perspective but also deviates from the OMOP-perspective’s ‘fitness for purpose’ with that purpose being accurate analysis of data. Accuracy in this case is thus, as we assume, only expected for certain types of questions that according to the designers are from their perspective relevant to be asked, and not for all types of questions that can be asked over exactly the same domain: thus the two questions asked, although valid from a realist perspective, are not supposed to be asked to OMOP CDM-compatible data repositories.

The root cause of the problem just sketched can be described as a confusion of types. A possible solution to this mis-representation of reality used in the above example may be addressed by expanding the CDM to include a Person-table, Provider-Role-table and a Patient-Role-table. These tables would distinctly identify the realization of patient and provider roles throughout the dataset and allow unique and accurate counting of instances of type PERSON. Constraining each role by a start and end timestamp should be used to temporally qualify roles linearly – allowing deeper analysis and thus increasing the CDMs ‘fitness’ (figure 2).

Another problem discovered is the opposite of the previous one. While the previous problem had to do with incorrect counting of particulars due to incorrect representation of uniqueness and confusion of types, this one has to do with cardinality constraints existing in the CDM. We have examined the Location-table and the Person-table to compare it with the RBO-perspective. As displayed in figure 3, the Person-table has a column named location_id which is defined as ‘A foreign key to the place of residency for the person in the location table, where the detailed address
The model’s definition (DDL) is designed in a way that will only allow one location to be assigned to a person at any given time. This is so because of the structure of the data model which is designed in such a way that a row in the person table can only have one reference to a row in the location table (Figure 3). This type of one-to-one relationships form the least costly joins that can be made in a relational database.

![Figure 3 – Relationship between Person and Location table](image)

From the RBO-perspective, relationships between particulars – the main relata – of which one is of type CONTINUANT always involve an extra particular which is an instance of TEMPORAL REGION. Assertions about such relationships should thus contain references to all three particulars: to both main relata and to the temporal region during which the relationship between the two particulars obtains. Entries in databases in which foreign keys are used as described above to indicate the address of a person, both persons and addresses being continuants, qualify as assertions about relationships that hold over a period of time. However, in the OMOP CDM, only two of the three required particulars are referenced: there is a reference to, say, Person 1 and to the residency of Person 1 (in this case an address – Address A) but there is no reference to the temporal region. Of course, at some point in time, the location_id may need to reference another address when Person 1 moves to another city or even across town – Address B at some other time. A secondary data use question such as “How many instances of type PERSON could have been exposed to pollution from this water source?” would only include whichever single location is associated with that particular PERSON and not consider all the places the person lived in the past. The OMOP CDM documentation does acknowledge that patients over time can have distinct locations, genders, etc., but “it is the responsibility of the data holder to select the one value to use in the CDM”\(^{22}\). A possible solution to correct the representation of reality in this case could be to change the cardinality of the Person-table to Location-table to allow multiple relationships between a person and a location. This can be accomplished by using a bridging table. In the example below we have added a start_dttm and end_dttm (Figure 4) column to the location-table so we can provide a TEMPORAL REGION reference. With this addition, a person’s locations can be tracked over time representing reality more accurately, and providing a stronger use case for secondary data use. The current CDM structure does not limit introducing a new location into the database, but it does not allow creating a historical transaction. If a person changes residency, the location of that residency would change removing the previous location and no record of this transaction would exist.

This is a reductionist problem involving temporality. Relationships amongst particulars one of which is a continuant obtain in some temporal region and it is important for the model to be able to capture this feature of reality. We can analyze other fields using this same logic such as the Gender-column which shows the exact same problem – gender can change over time and this reality is important for secondary data use and accurately representing reality.
Conclusion

It is our belief that we can compare data model designs to designs based on ontological realism to not only show their restrictions but improve their ‘fitness for purpose’. Through an RBO-perspective we have identified thus far two examples of misrepresentation of reality – the confusion of types and cardinality problems related to temporal regions around particulars. Some may argue against the importance of these misrepresentations but as realism based ontologists we argue that representing data accurately to reality directly affects real world secondary data use requirements. By applying realism based ontology, we can ultimately increase the ‘fitness for purpose’ of data models and their respective requirements. As such, a practical implementation of RBO-perspectives can be incorporated into the design of data models improving their accuracy to represent reality. Many aspects of the RBO field have been theoretical in the past and thus, have not been applied in a practical way to existing CDMs. We believe that the maturity of the work that is being done in the RBO fields, specifically in the biomedical domains, and the advances in database technologies, presents a unique opportunity to develop the next generation of data models.

Acknowledgement: This work was supported in part by Clinical and Translational Science Award NIH 1 UL1 TR001412-01 from the National Institutes of Health.

References

A Simulation Study on Handoffs and Cross-coverage:
Results of an Error Analysis

Katherine Blondon, MD, PhD1,2, Marzia Del Zotto, PhD1, Jessica Rochat,
Mathieu R. Nendaz, MD, MHPE1,2, Christian Lovis, MD, MPH1,2
1University Hospitals of Geneva, Geneva, Switzerland; 2University of Geneva, Geneva, Switzerland

Abstract

Handoffs and cross-coverage are necessary for maintaining the continuity of patient care, yet both are potential sources of error, and may threaten patient safety and care. Handoffs are the transfer of patient information and accountability from one provider to another. Cross-coverage is the management of patients, of whom physicians have little or no prior knowledge of, during nightshifts. We observed how physicians give a sign-out after receiving a handoff in a simulated session of an evening handoff and start of nightshift. We collected data from thirty physicians from an academic medical center as they signed out six patients after responding to nurse calls. An error analysis of the sign-out data revealed 42 errors overall, with 28 omissions and 14 “erroneous data” errors. We then propose ways to prevent these errors through modification of the electronic medical record and support tools, and through higher awareness of human factors.

Introduction

Failures in communication between health care professionals and diagnostic reasoning errors pose a potential threat to patient safety and quality of care. Root cause analyses of sentinel events have shown an implication of communication issues in up to 70% of cases in a report by the Joint Commission on Accreditation of Healthcare Organizations1. Communication is particularly important during moments of transition of care, or handoffs, which are necessary to maintain the continuity of care, particularly among hospitalized patients. Handoffs are defined as the transfer of patient information and accountability from one healthcare provider to another2. Handoffs take place between day and night teams, or between wards or departments (e.g., between the emergency department and the internal medicine ward). The handoff practice varies widely between departments and institutions, and refers to in-person, verbal handoffs, phone calls, written handoffs or a combination of these. Recommendations for improving handoffs have emphasized the need to standardize the process2, which has been successfully reported with the “I-PASS handoff bundle”3. Recent studies on handoffs have shown the benefits of improving the handoff process on medical errors4, 5.

Cross-coverage is the process used during nightshifts, when the night team pursues the management of patients in the wards6. Night physicians face additional challenges for patient care, as (1) they manage a larger number of patients, (2) they are unfamiliar with the majority of patients, and therefore need to depend on prior assessments and plans performed by the day colleagues, and (3) work with less staffed teams (both nursing and physician teams), with (4) subsequent lower supervision during the night than during the day. Furthermore, night physicians may also manage patients for whom they have not received a handoff. Therefore they need to be able to rapidly extract pertinent information from the medical records, and other available resources (nurses, or patient summaries). This rapid extraction of pertinent content from the medical records has been called “chart biopsy.”7

Paper summaries are an interesting support tool for handoffs, because it allows the receiver to have an overview of the patients’ problems. Systems like UW Cores (University of Washington computerized rounding and sign-out system)8 have shown their effectiveness in improving the workflow process and handoff. During focus groups from a prior study with residents and supervising physicians from our division9, discussions around support tools for handoffs raised concerns about errors and redundancy. Paper-based tools, even if auto-populated, are only an instantaneous snapshot of a patient’s medical chart, and the physicians therefore were concerned that they may be a source of error, if modifications appeared after printing the summary. Furthermore, since logging into the chart is required to see the latest vital signs or for prescriptions, most participants did not see an immediate benefit in having a potentially unreliable paper print-out, even if it could perhaps allow for a more ubiquitous or quicker access to patient information. Unsurprisingly, many physicians actually use hand-written notes (on blank paper or on documents such as the admission notes) as support tools throughout the day, which may have an even higher rate of imprecision or error.
Prior studies on handoffs or cross-coverage have for the large part been observational studies or retrospective cohort studies. These designs have limitations, such as lack of comparability and patient specificity, or recall bias. We therefore conducted a simulated handoff and chart biopsy study, an approach that allowed us to use standardized cases with pre-defined nursing cues, and direct observations of patient management and electronic health record (EHR) use.

The aim of the present study was to study the handoff process among residents and supervising physicians, focusing on the potential errors, to identify their origins. We hypothesized that errors during the handoff simulations could arise from the handoff process, the data collection during the chart biopsy, or the clinical reasoning during cross-coverage. We then explore ways to help avoid these errors through EHR design and medical training. The analysis of the handoff quality in terms of the type of support (paper summary, EHR or none) will be presented in a separate report.

**Handoffs in General Internal Medicine at our institution**

At the University Hospitals of Geneva, the medical teams (2 residents, 2 medical students, and 1 supervising physician) of the Division of General Internal Medicine each manage about 15 to 20 patients. The teams are divided into two zones of about 100 patients, and one nightshift physician cross-covers each zone. A senior physician from the Emergency Department, who is not otherwise involved in the patients’ care, supervises nightshift residents. Therefore, night and day teams are completely different, and rely on handoffs, notes in the medical charts, and interactions with the nursing teams, to maintain the continuity of patient care.

Handoffs take place in a meeting room away from the wards, and last often less than 15 min for the whole zone. This is achieved by only handing off a selection of patients to the night team, selected according to the following criteria: (1) Patients whose condition has recently deteriorated (unstable), or who have recently developed new symptoms of yet uncertain diagnosis, and who therefore require monitoring, (2) patients who require pending action (e.g., following up on lab or imaging results, or pending consultations by a specialist), (3) patients who are considered complex, or whose management is unusual, (4) patients who are receiving end of life care, or (5) patients for whom the day team anticipates potential complications overnight. Simple, stable patients with no expected complications are not handed off to the night team. A locally developed mnemonic in French (Code IDEALE for CODE status, Identity, main Diagnosis, Evolution, expectAtion, to-do List and quEstions) is sometimes used to introduce handoffs to students or new residents.

**Methods**

*Participants:* After approval from the local ethics committee, we recruited volunteer residents (R) and supervising physicians (SP) in General Internal Medicine from our academic hospital between January and November 2015 through convenience sampling.

*Simulation:* Our study design reproduced a Friday evening handoff session in our Internal Medicine wards, and beginning of evening shift for the participant. The simulation had three phases: a handoff phase, a chart biopsy phase, and a sign-out phase. For clarity in this paper, we used the term “handoff” when the participant was the receiver of information, and the term “sign-out” when the participant was the information giver. The overall duration of the simulation was about 45 min.

- In the **handoff** phase, one investigator (KB) gave a standardized handoff of four randomly chosen patient cases, allowing time for questions and interactions with the participant. Participants were randomly selected to receive a paper summary tool, have access to the EHR or to be the control group, with only blank paper for note taking (available in all 3 groups).
- In the **chart biopsy** phase, the participant was left alone to review the EHR for the beginning of a nightshift. During this phase, the participant received four calls from nurses, two of which were about patients that had been handed off, and two for patients without handoffs. We then informed the participant that her shift needed to be interrupted, and asked her to sign-out to a colleague. This phase lasted about 25 min, or whenever the participant felt prepared to sign-out, to avoid time pressure bias.
- In the **sign-out** phase, participants chose which patients they wanted to discuss, and presented them to an investigator (KB) who had no prior knowledge of the patients.

*Data collection:* We collected demographic data on the participants, such as level of experience and years of practice. We audio-recorded and transcribed the simulations. For this paper, we focused on the sign-outs, which were de-identified for subsequent analysis.
Scenario preparation: We created 8 fictitious clinical cases of commonly encountered diagnoses in internal medicine wards, inspired from real patient cases. A brief summary of the cases is provided in Table 1. For each scenario, we prepared a mock-up prototype based on our local electronic medical record (EHR), a nursing call, and a paper summary. We chose to use mock-ups of the EHR rather than real patient charts for several reasons. First, we wanted to be able to control all the information that participants had access to for each patient. Second, the mock-up helped avoid an additional cognitive load for participants, by adjusting the chart dates for the study. In a prior study with a similar design \(^{10}\), participants found the date issue confusing, especially since each patient had different hospitalization dates. Third, we did not want any post-test date information to be visible in the patient chart that could affect EHR use, or clinical reasoning (ie., "patient deceased", or test results from a later date). Although using a mock-up restricted the EHR functionalities, (i.e., search engines in the EHR were not functional, CT scans were limited to a few screen shots), the prototype allowed all the main interactions with the EHR, and all the dates in the EHR were adjusted to the dates of the participants’ sessions to decrease confusion and to optimize the realism of the simulation. Seven of the cases were used in the handoff randomization, whereas the medication error situation (Case E) was used solely in the randomization of nurse call, and thus was never handed off. Cases are listed in Table 1.

Table 1. Overview of patient cases and content of nurse calls

<table>
<thead>
<tr>
<th>Case</th>
<th>Main diagnoses</th>
<th>Nurse call</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Inaugural decompensated heart failure due to ischemic heart disease. Diabetesthe with end-organ complications, perforating ulcer</td>
<td>Retrosternal chest pain</td>
</tr>
<tr>
<td>B</td>
<td>Liver failure in a patient with Child C cirrhosis. Presented hemoptysis at admission due to anticoagulation for recent pulmonary embolism</td>
<td>Abdominal pain</td>
</tr>
<tr>
<td>C</td>
<td>Pneumonia in a patient with metastatic prostate cancer, lung atelectasis of undetermined origin and sleep apnea.</td>
<td>Chest pain</td>
</tr>
<tr>
<td>D</td>
<td>Acute renal failure and electrolyte imbalance in a patient with an ileostomy, and severe chronic obstructive pulmonary disease</td>
<td>Shortness of breath</td>
</tr>
<tr>
<td>E*</td>
<td>Uroepisis in a hypertensive patient who tends to fall, currently has a broken right hand</td>
<td>Medication error</td>
</tr>
<tr>
<td>F</td>
<td>Fortuitous discovery of severe hyponatremia, patient with suspected cirrhosis</td>
<td>Fever</td>
</tr>
<tr>
<td>G</td>
<td>Pneumonia with empyema in a young smoker, who presents pain due to his chest tube</td>
<td>Skin rash</td>
</tr>
<tr>
<td>H</td>
<td>Decompensated Child C cirrhosis due to G-I bleed and sepsis in a patient with chronic alcoolism, liver failure and encephalitis, recently transferred from the ICU. Currently still delirious. Awaiting paracentesis results for recurrent fever.</td>
<td>Low blood pressure</td>
</tr>
</tbody>
</table>

*This case was not handed off during the study, and was solely used for nurse calls

Some explanations about the cases help understand the encountered difficulties. In Case A, the participant received a call about acute chest pain, in a patient admitted for an inaugural episode of heart failure from ischemic heart disease. Patient management was quite simple in this case, but needed to be conducted urgently. In this scenario, the main difficulty resided in the retrieval of information about all the cardiology work-up the patient had received since his admission.

Case B was about a patient admitted ten days earlier for liver failure, who presented bleeding (hemoptysis) shortly after being admitted. His anticoagulant therapy was interrupted for a few days, despite the recent (<3 months ago) diagnosis of pulmonary embolism, as discussed with the vascular medicine consultant. The plan was to begin the anticoagulation therapy shortly, since the bleeding had stopped.

Case C was a patient with a history of metastatic prostate cancer who was admitted for pneumonia. The nurse calls about recurrent chest pain, which is of non-cardiac origin, with a range of differential diagnoses.

Case D was a patient with past medical history list that was not updated in the patient summary (it did not include chronic obstructive pulmonary disorder), while the results of the pulmonary function test were reported in the updated problem list, and showed sever lung disease. This discordance was noticed rapidly, and led to questions during the initial handoff, or later verification in the EHR.
Case E was a patient admitted for sepsis, with no particular concerns. She unfortunately takes her neighbor’s medication, leading to the ingestion of a high dosage of beta-blocker and an oral diabetes medication. The erroneous information for her was the indication about the ward she was in.

The patient in Case F was admitted for hyponatremia, and presents with new nocturnal agitation, with a wide range of possible diagnoses. Case G patient was a patient with complicated pneumonia, who had an antibiotic switch the day of the handoff, with the paper summary that had not been updated. All but one participant identified this error during the initial handoff. And finally, case H was a complex scenario of a patient with liver failure and many complications, who presented recurrent fever. His hospital stay was the longest, since he’d already been in the intensive care and intermediate care units.

Paper summaries were a novel feature for our participants, as they are not available in the current EHR system. We designed the summaries as potentially auto-populated fields from the EHR. We included patient identity and age, room and bed, code status, comorbidities, problems and to-do list. We also included hospital duration, which is not easily visible in the EHR. These paper summary features were discussed and validated in focus groups with Rs and SPs in a prior study on handoffs in our institution. The paper summaries were tested on pilot participants for this study, and were subsequently shortened. Dates of the summaries were also adjusted to the participants’ sessions.

To address concerns raised about the reliability of paper summaries, we included imprecisions and even discordant information in the paper summaries. For example, we reported a patient on one antibiotic, which was actually switched the day of the simulation sessions. This type of error is commonly encountered as a missed update, or an inappropriate copy-paste. In another case, we failed to update the past medical history list (probable COPD, no pulmonary function testing) after working-up a patient, which revealed severe COPD.

The verbal handoffs were also standardized for each case, with answers for anticipated questions from the participant. Other questions were answered with vague responses, which were meant to lead the participant to look up the information in the EHR. Although the paper summaries contained imprecisions or errors, all verbal handoffs contained the correct information.

Analysis: In this study, we studied all the elements given in the sign-out phase, and focused particularly on errors. Errors were defined as omissions, or erroneous information that was given during the sign-out. We also analyzed the imprecisions in the sign-out phase. The definitions of these terms for our study are presented in Table 2. An expert physician conducted a thematic analysis, and coded the transcripts using Atlas.ti v1 for Mac, taking into consideration the handoff status and nursing call interventions. A second expert physician cross-checked the initial coding. Differences were discussed to reach a consensus.

Table 2. Definitions

<table>
<thead>
<tr>
<th>Terms</th>
<th>Definitions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Error</td>
<td>Missing information, such as fever that was not reported, with a clinical impact</td>
</tr>
<tr>
<td>Omission</td>
<td>Facts that were mentioned in the sign-out, but were not provided in any of the data sources (handoff, nurses, paper summaries or EHR).</td>
</tr>
<tr>
<td>Imprecison</td>
<td>Facts that were incomplete or approximate, but with little or no impact on the comprehension or management of the case. For example, an age reported as 80-something rather than 83 years would be an imprecision.</td>
</tr>
</tbody>
</table>

We report the sign-out durations by level of expertise. We use descriptive statistics to present the errors and imprecisions. We compared the mean number of errors in cases between the 3 arms with crude and adjusted (for experience level) linear regression models with robust standard errors. We then analyzed whether the sources of errors were in relation to the EHR data, paper summaries, annotations, or initial verbal handoff. Finally, we studied how to prevent these errors, through content and design of the EHR or paper summaries, or through human factors.
Results

Participants
We recruited 30 participants, 16 residents and 14 supervising physicians (Table 3) with an average of 5.4 years of experience. Overall, SPs had 6.5 years more clinical experience than the Rs. Only two of the participants reported having received handoff training. Six of the interns and seven of the supervising physicians had heard of a locally developed handoff mnemonic. Table 3 shows that the randomization resulted in a slightly larger paper summary group, who also had more supervising clinicians.

Participants were all familiar with the local EHR system.

Sign-out description
The sign-out phase duration was 12.2 min on average (range 6.0 – 30.0), with a non-significant difference between SPs (11.3 min) and Rs (13.0 min), with p=.33. One SP participant, P12, chose to present only 3 patients during sign-out, and therefore had the shortest duration. All other participants presented the six patients.

A junior SP participant, P17, was the other extreme of the sign-out duration. He was particularly thorough during the sign-out process, and took time to look up information in the EHR to complete his sign-out. He took nearly five times the amount of time of the shortest sign-out for six patients (30 min vs 6.5 min), despite some pressure from an investigator, and had 2 errors and 5 imprecisions. P17’s explanation for not being able to be more concise was the complexity of the cases.

Errors and imprecisions
According to our definitions of errors, we found 42 errors overall among the 180 sign-outs, or 23.3 per 100 cases. Twenty-eight of the errors were omissions, 14 were erroneous facts. There were 100 imprecisions in the 180 sign-outs. Eight discordances between coders (all for imprecisions) were discussed to reach a consensus. We present the errors and imprecisions per case in Table 4.

Table 3. Study participant characteristics.

<table>
<thead>
<tr>
<th></th>
<th>EHR</th>
<th>Paper</th>
<th>Control</th>
</tr>
</thead>
<tbody>
<tr>
<td>n (residents)</td>
<td>9 (6)</td>
<td>11 (4)</td>
<td>10 (6)</td>
</tr>
<tr>
<td>Female, n</td>
<td>4</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>Hand-off training, n</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Median years experience (range)</td>
<td>4.0 (0.2-10)</td>
<td>8.0 (0.2-20)</td>
<td>4.0 (0.6-15)</td>
</tr>
<tr>
<td>Nightshift experience (%)</td>
<td>89</td>
<td>91</td>
<td>80</td>
</tr>
</tbody>
</table>

Table 4. Errors and imprecisions by case

<table>
<thead>
<tr>
<th>Case A</th>
<th>Case B</th>
<th>Case C</th>
<th>Case D</th>
<th>Case E*</th>
<th>Case F</th>
<th>Case G</th>
<th>Case H</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Omissions</strong></td>
<td><strong>Errorenous facts</strong></td>
<td><strong>Total</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>4</td>
<td>2</td>
<td>0</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td>6</td>
<td>28</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>5</td>
<td>3</td>
<td>3</td>
<td>14</td>
</tr>
<tr>
<td>11</td>
<td>7</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>5</td>
<td>3</td>
<td>9</td>
<td>42</td>
</tr>
<tr>
<td><strong>Imprecisions</strong></td>
<td><strong>Total</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>17</td>
<td>14</td>
<td>6</td>
<td>7</td>
<td>6</td>
<td>16</td>
<td>22</td>
<td>100</td>
</tr>
</tbody>
</table>

*Case E was solely used for nurse calls.

Cases A and H had more errors than the other cases, even though cases B and G also had more imprecisions. This can be visualized more easily in Figure 1.

Figure 1. Errors and imprecisions by case
As the number of participants of the three types of support tools differed among the groups, we present error and imprecision rates per case in Table 5. The 0.23 mean level of total errors per case did not differ between the 3 arms, in the crude (p=0.98) and adjusted models (p=0.99).

Overall, the small discordances or errors in the paper summaries were identified by the participants, and were not the cause of errors in the sign-out.

| Table 5. Errors and imprecision rate per case for each type of support |
|-----------------|----------------|----------|-----|
| Sources of errors | EHR access | Paper summary | Control | Total |
| Omissions | 0.13 | 0.17 | 0.17 | 0.16 |
| Erroneous facts | 0.09 | 0.06 | 0.08 | 0.08 |
| Total errors | 0.22 | 0.23 | 0.25 | 0.23 |
| Imprecisions | 0.43 | 0.56 | 0.67 | 0.56 |

Sources of errors

A closer study of the errors and their sources reveals that errors arose at different times of the handoff process. During the verbal handoff, physicians may receive erroneous information from their colleague, or they may receive correct information, but misunderstand the facts, or simply record them imprecisely. One physician had trouble giving his sign-out because he had trouble re-reading his own notes, for example. Although we standardized the handoff content, participants were given the opportunity to ask questions about each patient before moving to the next patient. These interactions were important to clarify information, or to address issues that the received anticipated for the night. The questions asked ranged from medication specifications (“Which antibiotic is he on?”) to lab test results, or contingency plans. For example, participants were asked to follow-up on a lab test for ascites in Case H, and many participants asked if the day team had discussed which antibiotic to prescribe if the liquid showed signs of peritonitis.

The analysis of the erroneous facts in our dataset revealed areas of weakness in the handoff and chart biopsy process. Four of the erroneous facts were related to comprehension of the timeline of events. For example, P21 mistook the bleeding complication in Case B that occurred at admission one week prior as an event of the day, and therefore found the work-up to be incomplete, and made erroneous recommendations. In this example, the error was initiated during the verbal handoff, despite P21 having received paper summaries. Another participant mixed the data from two different patients during the sign-out, not realizing immediately that the EHR was open on another patient’s chart. The other erroneous facts were in the recommendations given, due to clinical reasoning or medical knowledge.

In the analysis of the omissions, we found that these occurred at different parts of the handoff process. Some participants forgot to report nurse calls, others did not document key points such as vital signs when responding to nurse calls. And yet other omissions related to the retrieval of various reports in the EHR, when this information had not been integrated in the progress notes.

Sign-out structure

Overall, participants tended to present patients by ward during the sign-out, rather by degree of urgency or order of initial handoff. Some physicians were a little disorganized, and sometimes only remembered some patient information after having moved to the next patient, such as in the case of P1: near the end of the sign-out, P1 suddenly came back to the first patient, asking whether he’d reported that this patient had also presented fever, and that a septic work-up needed to be done.

Discussion

Handoffs and cross-coverage have both been identified as moments of potential threat for quality and safety of patient care. Observational studies have studied incident reports, rapid team responses, or medical errors in the patient charts. Our error rate of 23.3 per 100 cases is in the range of other studies on handoffs and errors. As a comparison, Starmer et al reported an overall rate of medical errors of 33.8 per 100 admissions, and 24.5 per 100 admissions in their studies on handoffs and medical errors.

Our statistical analysis of the errors did not show any differences among any support tools or no support. This could be due to the small number of errors in the study, since our study was not powered for this outcome. Also,
the sources of error are not all related to information retrieval or interpretation. Erroneous facts due to clinical reasoning or medical knowledge deficits may not necessarily be improved by changing the type of support tool.

Interpretation of our error and imprecision rates need to be performed with caution. Our analyses examine differences in facts from our data sources, but not all facts are relevant in a handoff. In fact, physicians need to select the pertinent features in a patient case to remain concise. Omissions therefore may not all be bad. The difficulty lies in the selection process of the pertinent features to disclose.

Furthermore, imprecisions are probably related to the high cognitive load of the handoff process. Beyond the case complexity in terms of clinical reasoning, the EHR and available resources can also lead to difficulties. As we found in the results, one of the cases required finding recent reports of the cardiology work-up (Case A). It is this step that led to errors for our participants, because pertinent documents were mixed with all the other documents. Although the search function of our EHR mock-up was limited, compared to the EHR the participants were used to, we note that the participants did not take the extra time to find the documents. When the results were not easily retrievable from the progress notes, for example, some participants presumed that the results were not yet available, and abandoned their search.

**Human factors and training**

Human factors play an important role in the handoff process. Although higher awareness of the challenges of the handoff process could be addressed through training, the participants of our study seemed to reflect the current state of handoff training. Only two participants (SPs) had received any handoff training prior to this study. Most participants had heard of a locally developed mnemonic for handoffs in our department. Although none of the participants actually used this mnemonic, they found it useful to illustrate or to teach handoffs to students or peers.

Our findings show that errors generated early in the handoff process can persist throughout the sign-out, simply because there is little verification of the information received. Communication failures are very common in clinical care, present in up to two thirds of preventable events. Therefore verification of verbal information is essential, and is emphasized in successful handoff training programs like I-PASS. Moreover, training in handoffs should also draw awareness on the reliability of note-taking: handwriting styles, speed, and comprehension can all affect the quality of the notes. This reliance on notes and paper support tools led one participant to state contradictory facts about a patient, reading from the past medical history list, and then moving on the problem list, without realizing this. The other participants reported the updated status from the problem list only.

Implications for training based on the prior example would be to make sure participants use their clinical reasoning skills to choose the pertinent elements when signing out. Some participants briefly explained their reasoning process when giving recommendations. Having a shared understanding, or at least having the opportunity to clarify certain points of the handoff is commonly included in handoff recommendations. Another important manner to prevent errors from the onset would be for the receiver to summarize key elements for each patient, to ensure correct comprehension. This is the last component of the I-PASS handoff bundle.

**Support tools functionalities**

The low verification of facts by the users emphasizes the need for any handoff support tool to be able to present the most recent information available in the EHR, and to allow manually triggered updates. Although synchronizing support tool data with the EHR helps avoid transcription errors, it could also potentially be a source of errors, if they override previous information, lack traceability or provide information that is not yet processed by the physician in charge. For example, if the support tool presents a management plan based on an initial set of lab results, lack of a notification for new results may engender a misunderstanding about the reasoning underlying the plan. Also, there should be little need for repetitive documentation that could favor copy-pasting, such as for problem lists, or any other kinds of lists.

Having a patient’s information summarized in one place helps maintain a structured handoff. The structure of the handoff is important to share mental models of the patients, and to avoid having to jump back to previous patients to provide additional information. While most participants were able to maintain a patient-by-patient sign-out, some got confused with their annotation methods, particularly with the occurrence of new events. Having a designated place to write additional information, such as nursing calls, can help physicians stay organized. Although paper support tools are easier and often more efficient to annotate, the space to write is limited and is not standardized, making it more difficult for the annotations to be shared with the oncoming physician.
Design implications

In this section, we make some suggestions to help avoid errors during the handoff and chart biopsy process. Based on the errors identified in our analysis, we found several important measures to consider improving in the design and content of the EHR and the connected support tools. Our study design with standardized handoffs and repeated simulations with the same cases helped us identify and understand some of the potential sources of errors during handoffs and cross-coverage in more depth. Our findings also emphasize the importance of the initial verbal handoff, which can convey errors or imprecisions that are often not verified later in the process. In this manner, the handoff process resembled the children’s game of “telephone.” It is therefore particularly important to avoid introducing errors at any stage of the handoff or sign-out process.

Our EHR is a home-grown system, used by all healthcare providers in our hospital. Although data entry is specific for each type of professional, the content of the EHR is accessible to all. The EHR is structured in various tabs, and include the following:

- A list of reports, admission notes and discharge summaries
- Progress notes and consultation notes, which are unstructured, but usually written up as problem lists
- Vital signs, and administered medications in a chart
- Prescribed medications
- Labs
- Imaging
- Nursing notes

As in many EHR systems, most of the content of the tabs are presented in chronological order. The documents of each tab are presented separately, which means that physicians need to jump from one tab to another to concatenate the information from the reports, admission notes and progress and consultation notes. Collection of this data in the EHR, or “chart biopsy,” may be facilitated when the day team compiles the information in the progress notes. An eye-tracking study of physicians reading electronic notes, however, showed that large sections are ignored, which emphasizes the need to revise the content of electronic notes. Furthermore, reports from the end of the day shift however may not be included in the daily summary, if there is no “new report” alert (which can simply be bold characters, colored text, for example) in the EHR. One way to improve this process would be have better traceability of prescribed labs and work-up tests, and notifications for the final reports. These notifications should reach the patient’s physicians, day or night team, to help guarantee that all results can be acknowledged in a timely manner, and interpreted in the patient’s context.

Improving the design of the EHR system to better support the clinical workflow process would be another way to help prevent errors. The EHRs needs to support several different goals such as billing, quality monitoring and clinical documentation for patient care. The structure of the clinical documentation component in our EHR is similar to the paper charts we initially used, with various sections (i.e., tabs in the EHR). Designing a support tool for handoffs needs to support the workflow and reasoning process of physicians, pulling information from different areas of the EHR into a coherent overview of a patient’s clinical state and trajectory. Although a “snapshot” of the patient’s current state is supported in many EHRs via a dashboard summary page, the challenge lies in tracing the patient’s evolution during the hospital stay, which may not be well described in the progress notes.

When we consider these requirements, mobile solutions seem to offer opportunities for patient-safe options due to their ubiquitous use and possibilities of personalized alerts. Handoff support tools need to be readily available at all times, to provide the relevant information needed to review patient care throughout a shift. A thoughtfully designed mobile app can pull together information from different parts of the EHR, individualized according to user preferences and each patient’s clinical context. Mobile solutions can allow alerts to target the most appropriate user to help avoid alert fatigue.

Understanding the sequence of events that occurring during a hospital stay can be difficult, if relying solely on the EHR. Our EHR provides a detailed graph of vital signs and medications over time, but tracing the history of events can be tricky. Users may try different approaches from different tabs, using search engines in the prescription history, or skimming through the progress notes. Improving the timeline view, in particular using event-based visualization could help address this issue. Furthermore, some authors have looked at handoffs as narratives, that can help care-providers understand the patient’s story, and the sequence of events that have occurred during the hospitalization.
Improving the progress notes in the EHR could also help improve information retrieval. Progress notes are often structured as a problem list and plan, and allow physicians to attribute priorities to the patient’s problems, describe the plan, gather the results and provide contingency statements if needed. Currently, the tab structure in our EHR system makes sharing of information across tabs tedious, having to switch back and forth from one tab to another. Building a medical chart around the problem list, which integrates the important findings in a single place could help avoid excess switching. If we consider the clinical reasoning process, we blend parts of the history taking, the physical examination, labs and imaging results into each problem.

The analysis of “erroneous facts” errors revealed difficulties with timeline and event sequencing. During the verbal handoff, the receiver needs to combine many actions, listening and understanding the case, taking notes, anticipating complications, and having to deal with potential interruptions (phone calls, for example). Although receiving a summary is helpful, care must be taken to provide a simple outline with minimal administrative information to avoid cognitive overload. When participants need to read long paragraphs about each problem, their attention to the handoff giver is low, and can lead to misunderstandings.

Strengths and limitations
Handoffs, in particular the verbal handoffs, are difficult to study in a real clinical setting. Differences in patient cases, or changes in patient state do not allow for comparisons of performances. Therefore, we chose to use a simulation setting to reproduce the handoff and cross-coverage context. Although the number of cases was lower than the total average number of cases handed off in the wards, we believe that six cases allow us to address concerns of case specificity. This number is also sufficient to challenge the participants’ cognitive load, and allowed for a realistic, albeit busy, beginning of a nightshift. Standardization of the initial handoff also facilitated the comparison among participants. Although participants reported that complexity of cases was similar to real patients, standardization limits the variance and sources of complexity that complicate communication during handoffs.

The analysis for errors was based mainly on facts reported during the sign-out process. We did not include an analysis of the clinical reasoning process during patient management for this report.

Finally, this study focuses on internal medicine physicians, so the findings may not necessarily be generalizable to other medical specialties. The wide variation in handoff practices also limits its generalizability in other institutions or departments.

Conclusion
Based on a simulated handoff and chart biopsy study, we performed an analysis focused on the errors and imprecisions that were observed during the final sign-out, with no significant differences in error rate by type of support tool. Our findings emphasize several ways to help avoid errors in the handoff process, in the design of EHRs and of support tools, as well as by addressing human factors in future training programs. We propose facilitating the cross-coverage process by providing event-based medical charts, which regroup documents from related events together. Support tools for handoffs should be able to easily present updated patient data in a concise way and should facilitate note-taking. Furthermore, handoff training should raise awareness for over-reliance on support tools, and low verification of the collected information. We plan to apply these human factors training suggestions, and emphasize the narrative aspect of handoffs in our next study. Future studies on the handoff process should also include an analysis of the written handoff process.

Acknowledgements
We acknowledge V. Soulier and R. Wipfli for their contribution to this project.

References


Alex F. Bokov, Ph.D.\textsuperscript{1}, Gail P. Olin, M.S. R.N.\textsuperscript{2}, Angela Bos\textsuperscript{1}, Alfredo Tirado-Ramos, Ph.D.\textsuperscript{1}, Pamela Kittrell, R.N.\textsuperscript{1}, Carlayne Jackson, M.D.\textsuperscript{1}

\textsuperscript{1}University of Texas Health Science Center at San Antonio, San Antonio, TX; \textsuperscript{2}University of Texas Health Science Center at Houston, Houston, TX

Abstract

We present a method for rapidly ranking all distinct facts in an electronic medical record (EMR) system by how over-represented or under-represented they are in a patient cohort of interest relative to some larger reference population of patients in the same EMR. We have implemented this method as a plugin for i2b2, the open source data warehouse platform widely used in research health informatics. Our method is highly flexible in terms of what medical terminologies it supports and is vendor-independent thanks to leveraging the i2b2 star schema rather than any one specific EMR. It can be applied to a wide range of informatics problems including finding health disparities, searching for variables to include in a risk calculator or computable phenotype, detection of comorbidities, discovery of adverse drug reactions. The case study we present here uses this software to find unlabeled flowsheets for patients suffering from amyotrophic lateral sclerosis.

Introduction

Amyotrophic lateral sclerosis is a neurodegenerative illness affecting upper and lower motor neurons and anterior horn cells, culminating in death due to respiratory insufficiency or malnutrition \textsuperscript{1}. The progression of this disease is monitored via the amyotrophic lateral sclerosis functional rating scale revised (ALSFRS-R) \textsuperscript{2}, the eating assessment test (EAT-10) \textsuperscript{3}, and the Center for Neurologic Study Lability Scale (CNS-LS) \textsuperscript{4}. A CNS-LS score of 13 or greater has been validated [Moore et al. 1997] as an indicator of pseudobulbar affect (PBA). PBA is believed to be caused by degeneration of connections between the brainstem and the frontal cortex resulting in sudden mood swings and inappropriate laughter or weeping. This impacts patient quality of life and in ALS patients PBA is associated with a worse prognosis. It would be of great interest to be able to extract time series of these scores for multiple patients from the EHR for visualization, hypothesis testing, and predictive modeling with research-quality statistical software such a R \textsuperscript{5}. However, EHR systems are not designed for on-demand bulk data extraction. For many systems process is slow and laborious, with each data-pull requiring the local health IT team to write or modify custom reports. Another necessary step prior to longitudinal analysis of EMR data is a thorough characterization of the populations of interest in order to identify variables to later evaluate for inclusion in the statistical model as predictors or covariates.

![Figure 1. The i2b2 interface showing terms as they appear when normally annotated in the metadata and linked to the star-schema. Any combination of nodes in the hierarchical menu on the left of the screen can be dragged and dropped to the right side to create a query. This is the existing functionality of the i2b2 system. Our extension to it is not visible in this figure, but is presented below, in the Methods section.](image-url)
Integrating Informatics from Bench to Bedside (i2b2) is a data warehouse platform that offers a more flexible data model for research purposes. However, importing data into i2b2 from commercial EHR systems inevitably involves some degree of reverse-engineering and detective work. One common problem that occurs with nursing flowsheet data is that individual flowsheet components are imported into i2b2 but the mapping to their parent flowsheet names is not available. An more general methodological gap in health services research is the need for semi-automated cohort characterization, particularly of rich datasets such as EMR.

Figure 2. Design of Chi²notype plugin (light blue) and its integration with i2b2 (dark blue). The SQL back-end uses two dedicated tables (created in an account specified by the system administrator in a configuration file). The first table, PCONCEPTS, contains all unique combinations of PATIENT_NUM (the ID field i2b2 uses to distinguish patients) and CONCEPT_CD (the ID field i2b2 uses to distinguish EHR concepts). The second table, PCONCEPTS_COUNTS, is initialized with a prefix column (for filtering groups of concepts, as will be described below), a CONCEPT_CD column, a NAME column containing a short human-readable description of each fact (if such a description is available in i2b2), a TOTAL column (which represents the total number of distinct patients associated with each fact in PCONCEPTS), and a RFC_TOTAL column (which represents the value of the TOTAL column for each fact divided by the overall number of distinct patients). With each new patient-set that gets queried, a new pair of columns gets added to the PCONCEPTS_COUNTS table; these correspond to the TOTAL and RFC_TOTAL for that patient set but the names of all columns after the first two are based on the MASTER_ID, INSTANCE_ID, and RESULTSET_ID for that patient-set as recorded by i2b2’s audit log in whichever account i2b2 uses to store data (by default, I2B2DEMODATA). This way, if a patient-set is ever queried again, its existing columns will be found in PCONCEPTS_COUNTS and reused, thus speeding response times. Given such a table, it is trivial to calculate $\chi^2$ statistics, sort the results, and filter the output natively in SQL.

As part of the Greater Plains Collaborative Clinical Data Research Network [ref] we use the HERON open source ETL process [7] developed at the University of Kansas Medical Center. The approach HERON uses is to empirically derive an ontology using agglomerative hierarchical clustering by binary similarity score. While such an approach is a practical means to tackle the global problem of how to represent flowsheet data without prohibitively time-consuming manual curation, it has important limitations. First, it is of limited transferability to other sites; even sites that use the same EHR vendor as HERON’s differ in the meaning of a large fraction of the internal IDs that represent each flowsheet measure. Other sites cannot simply copy the flowsheet ontology as used by HERON--each site must run its own computationally-intensive clustering process. Furthermore, knowing that a given group of measures are similar to each other still does not always reveal what the data recorded in them actually means.

Here we demonstrate a bottom-up, incremental approach that is complementary to that of HERON: empirically finding under-documented EHR facts by means of an analytical tool called Chi²notype which we created to interoperate with i2b2. Chi²notype’s usefulness is not limited to flowsheets -- here we also use it to characterize the demographics, diagnoses, and medications of ALS patients in our system compared to a sub-cohort of those patients who suffer from PBA.

**Methods**

**Cohort** We searched i2b2 for all patients for whom there has ever been reported an ICD9 8–9 or ICD10 diagnosis code of 335.20 or G12.21 respectively excluding problem-list modifiers of "Deleted" or "Resolved". This resulted in a
patient-set of 839 individuals. In i2b2, patient-sets refer to user-defined groups of patient-IDs (de-identified in our case) that can be saved and re-used as inclusion/exclusion criteria for more complex queries and also as input for various visualization and reporting i2b2 plugins. This patient-set was compared to all patients in the EMR system to find data-elements unique to ALS. A second patient-set was generated with the diagnosis inclusion criteria but in addition one or more of the following: diagnosis of emotional lability (799.24 and R45.86 for ICD9 and 10 respectively); or pseudobulbar affect (310.81 and F48.2 for ICD9 and 10 respectively); or a CNS-LS total score \( \geq 13 \). This ALS/PBA patient set was compared to the overall ALS patient set in order to obtain a demographic breakdown for both and to discover features that significantly are different for ALS patients with PBA relative to ALS patients in general.

**Software** We wrote an i2b2 plugin, Chi2notype, for automated cohort-characterization. Chi2notype ranks all EMR facts in the OBSERVATION_FACT table of i2b2’s star schema by their prevalence in a given patient-set relative to a second patient-set that is a strict super-set of the first (by default, all patients that have encounter-level data in i2b2).

In other words, our tool makes it possible to see how over- or under-represented each EHR fact is in any user-defined group of patients.

**Workflow** The basic pattern of specifying a cohort, then a sub-cohort based on it, and then determining EMR data elements significantly over- or under-represented in the sub-cohort relative to the larger cohort can directly produce a demographic table or give hints about the interpretation of under-documented data elements. It can also be used as a building block for more complex workflows. At a high level, the basic steps for retrospective data analysis of EMR data can be viewed as:

1. Convert verbal/conceptual cohort inclusion/exclusion criteria into an unambiguous computer executable query (e.g. i2b2) that relies only on data that is available.
2. Likewise translate predictors, covariates, and response variables into another unambiguous computer executable query. Again, adjust plan for feasibility and interpretability.
3. When the two above queries are finalized, use them to pull visit-level data (for self-service systems such as at our institution) or to create a data request for the appropriate informatics team.
4. Proceed with analysis plan-- development and validation subsets, model specification and variable selection, etc.

Chi2notype facilitates step 1 by making the researcher aware of disparities between the cohort and sub-cohort which are potential confounders but in many cases can be addressed by a balanced sampling strategy or by inclusion as covariates in the statistical model. Disparities that have a plausible link to the outcome of interest can be added to the data element specification of step 2 so they would be available for evaluation as additional predictors during the variable selection of step 4. Likewise, the characterization process can also suggest possible outcomes of secondary interest to include.

The ranking is done by odds ratio. A \( \chi^2 \) statistic is also calculated for each EMR data element as follows:

\[
\sum \frac{|N_{ij} - E_{ij}|^2}{E_{ij}} \sim \chi^2 \text{ where } E_{ij} = \frac{N_j \cdot N_i}{N} \text{ and...}
\]

- \( X_{all} \) = any set of patients from the EHR system
- \( N \) = total number of patients in \( X_{all} \).
- \( \hat{X}_i \) = a cohort selected from \( X_{all} \) such that \( \sum \frac{|N_{ij} - E_{ij}|^2}{E_{ij}} \sim \chi^2 \).
- \( N_{ij} \) = total number of patients in \( X_i \).
- \( N_{i,j} \) = any fact from the EHR
- \( N_{i,j} \) = total number of patients in \( X_{all} \) who have fact \( j \) recorded at least once in their record.

460
Table 1. A demographic breakdown of the overall ALS cohort side-by-side with the ALS + PBA sub-cohort. None of the difference in prevalence were statistically significant.

<table>
<thead>
<tr>
<th>Description</th>
<th>#ALS</th>
<th>% ALS</th>
<th># ALS</th>
<th>% ALS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ethnicity</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-Hispanic</td>
<td>488</td>
<td>58.16%</td>
<td>81</td>
<td>54.00%</td>
</tr>
<tr>
<td>Hispanic</td>
<td>189</td>
<td>22.53%</td>
<td>43</td>
<td>28.67%</td>
</tr>
<tr>
<td>Unknown/Other</td>
<td>143</td>
<td>17.04%</td>
<td>25</td>
<td>16.67%</td>
</tr>
<tr>
<td>Not Recorded</td>
<td>12</td>
<td>1.43%</td>
<td>0</td>
<td>0.00%</td>
</tr>
<tr>
<td>Answer Refused</td>
<td>7</td>
<td>0.83%</td>
<td>1</td>
<td>0.67%</td>
</tr>
<tr>
<td>Race</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White or Caucasian</td>
<td>626</td>
<td>74.61%</td>
<td>118</td>
<td>78.67%</td>
</tr>
<tr>
<td>African-American</td>
<td>18</td>
<td>2.15%</td>
<td>2</td>
<td>1.33%</td>
</tr>
<tr>
<td>Asian</td>
<td>10</td>
<td>1.19%</td>
<td>1</td>
<td>0.67%</td>
</tr>
<tr>
<td>Unknown/Other</td>
<td>148</td>
<td>17.64%</td>
<td>25</td>
<td>16.67%</td>
</tr>
<tr>
<td>Not Recorded</td>
<td>24</td>
<td>2.86%</td>
<td>0</td>
<td>0.00%</td>
</tr>
<tr>
<td>Answer Refused</td>
<td>13</td>
<td>1.55%</td>
<td>4</td>
<td>2.67%</td>
</tr>
<tr>
<td>Language</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>English</td>
<td>762</td>
<td>90.82%</td>
<td>139</td>
<td>92.67%</td>
</tr>
<tr>
<td>Spanish</td>
<td>27</td>
<td>3.22%</td>
<td>5</td>
<td>3.33%</td>
</tr>
<tr>
<td>Unknown</td>
<td>48</td>
<td>5.72%</td>
<td>6</td>
<td>4.00%</td>
</tr>
<tr>
<td>Other</td>
<td>2</td>
<td>0.24%</td>
<td>0</td>
<td>0.00%</td>
</tr>
<tr>
<td>Sex</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>363</td>
<td>43.27%</td>
<td>78</td>
<td>52.00%</td>
</tr>
<tr>
<td>Male</td>
<td>476</td>
<td>56.73%</td>
<td>72</td>
<td>48.00%</td>
</tr>
<tr>
<td>Marital Status</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Married</td>
<td>503</td>
<td>59.95%</td>
<td>101</td>
<td>67.33%</td>
</tr>
<tr>
<td>Single</td>
<td>127</td>
<td>15.14%</td>
<td>11</td>
<td>7.33%</td>
</tr>
<tr>
<td>Unknown</td>
<td>118</td>
<td>14.06%</td>
<td>22</td>
<td>14.67%</td>
</tr>
<tr>
<td>Divorced</td>
<td>43</td>
<td>5.13%</td>
<td>6</td>
<td>4.00%</td>
</tr>
<tr>
<td>Widowed</td>
<td>40</td>
<td>4.77%</td>
<td>8</td>
<td>5.33%</td>
</tr>
<tr>
<td>Legally Separated</td>
<td>6</td>
<td>0.72%</td>
<td>1</td>
<td>0.67%</td>
</tr>
<tr>
<td>Other</td>
<td>2</td>
<td>0.24%</td>
<td>1</td>
<td>0.67%</td>
</tr>
<tr>
<td>Distance of Residence from Health System</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt; 5 mi</td>
<td>94</td>
<td>11.20%</td>
<td>20</td>
<td>13.33%</td>
</tr>
<tr>
<td>5-10 mi</td>
<td>110</td>
<td>13.11%</td>
<td>18</td>
<td>12.00%</td>
</tr>
<tr>
<td>10-15 mi</td>
<td>132</td>
<td>15.73%</td>
<td>22</td>
<td>14.67%</td>
</tr>
<tr>
<td>15-20 mi</td>
<td>32</td>
<td>3.81%</td>
<td>6</td>
<td>4.00%</td>
</tr>
<tr>
<td>20-50 mi</td>
<td>109</td>
<td>12.99%</td>
<td>20</td>
<td>13.33%</td>
</tr>
<tr>
<td>50-100 mi</td>
<td>166</td>
<td>19.79%</td>
<td>22</td>
<td>14.67%</td>
</tr>
<tr>
<td>100-200 mi</td>
<td>107</td>
<td>12.75%</td>
<td>18</td>
<td>12.00%</td>
</tr>
<tr>
<td>State of Residence</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Texas (TX)</td>
<td>825</td>
<td>98.33%</td>
<td>146</td>
<td>97.33%</td>
</tr>
<tr>
<td>Other</td>
<td>13</td>
<td>1.55%</td>
<td>3</td>
<td>2.00%</td>
</tr>
<tr>
<td>Vital Status</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Living</td>
<td>688</td>
<td>82.00%</td>
<td>114</td>
<td>76.00%</td>
</tr>
<tr>
<td>Deceased</td>
<td>141</td>
<td>16.81%</td>
<td>36</td>
<td>24.00%</td>
</tr>
</tbody>
</table>

This is more informative than ranking EHR facts strictly by their prevalence in the patient-set of interest, since with such a ranking there would not be any way to know whether the prevalence of a fact is a phenomenon specific to this patient-set or to the entire population.

By default Chi2notype returns the top-10 and bottom-10 of all EHR facts that occur in at least 10 patients in the reference group. These defaults can be changed by the user (Figure 3), and instead of all EHR facts, individual categories can be selected (based on the SCHEMES table from the i2b2 metadata). The full result-set can be exported as a .CSV file that can be opened with spreadsheet software where custom user-specific filters and functions can be applied. The downloaded spreadsheets have the same column and row structure (i.e. aggregated data not individual visits, which is not purpose of this software, but we have previously published a different app for that scenario [ref]).

To deploy Chi2notype to a pre-existing i2b2 server, one would download it from https://github.com/UTHSCSA-CIRD/chinotytype/, place the individual components in the parts of the file-system directed by the instructions, set up the necessary accounts and permissions, log directly into the database back-end of i2b2, and start using it (it will create the tables it needs automatically when it is first used). This project is still under intense development, but we are happy to help via the GitHub ticket system, those who wish to deploy Chi2notype at their own i2b2 sites. Chi2notype depends only on the database tables used by i2b2 and has no dependencies whatsoever on the underlying EHR. So far this software has only been tested in a Linux environment against an ORACLE back-end, but the SQL code is written to be as close as practical to the ANSI SQL standard, so it should present minimal obstacles to porting to other RDBMS systems (and the open-source license under which we are distributing this software encourages such collaborative modifications).

Results

We filtered the Chi2notype output for the ALS patient-set so as to only return flowsheet measures. We found measures that, while exceedingly rare in the overall patient pool (< 0.1%) were present in 65-71% of ALS patients. Moreover, the exact same prevalence was shared by multiple measures, which suggested that each was obtained from a patient during the same session (Table 1).
These flowsheet components were among the many in our data warehouse for which the parent flowsheets were not automatically mapped by the ETL. However, when the concept codes identified by Chi2notype were searched in the raw EHR data, we found exact matches to the ALSFRS, EAT-10, CNS-LS flowsheets appearing in the EMR clinical front-end user interface. This allowed us to create new i2b2 concept paths which in turn make all these flowsheets and their individual components available as additional inclusion or exclusion criteria for researchers querying our data. Furthermore, another plugin we are developing makes it possible to extract the contents of these flowsheets as columns in a simple, tabular, time-series format side-by-side with other variables of interest to secondary analysis such as visit vitals, medications, and procedures.

We then used the total score from the CNS-LS flowsheet to identify ALS patients who at some point in their history had a score greater than or equal to 13 (or had other indicators of emotional lability among their diagnosis scores) Instead of asking "what flowsheet IDs that lack descriptions are over-represented in ALS patients relative to the whole patient population?", we were asking a broader question about a smaller population: "what diagnoses, medications, and lab values differ significantly between the 150 subcohort of patients who have ALS and PBA versus the overall cohort that includes all ALS patients?". These results are shown in Table 2, and a demographic summary of ALS and ALS/PBA patients is shown in Table 1.
We found, that among ALS patients with PBA, there was an increased incidence of dysphagia, fatigue, emotional lability, as well as difficulties with breathing and speech. Among the drugs that are significantly different in prevalence are dextromethorphan/quinidine which is used to treat PBA, Riluzole which is used to slow progression of ALS, Tizanidine which is used to alleviate muscle spasms, and Jevity which is a nutritional formula used percutaneous endoscopic gastrostomy (PEG) tubes for ALS patients who are suffering from dysphagia. The only data elements whose incidence was diminished among ALS+PBA patients were counts of above-reference values for plasma blood urea nitrate/creatinine and creatine kinase (relative to the total number of patients on whom those tests were done in the two cohorts respectively).

**Conclusion**

Chi²otype is an open-source plug-in for i2b2 that is capable of rapidly and intuitively generating a list of facts relevant to a patient-set of interest for quality control, cohort characterization, and narrowing down the list of candidate variables to consider for inclusion in analysis. We used this approach to identify flowsheet measures that had no human-readable names or annotations, purely on the fact that they are known to be over-represented in the patient-set of interest. We then compared a sub-cohort of ALS patients exhibiting signs of PBA to the overall ALS cohort and found increased prevalence of ALS complications in patients with PBA, consistent with the recently published results of Tortelli et al. [ref].

Chi²otype allows the researcher to quickly find EMR variables that have the strongest positive and negative correlation with membership in a cohort of interest for the purposes of data characterization and rational selection of variables for inclusion in an analytic dataset.

**Acknowledgments**

This work was supported by PCORI grants CDRN-1306-04631 and CDRN-1501-26643NIH

**References**


Objective: Build and validate a clinical decision support (CDS) algorithm for discharge decisions regarding referral for post-acute care (PAC) and to what site of care. Materials and Methods: Case studies derived from EHR data were judged by 171 interdisciplinary experts and prediction models were generated. Results: A two-step algorithm emerged with area under the curve (AUC) in validation of 91.5% (yes/no refer) and AUC 89.7% (where to refer). Discussion: CDS for discharge planning (DP) decisions may remove subjectivity, and variation in decision-making. CDS could automate the assessment process and alert clinicians of high need patients earlier in the hospital stay. Conclusion: Our team successfully built and validated a two-step algorithm to support discharge referral decision-making from EHR data. Getting patients the care and support they need may decrease readmissions and other adverse events. Further work is underway to test the effects of the CDS on patient outcomes in two hospitals.

Introduction

In response to the Readmission Reduction (1) and Value Based Purchasing Programs (2) hospitals are increasingly collaborating with post-acute care (PAC) services to mitigate readmission risk. Between 1996 and 2010, discharge of patients to PAC settings, one of the fastest growing sectors of the Medicare program, showed a relative increase of approximately 50%. (3,4) Partnering with and shifting care to the post-acute care setting increases the importance of properly identifying the right patients for the right settings.

Discharge planning (DP), a process undertaken during the hospital stay, is used to proactively identify patients’ needs after hospital discharge and to begin a smooth transition from hospital to home or other care settings. Yet, several issues exist that have the potential to undermine quality of care and patient safety. DP is a complex process conducted by a multidisciplinary healthcare team, (i.e., physicians, nurses, social workers, and physical therapists), in collaboration with patients and their caregivers. Older adults with multiple comorbidities, complex treatment regimens, and the rate at which they churn through the hospital system add to the complexity of decision making. (5-7) Despite best efforts, the DP process is often rushed, with clinicians frequently not having all the necessary information or time to make optimal decisions regarding PAC referrals. (8) Furthermore, studies have revealed great variability in DP assessments and models, (9) clinician risk tolerance and decision-making regarding PAC referrals (10-15) and insurance barriers and a lack of standards for post-acute care decision making during this critical time. (15) Inconsistent practices for DP and identification of the need for PAC services can result in either over-referral for PAC services, a significant Medicare expense, or under-referral, leaving patients with unmet needs. (16) The need for decision support tools to bring structure to this non-standardized, but critical, process is imperative.

Currently, few CDS tools exist that provide recommendations for care after hospital discharge. Barsoum and colleagues developed and validated a tool to identify patients that should not be discharged directly to home after total joint arthroplasty. (17) Holland and team (18) created a tool to identify high-risk patients in need of focused discharge planning. Tseng and colleagues demonstrated improved DP when using a systematic assessment tool versus traditional assessment, (19) while others developed a tool to identify surgical patients who should not be discharged to home. (20) Our earlier work with a CDS called the Discharge Decision Support System (D2S2) identified six statistically significant factors associated with likelihood for PAC referral of older adults, including patient’s age, less than excellent self-rated health, no or intermittent help at home, major walking limitations, depression symptoms and number of comorbid conditions. (10) The D2S2 was commercialized and is translated into practice in 38 hospitals. However, to our knowledge there are currently no CDS tools reported in the literature that provide recommendations for PAC referral for a general hospital population and considers the specific site of PAC services needed. The CDS described here meets those requirements.

Objectives

Our objective was to build and validate an expert clinical decision support system (CDS) for the discharge referral decisions of whether or not to refer patients for post-acute care and if so, to what level of care. The study aim was to
define and validate the most significantly predictive model of factors to mimic nationally-based, multi-disciplinary experts’ post-acute care referral decisions. The hypotheses (H) for model building were: (H1) There will be a statistically significant correlation between the evidence-based factors (case study information) and the experts’ yes/no referral decisions. (H2) There will be a statistically significant correlation between the evidence-based factors and the site of referral. The hypotheses for the model validation were: (H3) The model will predict the expert referral yes/no decision with sensitivity, specificity and positive predictive value of greater than 80%. (H4) The model will predict the site of referral with sensitivity, specificity and positive predictive value of greater than 80%.

Methods

Setting. The data for the algorithm came from the electronic health records (EHR) of patients cared for in six hospitals located in the New England, Mid-Atlantic and Midwest regions of the United States. The hospitals ranged in size from community, regional, to a quaternary academic medical center. Selected hospitals all had a comprehensive, structured documentation system for nursing called Knowledge Based Charting (KBC) within their larger EHR (described below). The study was approved by the University Institutional Review Board (IRB) and the IRB of two hospitals that required review.

Sample. Each hospital provided retrospective, de-identified EHR data for hospitalized patients age 55 and older cared for on medical and surgical units and critical care. They excluded records coded as observation stays and admissions to skilled rehabilitation units, obstetrics and pediatrics. The study team received the files via secure file transfer. Data retrieval occurred between October 2011 and July 2012 and contained 5,333 patient records meeting our eligibility criteria. Hospitals provided between 908 to 1751 patient records each. The team completed an extensive cleaning and re-coding process, described elsewhere, and then merged the files into one uniform dataset. To assure a nationally representative sample, the statistician drew the final study sample of 1,496 patients from the 5,333 to obtain a representative distribution of the 16 most common primary diagnoses of hospitalized patients. The EHR data populated the electronic case studies used for expert review as described below.

Electronic case study contents. Programmers created structured case studies from the 1,496 patient records. The data for the case studies came from the KBC Adult Patient Profile where nurses documented a thorough admission assessment from patients and caregivers and the Assessment/Intervention Flowsheet contained documented assessments from each shift throughout the hospital stay (e.g., cognitive status, functional status, fall risk). We selected 71 data elements from these documents based on our prior research and the literature to represent the factors that describe the patients’ health, admission course, and are associated with post-acute needs and outcomes. The Orem Self-Care Deficit Theory provided the organizing framework for the case studies. Orem posits that nursing care is appropriate when the person is not able to engage in self-care. This theory appropriately supported our study because we believe that patients who cannot perform self-care after hospital discharge need post-acute care. Orem describes 10 categories of basic conditioning factors that may affect self-care: age; gender; developmental state; health state; sociocultural orientation; health care system factors; family systems factors; patterns of living; environmental factors; and socioeconomic factors. For example, the category of health state included primary diagnoses, co-morbid conditions, medications, procedures, wounds and other elements listed in Table 1.

Table 1 Case Study Variables

<table>
<thead>
<tr>
<th>Ability to Learn</th>
<th>Comorbidity Name</th>
<th># of Discharge Medications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Admission Type</td>
<td>Consciousness Level</td>
<td>Nutrition Risk Eating Poorly</td>
</tr>
<tr>
<td>Age At Admission</td>
<td>Discharge Medication</td>
<td>Nutrition Risk Weight Loss</td>
</tr>
<tr>
<td>Ambulation Function Changed</td>
<td>Dressing Function Changed</td>
<td>Orientation</td>
</tr>
<tr>
<td>Ambulation Current</td>
<td>Dressing Current</td>
<td>Pain Rating Rest</td>
</tr>
<tr>
<td>Ambulation Prior</td>
<td>Dressing Prior</td>
<td>Past 6 Month ED Visit</td>
</tr>
<tr>
<td>Assistive Equipment Used</td>
<td>Eating Function Changed</td>
<td>Past 6 Month Hospital Stay</td>
</tr>
<tr>
<td>Barriers Follow Medication Schedule</td>
<td>Eating Current</td>
<td>Personal History Family History</td>
</tr>
<tr>
<td>Bathing Function Changed</td>
<td>Eating Prior</td>
<td>Presence of a Wound</td>
</tr>
</tbody>
</table>
Multi-Disciplinary Experts. Physicians, and masters or doctorally prepared nurses, social workers, and physical therapists were recruited via snowball sampling to review the case studies and provide their decision making expertise. The experts accessed the case studies via a website built to display their assigned case studies, track their progress, and capture their decisions. A detailed description of expert recruitment, the web-based expert knowledge elicitation process, and Delphi rounds is published elsewhere. 

Stratified by geographic region of the United States (north, south, east and west) and discipline (doctor, nurse, physical therapist, social worker), the statistician randomly assigned 171 experts to teams of three to review and judge approximately 13-30 unique case summaries each. Experts were paid as independent consultants per case. Agreement of two out of three experts was accepted as consensus. If all three disagreed on the site of care, the cases went through up to two Delphi rounds seeking majority agreement (2 out of 3). If agreement was not reached the case was not used in the “where to” modeling. Along with their “yes/no refer” decisions and “where to” decisions (i.e. home care, inpatient rehabilitation, skilled nursing facility, nursing home, or hospice) experts identified the factors within each case study that supported their decisions. We used these data to create and validate the two-step expert algorithm.

**Data Analysis**

We randomly separated the 1496 cases into a training dataset and a validation dataset. We used a number of analytical methods to determine the best model for each decision. These included: 1) Penalized/Regularized Logistic Regression Models performed in R using the glmpath package and 2) classification and regression trees (CART) performed in R using regression trees, or the ctree command included in the party package. Factors were added to models based on increasing the predictive ability of the algorithm above that already been obtained, while penalizing for the complexity of the model and structure of the data. Separate models were examined where (a) factors were looked at as "blocks" to be included/excluded together, (b) as indicators for level, and (c) based on severity cut-points. The optimal model was statistically determined based on area under the curve (AUC), and Akaike and Bayesian information criterion (AIC & BIC).

**Modeling yes or no refer.** Patient characteristics selected by experts as important for their decision were not as informative for reducing the data and pinpointing important variables as we had initially hoped. In the final round, for each case by each expert, they selected from one to 71 different items. In all, the experts marked 4,488 characteristics (variables) as important within the 1,496 cases with a mean of 23 per case (SD=10.75). Furthermore, the make-up of important/not-important for referral decision was statistically significant for all variables. Consequently, all 71 variables (Table 1) were initially candidates for inclusion in the prediction models.

---

<table>
<thead>
<tr>
<th>Bathing Current</th>
<th>Education</th>
<th>Primary diagnosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bathing Prior</td>
<td>Employment Status</td>
<td>Race</td>
</tr>
<tr>
<td>Braden Score</td>
<td>Ethnicity</td>
<td>Self-Rated Health</td>
</tr>
<tr>
<td>Caregiver Ability</td>
<td>Fall Risk Score</td>
<td>Surgical Procedure Term</td>
</tr>
<tr>
<td>Caregiver Availability: day of week</td>
<td>Financial Concern</td>
<td>Toileting Changed</td>
</tr>
<tr>
<td>Caregiver Frequency of Availability</td>
<td>Gender</td>
<td>Toileting Current</td>
</tr>
<tr>
<td>Caregiver Presence (yes/no)</td>
<td>Hearing</td>
<td>Toileting Prior</td>
</tr>
<tr>
<td>Caregiver Availability: time of day</td>
<td>Home Accessibility</td>
<td>Transferring Changed</td>
</tr>
<tr>
<td>Caregiver Relationship</td>
<td>Incision Appearance</td>
<td>Transferring Current</td>
</tr>
<tr>
<td>Caregiver Understanding</td>
<td>Incision Location</td>
<td>Transferring Prior</td>
</tr>
<tr>
<td>Caregiver Willingness</td>
<td>Length of Stay</td>
<td>Vision</td>
</tr>
<tr>
<td>Communication Function Changed</td>
<td>Lives With</td>
<td>Wound Type</td>
</tr>
<tr>
<td>Communication Current</td>
<td>Living Arrangement</td>
<td></td>
</tr>
<tr>
<td>Communication Prior</td>
<td>Marital Status</td>
<td></td>
</tr>
<tr>
<td>Caregiver Understanding</td>
<td>Mental Health</td>
<td></td>
</tr>
<tr>
<td>Caregiver Willingness</td>
<td>Number of Comorbid Conditions</td>
<td>Table 1</td>
</tr>
</tbody>
</table>
Penalized regression uniformly gave the best predictive ability with variable selection based on the BIC criteria. We tested the robustness of the selected variables using 1,000 bootstrap samples. If the predictor did not remain important in at least 50% of the bootstrap samples, its exclusion was tested as a potential refinement of the model. It was retained if its removal affected other variables in the model even if it appeared in <50%. Refinements included: excluding primary diagnosis variables, excluding specific comorbidities, and including combinations of variables as suggested by clustering or CART.

For each of the possible refinements, the ROC curve produced was compared to the ROC curve for the “baseline” model using statistical tests based on both U-statistics and bootstrapping via the pROC package. The “best” (least complex) model with similar statistical properties to the “baseline” model was found to be the one with primary diagnosis and comorbidity variables excluded, and included one of the combinations that was suggested by CART.

**Modeling Where to Refer.** The model for referral location included the variables within 1,204 cases the experts determined the outcome was “yes” refer for care. We excluded 17 cases with uncertain location, even after two Delphi rounds. A number of steps led to the final model: 1) Models for predicting the five referral sites were generated. All methods produced poor performing models, with some models doing worse than chance. 2) Hospice was excluded since only 66 patients were determined as a hospice referral by the experts and the cost of misclassifying a patient to hospice could be high. The main error in prediction was among the facility-based locations. Based on results from steps one and two, we collapsed the referral site outcomes to create the dichotomous referral outcomes of facility versus home care. Since cases who currently reside in a facility were highly likely to return to a facility, these subjects were referred to a facility in the first split of the model. Penalized Logistic Regression models best performed prediction for the remainder of the cases with the outcome facility versus home care.

**Results**

The 1,496 patients were on average 74.3 years old (11.25 SD, range 55-103). Gender distribution was 55.5% female, 83.4% White and 12.1% Black or African American and 98.3% were non-Hispanic. Among a large variety of primary diagnoses, the most common diagnoses were Pneumonia, Atrial Fibrillation, Acute Exacerbation of Obstructive Chronic Bronchitis, Unspecified Septicemia, and Acute Kidney Failure. Average number of secondary diagnoses and co-existing conditions was 10.29 (range 1 to 32, SD=5.73). Table 2 provides a fuller description of the sample.

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Means (SD) [Range] or n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>74 (11.25), [55 – 103]</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>665 (44.45)</td>
</tr>
<tr>
<td>Female</td>
<td>831 (55.55)</td>
</tr>
<tr>
<td>Race</td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>1247 (83.36)</td>
</tr>
<tr>
<td>Black</td>
<td>181 (12.10)</td>
</tr>
<tr>
<td>Other</td>
<td>27 (1.80)</td>
</tr>
<tr>
<td>Missing</td>
<td>41 (2.74)</td>
</tr>
<tr>
<td>Ethnicity</td>
<td></td>
</tr>
<tr>
<td>Hispanic/Latino</td>
<td>25 (1.67)</td>
</tr>
<tr>
<td>Non-Hispanic Latino</td>
<td>976 (65.24)</td>
</tr>
<tr>
<td>Missing</td>
<td>495 (33.09)</td>
</tr>
<tr>
<td>Education</td>
<td></td>
</tr>
<tr>
<td>Elementary</td>
<td>125 (8.36)</td>
</tr>
<tr>
<td>High school</td>
<td>763 (51.00)</td>
</tr>
<tr>
<td>College</td>
<td>378 (25.27)</td>
</tr>
<tr>
<td>Graduate/Postgraduate</td>
<td>88 (5.88)</td>
</tr>
<tr>
<td>Missing</td>
<td>142 (9.49)</td>
</tr>
<tr>
<td>Employment Status</td>
<td></td>
</tr>
<tr>
<td>Employed</td>
<td>199 (13.30)</td>
</tr>
<tr>
<td>Not employed</td>
<td>208 (13.90)</td>
</tr>
<tr>
<td>Retired</td>
<td>936 (62.57)</td>
</tr>
<tr>
<td>Missing</td>
<td>153 (10.23)</td>
</tr>
</tbody>
</table>
Hypothesis one (H1). H1 was supported. H1—There is a statistically significant correlation between the evidence-based factors (clinical information) and the experts’ yes/no referral decisions. Of the 71 case summary variables, 16 variables (each associated with particular values or interaction terms), were identified as significantly correlated and important to the decision of whether or not to refer the patient for post-acute care. Significant variables include, for example, employment status, fall risk, several activities of daily living, and number of comorbid conditions. An optimal cut off score indicated the difference between refer and do not refer. Table 3 lists the variables important for the yes/no refer prediction and the percentage of 1,000 bootstrap samples in which they appeared. The remaining factors were discarded as their inclusion did not sufficiently improve the predictive ability of the algorithms while increasing the complexity of the models. In some instances, the same variable with different values yielded significant correlations.

Table 3: Predictors of the expert’s decisions to refer patients for post-acute care or not and the percentage of 1,000 bootstrap samples in which they appeared.*
Employment Status: retired or working 92.5
# of hospital stays within past 6 months 51.2
Fall Risk Score 97.6
Equipment 94.9
Home Accessibility 52.8
Wound Present 68.3
Ambulation Current-level of ambulation function on admission 99.9
Ambulation Change-level of decline in ambulation function from level A to B by discharge 85.7
Ambulation Change-level of decline in ambulation function from level B to C by discharge 58.3
Transfer Change-level of decline in transfer function from level A to B by discharge 80.3
Transfer Change-level of decline in transfer function from level B to C by discharge 81.0
Bathing Change-level of decline in bathing function from level A to B by discharge 84.6
Eating Prior-level of eating function on admission 50.8
Number of Comorbidities 98.0
Caregiver Presence: yes/no 56.1
Discharged on Narcotics 52.9
Interaction Ambulation Current X Transfer Change X Caregiver Presence 99.9

*detailed co-efficients, cut off scores, and levels of measurement are omitted to protect intellectual property

Hypothesis two (H2). H2 was partially supported. H2-There is a statistically significant correlation between the evidence based factors and the site of referral. Models for predicting the five referral sites (home health care, in-patient rehabilitation facility (IPF), skilled nursing facility (SNF), nursing home (NH) and hospice) were generated but they proved to be poor performers. Based on the results, we collapsed the site outcomes to depict facility (SNF, IPF, NH) versus home health care referral thus achieving a satisfactory model. For this model there are 13 variables, each associated with particular values or interaction terms, identified as important to the decision of where to refer those patients. Significant variables include for example caregiver availability, pressure ulcer risk, and several activities of daily living, and number of comorbid conditions. An optimal cut off score indicated the difference between home health care referral versus facility. Table 4 lists the variables important for the where to prediction and the percentage of 1,000 bootstrap samples in which they appeared. The remaining factors were discarded as their inclusion did not sufficiently improve the predictive ability of the algorithms while increasing the complexity of the models. In some instances, the same variable with different values yielded significant correlations.

Table 4. Predictors of the experts’ decisions of where to refer the patients for post-acute care (home care versus facility care) and the percentage of 1000 bootstrap samples in which they appeared.*

<table>
<thead>
<tr>
<th>Factor</th>
<th>% Bootstraps</th>
</tr>
</thead>
<tbody>
<tr>
<td>Braden Score (pressure ulcer risk)</td>
<td>95.0</td>
</tr>
<tr>
<td>Fall Risk Score</td>
<td>60.1</td>
</tr>
<tr>
<td>Ambulation Current-level of ambulation function on admission</td>
<td>74.3</td>
</tr>
<tr>
<td>Ambulation Current-level of ambulation function on admission</td>
<td>95.6</td>
</tr>
<tr>
<td>Ambulation Change-level of decline in ambulation function from level A to B by discharge</td>
<td>85.8</td>
</tr>
<tr>
<td>Transfer Current-level of transfer function on admission</td>
<td>37.6</td>
</tr>
<tr>
<td>Transfer Change-decline in level of transfer function in level A to B by discharge</td>
<td>81.6</td>
</tr>
<tr>
<td>Toileting Current-level of toileting function on admission</td>
<td>64.3</td>
</tr>
<tr>
<td>Toileting Current-level of toileting function on admission</td>
<td>60.6</td>
</tr>
<tr>
<td>Bathing Current-level of bathing function on admission</td>
<td>71.3</td>
</tr>
<tr>
<td>Bathing Change-level of decline in bathing function from level A to B by discharge</td>
<td>69.8</td>
</tr>
<tr>
<td>Eating Prior-level of eating function on admission</td>
<td>52.0</td>
</tr>
</tbody>
</table>
Hypothesis three (H3). H3 was supported. H3-The model predicts the expert referral (yes/no) decision with sensitivity, specificity and positive predictive value of greater than 80%. The Area under the Curve (AUC) was 91.5%, sensitivity was 90.1% (95% CI: 88.1-91.9), specificity was 76.9% (95% CI: 71.0-82.0), positive predictive value was 94.2% (95% CI: 92.5-95.6), negative predictive value 65% (95% CI: 59.2-70.6). While specificity was slightly lower than 80%, increasing specificity resulted in less acceptable sensitivity and positive predictive values and a value of >70% was determined to be highly acceptable in this case, since the risk is to over refer which would not do harm to patients, however, it may have cost implications. The training set had 1251 cases; the validation set 245 cases.

Hypothesis four (H4). H4 was partially supported. H4-The model predicted the type of setting (facility vs home care) with sensitivity, specificity and positive predictive value of greater than 80%. The AUC was 89.7%, sensitivity was 89.2% (95% CI: 84.0-93.2), specificity was 68.0% (95% CI: 53.3-80.5), positive predictive value was 91.6% (95% CI: 86.7-95.1), negative predictive value was 61.8% (95% CI: 47.7-74.6). While specificity was lower than 80%, increasing specificity resulted in less acceptable sensitivity and positive predictive values and a value near 70% was determined to be acceptable in this case; since the risk is to refer to facility care versus home care which would not do harm to patients, however, it may have cost implications. The model was fitted to the validation dataset, using the same beta coefficients and cut-point as found in the validation dataset.

Discussion.

In this study we developed and validated a two-step expert derived algorithm to guide decision making for which patients need a post-acute care referral and which type of setting best matches their needs. Discharge referral decision support algorithms are useful to remove barriers to care imposed by subjectivity and variation in decision-making. We derived our algorithm through expert consensus based on patient need rather than on policy or insurance, which can serve as sources of bias or barriers to appropriate care. The importance of this is shown by Lockery and colleagues who found that patient, family or professional involvement in the DP process did not directly affect discharge placement. Rather DC placement may more accurately reflect insurance status. For those in lower socioeconomic groups with Medicaid that provides less restrictive home health benefits their care is shifted toward home based services, while Medicare policies support institutional care such as SNF or IRF. We purposely did not provide insurance information to our experts and we instructed them to base their decisions solely on need.

More recently, with the advent of the Affordable Care Act with Bundled Payment Programs and Accountable Care Organizations, insurers and providers are seeking the lowest cost site of care rather than the site that may best meet patients’ needs. Such policy shifts reinforce the need for and value of evidence-based decision support that identifies patients based on their clinical characteristics and not on insurance or local conventions. Over time, comparisons of patient outcomes when our algorithm recommends care versus where patients actually go after discharge will demonstrate the impact of such policy shifts on patient outcomes.

Our algorithm may be helpful to discharge planners in assisting with the assessment of patients for post-acute care. A survey of 37 social workers conducting DP in 36 hospitals reported that assessment was their most important and time consuming task with assessment of home support and help with activities of daily living (ADLs) the most demanding and they spent less time on counseling and more time on concrete tasks such as determining services. Automating the assessment of these two concepts with CDS may be of great value in decreasing the work and cognitive load for discharge planners. Wolock and colleagues found discharge planners dealt with psychosocial problems and relationship issues only when they interfered with discharge and large caseloads prevented discharge planners from having enough time to do it all. CDS could lift some of this burden by automating the assessment process, alerting discharge planners of high need patients and recommending levels of care as first steps and a “heads up”, thereby allowing more time for the important counseling interventions that engage patients and caregivers in shared decision-making. Our algorithm operates off of data collected by nurses during routine patient care at admission and throughout the stay, making efficient use of information and negating the need to spend time collecting it again.

The factors deemed important within the algorithm reflect need for post-acute support and those associated with risk for readmission and other poor discharge outcomes. The algorithm cumulatively brings these factors together to determine the patients’ need for post-acute care and suggest whether a facility level (SNF, IPF, NH) or home health
care is most likely to meet their needs. The algorithm heavily weighs the performance of ADLs as critical factors. Multiple studies reinforce the importance of ADL function demonstrating the association between ADL limitations, discharge placement and discharge outcomes.\(^{33,34}\)

In our study, experiencing a decline in ambulation, transferring, and bathing was associated with referral to facility-based care versus home care. Similarly, Mason and colleagues found that the need for skilled nursing care or occupational therapy versus custodial care for ADLs differentiated patients discharged with home care versus facility level care respectively.\(^{35}\) Similar to our model, an algorithm that predicts SNF or inpatient rehabilitation versus discharge to home for colorectal cancer surgery patients includes number of comorbidities and number of previous hospital stays as predictors.\(^{36}\) Unique to our model, having a caregiver was important to the referral decision and having a caregiver who was available 24 hours a day/ seven days per week was important regarding the level of care, facility versus home based.

Our previous algorithm, the D2S2 identifies patients who need post-acute care.\(^{10}\) Translation to practice demonstrated significant declines in readmissions when using the D2S2 compared to not, after controlling for covariates and time.\(^{37,38}\) This study built on that work with a larger, more diverse sample and robust set of data elements and created a two-step algorithm to determine the level of care for those identified for referral. The data elements are available shortly after admission and require no additional data collection. We are testing the new algorithm in a quasi-experimental, pre/post design and comparing patient outcomes using propensity modeling in two hospitals.

Study limitations include a sample limited to 1,496 patients that contains only the most common diagnosis for hospital admissions. We also could not account for every patient factor that may be important to these decisions. For example, the patient and caregiver preferences were not included. Also, the experts were unable to “see” the patient but made their decisions based on a written case study and they made their decisions independently whereas in real life they often collaborate as a team.

**Conclusion.**

Using electronic health record data and expert opinion, our team successfully built and validated a two-step prediction algorithm to support discharge referral decision-making. We achieved highly satisfactory predictive summary statistics on both steps of the algorithm and the factors within it match those commonly identified as associated with post-discharge service need and outcomes. Evidence-based CDS for discharge planning may alleviate some of the workload for discharge planners enabling them to spend more time counseling patients and engaging in shared decision-making. The algorithm was embedded into the EHR of a hospital and tested in a quasi-experimental pre/post design. Next steps include analysis of the experimental data to determine whether the algorithm has an impact on referral rates, site of referral, types of patients referred, and most importantly patient outcomes such as readmission and emergency department use.

**Acknowledgement.**

Research reported in this publication was supported by the National Institute of Nursing Research of the National Institutes of Health under Award Number R01NR007674. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

**References**


Creating Clinical Fuzzy Automata with Fuzzy Arden Syntax

Jeroen S. de Bruin, PhD, MSc¹,², Heinz Steltzer, PhD, MD³, Andrea Rappelsberger, MSc¹, Klaus-Peter Adlassnig, PhD, MSc, FACMI, MIAHSI¹,²

¹Section for Artificial Intelligence and Decision Support, Center for Medical Statistics, Informatics, and Intelligent Systems, Medical University of Vienna, Vienna, Austria; ²Medexter Healthcare, Vienna, Austria; ³Trauma Hospital Vienna South, Vienna, Austria

Abstract

Formal constructs for fuzzy sets and fuzzy logic are incorporated into Arden Syntax version 2.9 (Fuzzy Arden Syntax). With fuzzy sets, the relationships between measured or observed data and linguistic terms are expressed as degrees of compatibility that model the unsharpness of the boundaries of linguistic terms. Propositional uncertainty due to incomplete knowledge of relationships between clinical linguistic concepts is modeled with fuzzy logic. Fuzzy Arden Syntax also supports the construction of fuzzy state monitors. The latter are defined as monitors that employ fuzzy automata to observe gradual transitions between different stages of disease. As a use case, we re-implemented FuzzyARDS, a previously published clinical monitoring system for patients suffering from acute respiratory distress syndrome (ARDS). Using the re-implementation as an example, we show how key concepts of fuzzy automata, i.e., fuzzy states and parallel fuzzy state transitions, can be implemented in Fuzzy Arden Syntax. The results showed that fuzzy state monitors can be implemented in a straightforward manner.

Introduction

By definition an automaton is “an abstract state-determined machine designed to follow automatically a predetermined sequence of operations or respond to encoded instructions”. Automata are prime examples of general systems over discrete spaces [1]. The concepts of state and transition are core aspects of automata theory. The best known class of automata are deterministic finite-state automata, in which systems can be in exactly one of a finite number of states at any given time. These states are labeled using natural language concepts and/or classifications. Transitions between system states (i.e., linguistic concepts) are triggered by an external input, and may be presented as a function mapping the current state and the input into the next state [1].

Although the concept of a discrete state is attractive due to its simplicity in terms of implementation, it is of limited use for modeling situations and contexts where discrete states cannot be accurately defined using natural language concepts, or where knowledge about the situation is imprecise or incomplete. For instance, an individual’s health cannot be accurately classified by discrete, mutually exclusive concepts at the sole end of the spectrum, such as healthy or sick, nor would it be practical to introduce a myriad of states between these spectrum poles. To address these shortcomings in classification with linguistic concepts, Zadeh introduced fuzzy sets and fuzzy logic [2]. With fuzzy sets, the relationship between linguistic terms and measured or observed data is expressed as a degree of compatibility (DoC) calculated by fuzzy membership functions rather than a discrete, dichotomous classification. In this context, a DoC formally models the unsharpness of the boundaries of linguistic terms. Propositional uncertainty due to incomplete knowledge of relationships between clinical linguistic concepts is modeled with fuzzy logic, allowing for approximate reasoning instead of exact rule inference.

The incorporation of fuzzy sets and fuzzy logic into automata gives rise to the notion of fuzzy automata, which are able to handle uncertainty and continuous space [3,4]. With fuzzy automata, a system can be in more than one state at the same time, whereby the applicability (or membership) of each state is expressed as a DoC. Because of this multi-state membership, multiple simultaneous transitions of state are also possible. The system as a whole is far more expressive because the potentially endless combinations of state values (i.e., applicability of linguistic concepts) can imply richer, more abstract linguistic concepts. Furthermore, multiple (partial) parallel transitions cause transitions between linguistic concepts to be more gradual and thus more intuitive.

The properties of fuzzy automata discussed above, they are well suited for the use in (automated) clinical monitors [5]. In clinical monitoring, streams of patient data are measured in (occasionally brief) time intervals and presented to the clinician. Given the increasing body of patient data being measured in various time intervals, the monitor is growing more complex, both in terms of the number of presented data elements as well as the interpretation of combinations of different data elements. In other words, the more powerful a monitor becomes (from a functional perspective), the greater is the risk of errors of omission in the interpretation of data. By using a fuzzy automaton in
clinical monitoring for the interpretation and aggregation of measured patient data over time, the dimensionality of input data can be reduced and presented in semantically meaningful, clinically relevant linguistic concepts over fixed time intervals, along with a gradual indication of their applicability to the patient.

Standardized technical communication with hospital information systems and electronic health records is an equally important element of meaningful semantic communication with users. The acceptance and dissemination of the aforementioned monitoring systems could be improved by employing known communication standards, especially when the systems are to be embedded into existing clinical monitors. A widely used standard for computerized knowledge representation and processing is Arden Syntax [6], which is a programming language for the collection, description, and processing of medical knowledge in a machine-executable format. Since Arden Syntax version 2.9 was augmented by formal constructs based on fuzzy set theory and fuzzy logic, it is now possible to implement a fuzzy automaton in a clinical system using a standard that intrinsically supports fuzzy methods.

In this paper, we report our preliminary experience concerning the implementation of a clinical fuzzy automaton using Arden Syntax 2.9 (Fuzzy Arden Syntax). As a use case we re-implemented a previously published monitoring system named FuzzyARDS [7], which is a clinical monitoring system for patients suffering from acute respiratory distress syndrome (ARDS) and is based on the DiaMon-1 framework [8]. For easier comprehension, we provide an overview of Arden Syntax and the fuzzy methods incorporated in Fuzzy Arden Syntax, as well as a short description of the DiaMon-1 framework and the FuzzyARDS monitor. Then, using the re-implementation of parts of the FuzzyARDS monitor (referred to henceforth as FuzzyArdenARDS) as an example, we show how key concepts of fuzzy automata, such as fuzzy states and parallel fuzzy state transitions can be implemented in Fuzzy Arden Syntax.

Methods

Arden Syntax

Arden Syntax is a standard for computerized medical knowledge representation and processing. Arden Syntax knowledge bases are fragmented into medical logic modules (MLMs), which are collections of medical rules and knowledge for the purpose of making at least one medical decision [9]. Several properties make Arden Syntax well suited for the computerized representation of medical knowledge [10]. In fact, the program code in Arden Syntax resembles natural language. Thus MLMs are understood more easily by healthcare professionals. Moreover, pure medical knowledge is separated from more technical processing, which improves code transparency. Finally Arden Syntax supports various data types specific to medical documentation, such as time and duration types.

A complete overview of Arden Syntax is beyond the scope of this paper. However, for comprehension of the examples presented in this paper we will provide a short overview of Fuzzy Arden Syntax. For a complete description we refer to the Arden Syntax version 2.9 specification [6].

Each Arden Syntax MLM is hierarchically structured. At the top level an MLM is divided into the categories of maintenance, library, knowledge, and resources. Each of these categories contains category-specific slots. Slots in the maintenance category enable the MLM author to provide metadata on the MLM, such as the MLM title, author, or version. Slots in the library category provide contextual information about the MLM, such as the purpose of the MLM, an explanation of its functionality, and evidence-based resource citations. The medical knowledge of the MLM is implemented in the knowledge category. In the data slot, MLM parameters can be assigned to variables, and data from external sources can be obtained through curly braces expressions. The MLM’s program logic is implemented in the logic slot. Apart from implementing logic, processing can also be deferred through the invocation of other MLMs. Execution ends with a concluding statement which, if considered true, results in the execution of the action slot. Finally, localized messages can be constructed through the optional definition of the resources category.

Fuzzy Arden Syntax

Since version 2.9, the Arden Syntax supports formal constructs based on fuzzy set theory and fuzzy logic. We will present a selection of fuzzy extensions implemented in Fuzzy Arden Syntax in this section. Note that this is not a complete overview of all fuzzy concepts implemented in Fuzzy Arden Syntax; for a detailed description of fuzzy constructs implemented in Fuzzy Arden Syntax we refer to previously published work [11].

The truth value model of Arden Syntax was expanded in order to extend traditional methods of calculation and logic. Prior to version 2.9, the truth value model was dichotomous, supporting only the values “true” and “false”. However, truth values in Fuzzy Arden Syntax may now constitute any value in the range [0, 1]. True and false still exist in this model as the extremes of the range: “false” has a truth value of 0, whereas “true” has a truth value of 1. Based on this
truth value model, Fuzzy Arden Syntax incorporates fuzzy set data types, built-in propositional fuzzy logic operators, and degrees of applicability of conditional branches.

The fuzzy set data type can be used to model and quantify the unsharpness of boundaries in definitions of linguistic concepts. A fuzzy set is declared with two or more value tuples which define the fuzzy region(s). From these tuples, a linear membership function is constructed for the fuzzy set, which is used to calculate the DoC of measured or observed data with respect to the clinical linguistic concept under consideration.

With fuzzy logic operators, propositional uncertainty in relationships between linguistic clinical concepts can be modeled implicitly. Three basic propositional fuzzy logic operations are implemented in Fuzzy Arden Syntax – negation, conjunction, and disjunction. These are equipped to handle all truth values in the extended truth value model. By default, negation of a concept is implemented as 1 minus the truth value of that concept. The standard minimum function is used as the fuzzy conjunction operator, and the standard maximum function is used as the fuzzy disjunction operator.

The new truth value model called for an extension of the evaluation mechanism for conditional branches. Whereas only one conditional branch was executed at a time in the traditional model, in the extended model each branch whose condition amounts to non-zero is executed in parallel. To this end, all simple data types have a degree of applicability, which is set to 1 by default. However, when the program execution splits into multiple parallel branches, each branch is provided with its own set of duplicated variables, and each variable is assigned a degree of applicability which is equal to the relative truth value of the respective branch’s conditional expression. After execution of all conditional branches, variable values can either be aggregated or not. When the branches are aggregated (through the aggregate keyword at the end of the conditional block), duplicated variables are joined using a weighted average (applicability multiplied by the variable value). If not aggregated, the MLM will conclude with multiple return values, each with its own applicability.

As an example, consider the Fuzzy Arden Syntax MLM below in which we use fuzzy sets to classify the severity of ARDS. For the sake of brevity, we have confined the example to the knowledge category.

```plaintext
1   maintenance:   […]
2   library:  […]
3 knowledge:
4   type:   data_driven;;
5   data:    (pao2, fio2) := argument;; // blood gas and inspiration
6   priority:  ;
7   evoke: ;;
8   logic:
9 // Fuzzy set definitions
10  ARDS_severe := fuzzy set (100,1), (110,0);
11  ARDS_moderate := fuzzy set (100,0), (110,1), (190,1), (200,0);
12  ARDS_mild := fuzzy set (190,0), (200,1), (300,1), (310,0);
13
14 // Parameter analysis
15  if ((pao2 / fio2) is in ARDS_severe) then
16    msg := "Patient suffers from severe ARDS."
17  elseif ((pao2 / fio2) is in ARDS_moderate) then
18    msg := "Patient suffers from moderate ARDS."
19  elseif ((pao2 / fio2) is in ARDS_mild) then
20    msg := "Patient suffers from mild ARDS."
21  endif;
22
23 // Program conclusion
24  conclude true;;
25  action: return msg;;
26  urgency: ;;
27 end:
```
The logic in this MLM is based on the most recent ‘Berlin Definition’ of ARDS published in 2012 [12]. According to this definition, the severity of ARDS is determined by the ratio of partial arterial oxygen pressure (PaO2) and the fraction of inspired oxygen (FiO2). Based on the resulting outcome, ARDS is characterized as either mild (200 < PaO2/FiO2 ≤ 300), moderate (100 < PaO2/FiO2 ≤ 200), or severe (PaO2/FiO2 ≤ 100). However, patients with measured values close to these thresholds are also of interest. As such, we created fuzzy sets for all characterizations that extend beyond the defined thresholds (lines 10-12). Next, the PaO2/FiO2 ratio is compared with each fuzzy set, and the resulting truth values serve as conditions in conditional branches (lines 15-21). In case multiple truth values are non-zero, each of these conditional branches is executed. Furthermore, as the branches are not aggregated, this would cause the MLM to return multiple copies of msg, each with its own degree of applicability. For example, if PaO2/FiO2 were 106, this would cause the “severe ARDS” message to be returned with an applicability of 0.4 and the “moderate ARDS” message to be returned with an applicability of 0.6. This could be interpreted as a patient’s ARDS being between the “moderate” and “severe” state.

**FuzzyArdenARDS and DiaMon-1**

The FuzzyArdenARDS application was re-implemented based on the original FuzzyARDS automaton, which was constructed using the DiaMon-1 framework as discussed in [7,8]. This formal framework was developed in order to design monitors capable of abstracting continuously supplied, objectively observed raw data into aggregated, qualitative, linguistic concepts, such as stages of disease. Furthermore, monitors developed with this framework provide early indication of improvement in, or deterioration of, a patient’s health status because the monitors include smooth transitions between stages.

Applications in the DiaMon-1 framework are referred to as state monitors, which employ fuzzy automata to observe gradual transitions between different stages of disease. In this context, a state represents a (linguistic) representation of a patient’s health status or a specific stage of disease. Transitions provide possible pathways between states or stages, which are triggered by inputs such as time or measured data.

A fuzzy state monitor \( SM \) is formally defined as a 6-tuple \( \overline{SM} = (Q, \overline{q}_0, X, \delta, P, f) \). In this definition, the first four parameters jointly constitute the underlying fuzzy automaton: \( Q \) denotes a finite set of states, \( \overline{q}_0 \) is a (potentially) fuzzy subset of \( Q \) that marks the initial state, \( X \) is a finite set of input symbols, while \( \delta: Q \times X \to Q \) is a transition function that maps states and inputs onto states. Furthermore, \( P \) is the parameter value space over all observed parameters: \( p_1 \times \ldots \times p_n \), and \( f \) is a mapping function that maps parameter tuples from \( P \) to a fuzzy subset of \( X \).

Calculation of the monitor state at time point \( t (\overline{q}_t) \) proceeds through an inductive function based on \( \overline{q}_{t-1}, \delta, \) and \( f \). Suppose that at time point \( t, f \) yields fuzzy subset \( \overline{fs}_t \), which is a collection of truth values for all \( x \in X \). Then, using the extension principle [13], the state for each \( q \in Q \) at time point \( t, \overline{q}_t(q) \), can be calculated as follows:

\[
\overline{q}_t(q) = \left\{ \begin{array}{ll}
\lor \{ \overline{q}_{t-1}(q') \land \overline{fs}_t(x) | \delta(q', x) = q, q' \in Q, x \in X \} & \text{if } \delta^{-1}(q) \neq \emptyset \\
0 & \text{if } \delta^{-1}(q) = \emptyset
\end{array} \right.
\]

For the FuzzyArdenARDS application, the automaton states \( Q \) are shown in Table 1.

**Table 1.** Definition of the fuzzy automaton states in FuzzyArdenARDS.

<table>
<thead>
<tr>
<th>State</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Start</td>
<td>Initial state</td>
</tr>
<tr>
<td>Normal</td>
<td>Oxygenation is satisfactory, no additional effort needed.</td>
</tr>
<tr>
<td>Hypoxic</td>
<td>Oxygenation is too low.</td>
</tr>
<tr>
<td>Responding to high FiO2</td>
<td>Oxygenation was positively affected by high FiO2.</td>
</tr>
<tr>
<td>Not responding to high FiO2</td>
<td>High FiO2 did not have a desired effect.</td>
</tr>
<tr>
<td>Improved after hand bagging</td>
<td>Manual oxygenation through hand bagging has improved oxygenation.</td>
</tr>
<tr>
<td>Not improved after hand bagging</td>
<td>Hand bagging did not have the desired effect.</td>
</tr>
</tbody>
</table>

Note: FiO2, fraction of inspired oxygen.
As initial state $q_0$, the truth value for the Start state is 1, and 0 for all others. The set of input symbols $X$ comprises \{adequate oxygenation, hypoxemia, high FiO2, low FiO2, rapidly improving oxygenation, slowly decreasing oxygenation\}. The set of fuzzy state transition rules $\delta$ is shown in Table 2.

**Table 2.** Definition of the fuzzy state transition rules in FuzzyArdenARDS.

<table>
<thead>
<tr>
<th>Rule</th>
<th>Begin state</th>
<th>Transition condition</th>
<th>End state</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Start</td>
<td>Adequate oxygenation</td>
<td>Normal</td>
</tr>
<tr>
<td>2</td>
<td>Start</td>
<td>Hypoxemia</td>
<td>Hypoxic</td>
</tr>
<tr>
<td>3</td>
<td>Normal</td>
<td>Hypoxemia</td>
<td>Hypoxic</td>
</tr>
<tr>
<td>4</td>
<td>Hypoxic</td>
<td>Low FiO2 $\land$ Adequate oxygenation</td>
<td>Normal</td>
</tr>
<tr>
<td>5</td>
<td>Hypoxic</td>
<td>High FiO2 $\land$ rapidly improving oxygenation</td>
<td>Responding to high FiO2</td>
</tr>
<tr>
<td>6</td>
<td>Hypoxic</td>
<td>High FiO2 $\land$ hypoxemia</td>
<td>Not responding to high FiO2</td>
</tr>
<tr>
<td>7</td>
<td>Responding to high FiO2</td>
<td>Low FiO2 $\land$ slowly decreasing oxygenation</td>
<td>Improved after hand bagging</td>
</tr>
<tr>
<td>8</td>
<td>Responding to high FiO2</td>
<td>Low FiO2 $\land$ hypoxemia</td>
<td>Not improved after hand bagging</td>
</tr>
<tr>
<td>9</td>
<td>Not responding to high FiO2</td>
<td>Low FiO2 $\land$ hypoxemia</td>
<td>Hypoxic</td>
</tr>
<tr>
<td>10</td>
<td>Not responding to high FiO2</td>
<td>High FiO2 $\land$ adequate oxygenation</td>
<td>Responding to high FiO2</td>
</tr>
<tr>
<td>11</td>
<td>Improved after hand bagging</td>
<td>Adequate oxygenation</td>
<td>Normal</td>
</tr>
<tr>
<td>12</td>
<td>Improved after hand bagging</td>
<td>Hypoxemia</td>
<td>Hypoxic</td>
</tr>
<tr>
<td>13</td>
<td>Not improved after hand bagging</td>
<td>Hypoxemia</td>
<td>Hypoxic</td>
</tr>
</tbody>
</table>

Note: FiO2, fraction of inspired oxygen; $\land$, fuzzy logical conjunction operator (minimum function).

Three parameters have been considered for $P$: time, oxygen saturation $SaO2$ (as a noninvasive alternative to measuring $PaO2$), and FiO2. Finally, rules for fuzzy sets based on time, $SaO2$ and FiO2, which jointly constitute $f$ are presented in Table 3.
Table 3. Definition of the fuzzy automaton parameter-to-input symbol mapping in FuzzyArdenARDS.

<table>
<thead>
<tr>
<th>Input symbol</th>
<th>Rule (Start of fuzzy region)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adequate oxygenation</td>
<td>SaO2 above 97% (93%) for 5 minutes</td>
</tr>
<tr>
<td>Hypoxemia</td>
<td>SaO2 between 90% (87%) and 93% (97%) for 2 minutes</td>
</tr>
<tr>
<td>High FiO2</td>
<td>FiO2 above 60% for 30 seconds</td>
</tr>
<tr>
<td>Low FiO2</td>
<td>FiO2 below 60% for 30 seconds</td>
</tr>
<tr>
<td>Rapidly improving oxygenation</td>
<td>SaO2 increasing from 87–95% (85–99%) to 97–100% (93–100%) within 30–90 seconds</td>
</tr>
<tr>
<td>Slowly decreasing oxygenation</td>
<td>SaO2 above 96% (91%) steady or decreasing to 94% (89%) within 25 minutes</td>
</tr>
</tbody>
</table>

Note: SaO2, oxygen saturation; FiO2, fraction of inspired oxygen.

Data processing
The MLMs discussed in this report were created with the ARDENSUITE software [14,15]. The ARDENSUITE is a framework for medical knowledge representation and reasoning, which comprises an integrated development and test environment (IDE) and an ARDENSUITE server, including software modules for interconnecting with data sources (Figure 1).

Figure 1. Graphic depiction of the ARDENSUITE framework. Image adapted from [15].

With the ARDENSUITE IDE, users can write and compile MLMs via the authoring tool. If test data are available, users can also immediately test the implemented MLMs. After compilation, the MLMs are uploaded to the administration module of the ARDENSUITE server. The administration module is a management tool for compiled Arden Syntax projects and supports functionalities such as the activation or deactivation of MLMs in an application, or MLM version management. The core element of the ARDENSUITE server is the ARDENSUITE engine, which executes compiled MLMs. To facilitate access to MLM functionalities by arbitrary clients, the server provides service-oriented access through a web-service component. Using this component, MLM calls and data exchange are facilitated through the Simple Object Access Protocol (SOAP) and Representational State Transfer (REST) web-service standards. Web-service standards can also be used to connect the ARDENSUITE server with SOAP/REST-compatible external database sources through the ARDENSUITE server connector. With this module, external data sources can be accessed directly from within MLM files using query languages (such as SQL), which are then forwarded to the source data base management system using SOAP or REST web services.
Results

In this section, we discuss parts of the resulting MLM implementation of FuzzyArdenARDS. We implemented the core of the FuzzyArdenARDS application in two MLMs. In the first MLM, fuzzy sets are defined and truth values calculated using the SaO2 and FiO2 parameters. Each parameter is supplied as a list of values over the last half hour. The MLM is designed for semi-real-time processing, and is called every 30 seconds, thus analyzing 30 seconds of data at a time. Again, for the sake of brevity, we have confined the MLMs listed below to the knowledge category.

```plaintext
1 maintenance: [...]
2 library: [...]
3 knowledge:
4 type: data_driven;;
5 data: (sao2, fio2):= argument;;
6 priority: ;;
7 evoke: ;;
8 logic:
9 // Fuzzy set definitions
10 fs_adeq_oxy := fuzzy set (0.93,0), (0.97,1);
11 fs_hypoxemia := fuzzy set (0.87,0), (0.9,1), (0.93,1), (0.97,0);
12 fs_rio_begin := fuzzy set (0.85,0), (0.87,1), (0.95,1), (0.99,0);
13 fs_rio_end := fuzzy set (0.93,0), (0.97,1);
14 fs_sdo_begin := fuzzy set (0.91,0), (0.96,1);
15 fs_sdo_end := fuzzy set (0.89,0), (0.94,1);
16
17 // Parameter analysis
18 sao2_5mins := sao2 where they occurred within the past 330 seconds;
19 ade_oxy := minimum (sao2_5mins in fs_adeq_oxy);
20
21 sao2_2mins := sao2 where they occurred within the past 150 seconds;
22 hypoxemia := minimum (sao2_2mins in fs_hypoxemia);
23
24 fio2_30secs := fio2 where they occurred within the past 60 seconds;
25 high_fio2 := minimum (fio2_30sec) > 60;
26 low_fio2 := maximum (fio2_30sec) < 60;
27
28 sao2_30secs := sao2 where they occurred within the past 30 seconds;
29 for sao2_element in sao2_30secs do
30   reference_list := sao2 where ((time of sao2_element) - (time of them)) is within 30 seconds to 90 seconds;
31   for sao2_ref_element in reference_list do
32     lmax_tv := sao2_ref_element is in fs_rio_begin and sao2_element is in fs_rio_end;
33     rio := maximum (rio, lmax_tv);
34   enddo;
35 enddo;
36
37 sdo := 0;
38 for sao2_element in sao2_30secs do
39   reference_list := sao2 where ((time of sao2_element) - (time of them)) is not greater than 25 minutes;
40   for sao2_ref_element in reference_list do
41     lmax_tv := sao2_ref_element is in fs_sdo_begin and sao2_element is in fs_sdo_end;
42     sdo := maximum (sdo, lmax_tv);
43   enddo;
44 enddo;
45 conclude true;;
```
In the above MLM code snippet, truth values are calculated for each of the symbols in \( X \). First, all fuzzy sets are defined (lines 10-15). Then the truth value is calculated for the symbol *adequate oxygenation*. For this purpose, we first need to obtain SaO2 values for the last 5½ minutes; 30 seconds for evaluation and 5 minutes thereafter to calculate the truth value for all data in the evaluation phase according to the mapping in Table 3 (line 18). We then apply the data to the defined fuzzy set and take the minimum truth value as an aggregate result for the last 30 seconds (line 19). In a similar fashion, truth values are calculated for the symbols *hypoxemia* (lines 21-22), *high FiO2*, and *low FiO2* (lines 24-26). For the symbol *rapidly improving oxygenation*, a slightly different approach is needed. First, data elements for the first 30 seconds for evaluation are isolated (line 28). For each of the data elements, we then select prior elements or data elements within a period of 30-90 seconds before the timestamp of the analyzed elements (line 29-31). Data elements of both lists are applied pairwise to their respective fuzzy sets, and a logical conjunction of each pair is calculated (line 32-33). Finally, the maximum of all pairwise logical conjunctions is chosen as an aggregate value for the symbol *rapidly improving oxygenation* (line 34). In a similar fashion, a truth value is calculated for the symbol *slowly decreasing oxygenation* with a time period of 25 minutes.

The second MLM is used to perform the transitions. As parameters, a list of truth values for each state and a list of truth values for each input symbol are provided. Due to space constraints, we only show that part of the MLM implementation dealing with the transitions for the end states *Start*, *Normal*, and *Hypoxic*.

```plaintext
1  maintenance:   […]
2  library:    […]
3  knowledge:
4    type:   data_driven;;
5  data:
6    // States (array indices) enumerations:
7  AutomStates := OBJECT [Start, Normal, Hypoxic, RespHighFiO2,
8                      NotRespHighFiO2, ImpAfterHandBagging, NotImpAfterHandBagging];
9  states := new AutomStates with 1 seqto 7;
10
11    // Inputs (array indices) enumerations:
12  AutomInputs := OBJECT [AdequateOxy, Hypoxemia, HighFiO2,
13                      LowFiO2, RapidImpOxygenation, SlowDecOxygenation];
14  inputs := new AutomInputs with 1 seqto 6;
15
16  (state_tvs, input_tvs) := Argument;;
17
18  logic:
19    // Start state has no incoming transitions
20  state_start := 0;
21
22    // Normal state has three incoming transitions:
23  state_normal :=
24      // Rule 1: Start and adequate oxygenation
25      (state_tvs[states.Start] as truth value and
26      input_tvs[inputs.AdequateOxy] as truth value)
27      or
28      // Rule 4: Hypoxic and low FiO2 and adequate oxygenation
29      (state_tvs[states.Hypoxic] as truth value and
30      input_tvs[inputs.LowFiO2] as truth value and
31      input_tvs[inputs.AdequateOxy] as truth value)
32      or
33      // Rule 11: Improved after hand bagging and adequate oxygenation
34      (state_tvs[states.ImpAfterHandBagging] as truth value and
35      input_tvs[inputs.AdequateOxy] as truth value);
36
37    // Hypoxic state has five incoming transitions:
```
Truth values for the seven states in $Q$ are calculated in the MLM shown above. First, to improve MLM readability we constructed objects that enumerate the indices of the state and input value lists supplied through the argument (lines 6-16). The calculation of truth values for each state starts thereafter. By definition, the $Start$ state is only true at the first iteration of the automaton, followed by 0, as it has no incoming transitions. For each other state, the truth value is determined by a logical disjunction over all rules in Table 2 that have the respective state as an end state. The truth value for each rule, on the other hand, is calculated by a logical conjunction over the truth values of the state and transition conditions. As such, for the $Normal$ state a logical disjunction is calculated over transition rule 1 (lines 25-26), rule 4 (lines 29-31), and rule 11 (lines 34-35). The truth values for the other states are calculated in a similar fashion.

**Discussion**

In the present report, we showed how fuzzy state monitors as defined in [8] could be implemented using fuzzy methods supported by Fuzzy Arden Syntax, a standard for computerized knowledge representation and processing. When designing a clinical knowledge base, knowledge engineers work closely together with clinicians to construct the rules. However, the translation process from natural language to a computerized knowledge representation may be prone to error. The average clinician may be unable to, or not interested in, the validation of source code. The original FuzzyARDS program was written in a dialect of the object-oriented language Smalltalk, which is neither considered a mainstream medium nor is easily understood by those untrained in its application. Given that Arden Syntax rules closely resemble natural language, clinicians can verify the implemented knowledge in MLMs (more or less) easily without in-depth knowledge of modern programming languages.

Within the context of fuzzy state monitors, we introduced fuzzy sets as mapping functions that map raw data to clinical linguistic concepts, thereby yielding a degree of compatibility between 0 and 1. Based on these degrees, together with degrees of applicability for each state in a fuzzy automaton, we showed that multiple transitions could occur simultaneously, resulting in a new automaton configuration comprising (again) degrees of applicability for each state in the fuzzy automaton.

The interpretation of the automaton results is an important aspect. It was not discussed here because it would exceed the scope of this report and is dependent on individual applications. In the case of FuzzyArdenARDS, six parameters need to be interpreted in a pairwise manner: oxygenation state (normal vs. hypoxic), response to high FiO2 (response vs. no response), and response to hand bagging (improvement vs. no improvement). Given that values for each of
these states can be a DoC, a richer, a gradual interpretation follows from these pairwise interpretations. For instance, when the “normal” state has a DoC of 0.3 and the “hypoxic” state a DoC of 0.7, the patient’s current oxygenation state may be interpreted as moderately hypoxic.

The limitations of the present report are worthy of note. First, the study is limited to a single application, namely FuzzyArdenARDS. Other fuzzy state monitors and even other types of fuzzy automata need to be implemented with Fuzzy Arden Syntax to ensure the syntax is equipped to support a variety of fuzzy applications and methodologies. Second, the fuzzy sets that we implemented for the FuzzyArdenARDS automaton were “two-dimensional”. In other words, they only implemented fuzzy regions for one parameter. As some rules were defined over two parameters (SaO2 and a time duration), three-dimensional fuzzy sets that provide more accurate modeling for these rules will have to be implemented in the future. Finally, as of yet the program has only been tested on retrospectively collected data.

We reported on the first steps in implementing fuzzy state monitors with Fuzzy Arden Syntax. In the future, we plan to study and address the aforementioned limitations and continue to improve the use of Arden Syntax for real-time monitoring.

Conclusion

The native support of fuzzy methods allows the intuitive implementation of a fuzzy state monitor for clinical application with Fuzzy Arden Syntax.

References

Abstract

The increased adoption of Electronic Health Record (EHR) systems offers new opportunities for clinical research. The Health Insurance Portability and Accountability Act (HIPAA) mandates that medical records need to be stripped of personal identifiers in order to be shared. One particular challenge is how to handle free-text medical records. While many methods have been developed, there is a dearth of software tools than can be easily used by practitioners. We present deidentify, a new de-identification tool, which comes with a graphical user interface and runs on all operating systems. Evaluating its algorithm on a gold-standard corpus of nursing notes, we demonstrate its adequate performance with a recall of 0.919 and a precision of 0.645. Since deidentify comes with a pre-trained model, it can be used when no training data is available, but can also be manually configured. This way, it should be convenient to use for de-identification tasks.

1 Introduction

In recent years, we have witnessed a major transformation in the health care sector with the move to a wide array of clinical information systems such as Electronic Health Record (EHR) systems and Computerized Provider Order Entry (CPOE) applications. Hospital adoption of these systems has increased more than five-fold since 2008, with now more than nine in ten hospitals had adopted at least a basic EHR system.1

For a hospital or physician’s practice, implementing an EHR system often coincides with a move from paper to electronic documentation of progress notes and other documents created by the clinicians. There are many perceived benefits in the use of electronic health record systems. Menachemi & Collum identify three main types: benefits for clinical, organizational and societal outcomes.2 They present findings which demonstrate that EHRs can reduce clinical errors and drive down costs. Societal benefits arise because the electronic availability of patient data enables public health researchers to more easily conduct studies. In addition, most of this research can be done using much larger sample sizes than commonly used in the medical literature. This could lead to more significant and robust findings.

However, there is a growing concern in the medical community that the focus EHR systems place on structured data entry – while being efficient in saving costs and permitting subsequent analyses – does not enable doctors to effectively express clinically relevant information. Resnik concludes that “important clinical information, detail, and nuance would fail to be captured by an EHR standards discrete fields, with potentially serious consequences for the patient”. Instead, he recommends Natural Language Processing (NLP) techniques to extract meaningful information from free-text medical records.3

One main obstacle in working with free-text medical data is the need to de-identify or anonymize all identifiers which could be used to link a medical record to a certain individual. The Privacy Rule of the Health Insurance Portability and Accountability Act (HIPAA) restricts the distribution of all medical data containing protected health information (PHI), thereby limiting the ability of health care organizations to freely share their data with researchers.

HIPAA permits two methods for the de-identification of PHIs: the “Safe Harbor” rule lists 18 identifiers, which have to be removed in order for data to be counted as de-identified. Alternatively, an expert can testify that the employed statistical or scientific method provides only a small risk of allowing identification of individuals that are part of the study.

When working with free-text, the “Safe Harbor” rule provides a fairly straightforward approach for de-identification: Removing all identifiers that fall in one of the 18 categories is said to provide the necessary anonymity and suffices for health records to satisfy the HIPAA data sharing regulations and allows the data to be used for scientific studies.

In the next section, we review several approaches proposed in the literature to de-identify free-text medical records in compliance with the “Safe Harbor” rule. There has been a large body of research in this area, yet in our experience,
there is a lack of practical tools, which allow researchers and clinical practitioners to de-identify patient data out of the box and without large technical expertise. Many scientific papers propose new methodology without actual implementations, others provide accompanying software with many pre-requisites that have to be installed on the user’s operating system. Many were written a long time ago and appear to be unsupported by now.

In this paper, we present a new tool for de-identification of free-text medical data, which addresses the aforementioned concerns: The goal has been to develop a software which is easy to install, supported on all major operating systems (Windows, MacOS, and Linux) and provides a graphical user interface without requiring the user to be an expert in the domain of NLP or medical informatics.

Using cutting-edge web technologies, our deidentify tool is actively maintained, delivers performance comparable with state-of-the-art methods and is distributed under an open-source license. As a native app built on top of Chromium and the Node.js JavaScript run-time, deidentify allows for easy integration of third-party modules given its modularized nature. This should position it as a tool which can be adapted to changing regulatory frameworks and which can accommodate and integrate top-notch NLP techniques.

The paper is organized as follows: In the next section, we review the relevant literature and trace the development of methods used for de-identifying PHIs. In Section 3, we present the methodology used by our deidentify software and show how it scrubs records of all PHIs and removes or replaces them with made-up values. To assess the quality of our de-identification results, Section 4 describes the evaluation scheme used to test our method, the results of which are presented in Section 5. In Section 6, we reflect on the results and compare them to those of existing approaches advocated in the literature. Section 7 concludes the paper by providing an outlook for future research and an assessment of the strengths and weaknesses of our approach.

2 Related Work

Meystre et al. provide a recent overview on the literature of de-identification systems. They identify two broad categories: First, approaches which are predominantly based on pattern matching and/or deterministic rules. Second, de-identification applications mostly based on machine learning methods.

Historically speaking, most of the earlier research in the field can be assigned to the former category. The Scrub system developed by Sweeney deserves special attention. Scrub is one of the few systems which cannot only be used to replace identifiers by placeholders, but instead can replace them with automatically created surrogate data. The Scrub system ensures that there is a one-to-one mapping of PHI data to their replacements, which preserves the readability of the texts and their internal consistency. Unfortunately, the Scrub system is not available for download anymore, a feature it shares with quite a few of the other developed methods.

Of the more recent machine learning approaches, Meystre et al. identify several methods commonly used in the literature, such as Conditional Random Fields (CRF), Decision Trees, Maximum Entropy models and Support Vector Machines (SVM). For example, Aramaki et al. performed above average with a CRF model in the de-identification challenge organized by the i2b2 (Informatics for Integrating Biology and the Bedside) initiative.

The field has been moving towards usage of machine learning methods for some time now, because state-of-the-art methods outperform solely rule-based pattern matching approaches. At least for the more complicated PHIs, their performance differs by a substantial margin. The statistical techniques allow exploitation of complex relationships between PHIs and other text features, which even human experts might be oblivious to. In addition, they are more easily generalizable to different data sets. Yet, the results of the i2b2 de-identification challenge show that systems that combine machine learning with regular expressions yield the best results.

3 Method

PHI Detection

dependent uses a combination of pattern-matching and machine learning algorithms in order to detect all relevant identifiers. In doing so, our goal is simple: To provide an easy, computationally efficient procedure which can compete
with the approaches proposed in the literature. Incorporated into a desktop application, it allows users to perform de-identification tasks without a) prior programming experience, b) knowledge of the underlying procedure, and c) without having to tinker with different parameters in order to produce sound results. Furthermore, our system comes with a pre-trained model that does not require additional training data. We believe that the advantage of having a ready-to-use solution outweighs the downside of the marginal decrease in performance compared to tailored models, the more so since training data is often not available and generating it is a laborious and costly endeavor.

Identifiers such as phone numbers, fax numbers or dates are detected by various regular expressions. These have been designed to cover a wide range of possible ways the aforementioned identifiers can be expressed. For example, there are multiple formats of writing a date (1st of May, 05/01, 5/1 etc.), and the rules have been designed to handle this. A selection of them is displayed in Table 1.

Each record is searched for matches of any of the implemented regular expressions, where the steps of the procedure are ordered such that the most specific ones are checked first.

<table>
<thead>
<tr>
<th>Type</th>
<th>Regular Expression</th>
</tr>
</thead>
<tbody>
<tr>
<td>Email Addresses</td>
<td>/\b[A-Z0-9_.%+-]+@[A-Z0-9.-]+.[A-Z]{2,4}\b/gi</td>
</tr>
<tr>
<td>Phone Number</td>
<td>/+?[0-9]{1,3}[/\s.-]?\d{3}[/\s.-]?\d{4}/g</td>
</tr>
<tr>
<td>Fax Number</td>
<td>/+?[0-9]{7,}/g</td>
</tr>
<tr>
<td>SSN</td>
<td>/\d{3}-?\d{2}-?\d{4}/g</td>
</tr>
</tbody>
</table>

Table 1: Perl-style regular expressions to detect various patient identifiers used by the deidentify tool.

NER Detection via Conditional Random Fields

For the detection of persons, locations and organizations, we rely on the Conditional Random Field (CFR) model for Named Entity Recognition (NER) developed by Finkel et al.¹⁰ To motivate the use of CFRs, one should emphasize that NER — compared to classification tasks, which predict a single categorical response variable — requires a sequence model because the goal is to detect the underlying state for each word in a sequence of words. One popular choice has been Hidden Markov models (HMMs). HMMs model the sequence of T observed variables \( X = \{x_t\}_1^T \) and their underlying states \( Y = \{y_t\}_1^T \) by making the following simplifying assumption about the joint distribution of \( X \) and \( Y \): \( p(y, x) = \prod_{t=1}^T p(y_t|y_{t-1})p(x_t|y_t) \), with the initial state distribution being denoted as \( p(y_1|y_0) \).

The Markov property in this case refers to the assumption that each state only depends on its direct predecessor and is independent of all other states. As can be glanced from the factorization of the joint probability distribution, the other assumption of the model is that each observed value depends solely on its underlying state. Instead of using the current word as the single feature for each \( t \), it might be more useful to consider a feature vector instead. However, when \( x \) is high-dimensional, estimating the joint distribution becomes either intractable or requires one to place strong independence assumptions on the covariance matrix of the feature vectors.

A solution for this is to model the conditional distribution \( p(y|x) \) instead of the full joint distribution. This avoids making any assumptions about \( p(x) \) and is sufficient for prediction tasks. This rationale gives rise to Conditional Random Field (CFR) models. When \( G \) is a factor graph over \( Y \), a general conditional random field is a conditional distribution of the form

\[
p(y|x) = \frac{1}{Z(x)} \prod_{\Phi_A \in G} \exp \left\{ \sum_{k=1}^{K(A)} \theta_{ak} f_{ak}(y_a, x_a) \right\},
\]

where \( f_{ak} \) are real-valued feature functions, \( \theta_{ak} \) associated weights and \( Z(x) \) is a normalization function.¹⁰

CRF models do not assume independent features and allow for long-range dependencies in the data taking subsequent and preceding words into account, both very desirable properties for a NER system. On these grounds, Finkel et al.¹⁰
argue that CRFs are preferable to Hidden Markov Models (HMMs) and MaxEnt Markov Models (MEMMs) for NER
tasks.

The Stanford Natural Language Processing Group provides a Conditional Random Field (CRF) model with developed
feature extractors and offers models pre-trained on large corpora (CoNLL, MUC-6, MUC-7 and the ACE named
tility corpora), making them useful for application in many different domains, including the health care domain. The
Stanford CRF model uses the following features: the current word, the previous word, the next word, the current word
character n-gram, the current Part-of-Speech (POS) tag, the surrounding POS tag sequence, the current word shape
and the surrounding word shape sequence as well as indicators for the presence of the word in a left or right window
(with a window size of four).

In contrast to training a custom model, using a pre-trained model has the distinct advantage of allowing real-time
processing and alleviates the need for a large data corpus or labeled training data. If required, it is always possible
to switch out the default model used by deidentify for a custom-built model. In most situations this should not be
needed. In cases where single persons or organizations are not detected accurately, one can manually add them to the
database of known entities by means of a visual interface. The list of known entities is compared against the text to be
de-identified and any matches are replaced.

Replacement Strategy

As a result of the previous procedures, the approach retrieves a list of protected health information identifiers detected
in the analyzed free-text documents. These can subsequently be replaced by automatically generated substitutes,
tokens or type identifiers. For each identifier class, our software contains algorithms that can create an appropriate
replacement text. For example, if a detected identifier is a phone number, it is replaced by a randomly generated phone
number that does not belong to a real patient. For most of the identifiers, the replacements are created without taking
the original identifiers into account, i.e. the digits of the new phone number do not depend in any way on the original
phone number. Therefore, reverse engineering of the original values is prevented.

Replacing identifiers in isolation would likely render many texts unreadable, e.g. when a doctor repeatedly mentioned
in a document is replaced by references to multiple other individuals. In order to preserve the internal consistency of
the documents, we ensure that each occurrence of a single identifier is mapped to the same replacement by performing
a database look-up that checks whether the identifier has already been substituted before. If the result is positive, the
previously used replacement is used instead of generating a new one. Hence, each occurrence of a patient’s name will
be mapped to the same fictitious patient name. Thereby, the narrative and meaning of the documents are preserved.

<table>
<thead>
<tr>
<th>Original</th>
<th>Substituted</th>
<th>Identifier</th>
<th>Redacted</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dear Janine Keane, as we have discussed, I hereby send you the requested information about my patient, Julie Andrews. You can reach her via email (her address is <a href="mailto:jandrews@gmail.com">jandrews@gmail.com</a>) or via phone: 998 785 6756. Sincerely, Elijah Hunt, MD</td>
<td>Dear Rosie Copeland, as we have discussed, I hereby send you the requested information about my patient, Beatrice Burton. You can reach her via email (her address is <a href="mailto:lebewukak@halava.fi">lebewukak@halava.fi</a>) or via phone: (836) 230-3149. Sincerely, Jayden Bush, MD</td>
<td>Dear &lt;name&gt;, as we have discussed, I hereby send you the requested information about my patient, &lt;name&gt;. You can reach her via email (her address is &lt;email&gt;) or via phone: &lt;phone&gt;. Sincerely, &lt;name&gt;, MD</td>
<td>Dear **********, as we have discussed, I hereby send you the requested information about my patient, *************. You can reach her via email (her address is ****************) or via phone: ************. Sincerely, ***********, MD</td>
</tr>
</tbody>
</table>

Table 2: In the left column, a made-up note of a doctor to one of his colleagues is displayed. The remaining columns show the outputs of the different replacement strategies implemented in deidentify
The deidentify software is customizable in terms of which identifiers are replaced and the strategies employed to do so. For example, it is possible to replace only identifiers of a specific type or to specify that first names are mapped only to names from the same gender. To support longitudinal studies, the newest version of deidentify also comes with facilities to specify ranges for date replacements, such that consistency of the documents is preserved and one can appropriately place them in time. However, it must be noted that there is always a trade-off in terms of convenience on the one hand and privacy on the other: Obviously, the risk of identifying individuals when the employed replacement strategy is not truly random increases. However, we think that users should make their own judgments and that a tool should offer them options so that it can be adapted to their specific needs.

An example text alongside the different substitution strategies is displayed in Table 2.

Software

The current version of the software and the source code is available in a repository on GitHub: https://github.com/Planeshifter/deidentify/releases. The deidentify software is freely distributed under the GNU General Public License v2. Built upon Chromium and Node.js, it runs on all major operating systems (MacOS, Linux and Windows). Being agnostic to the used operating system, durability of the tool is ensured. Integrating the Stanford NLP CRF classifiers, a run-time version of Java 1.8 or later must be installed for deidentify to work. There are no other prerequisites. The software supports de-identification of texts stored in *.pdf, *.doc(x), or *.txt format.

The software comes with an easy-to-use point and click interface, which allows both de-identification and inspection of individual files as well as batch processing of all files stored in a given folder.

Processing the data is time-intensive due to the advanced machine learning techniques applied under the hood. On an Intel i7-3540M quad-core laptop with 3.00GHz and 8GiB of RAM, the software processed 170,000 words per hour, or 2,833 words per minute.

4 Evaluation

The quality of predictive models is often measured by their recall, precision, and false positive rate. In our context, recall (also known as sensitivity) is the proportion of PHIs correctly identified throughout the doctor’s notes; precision (also known as positive predictive value) represents the proportion of correct findings among the terms identified as PHI. Finally, the false positive rate depicts the proportion of wrongly identified PHIs among non-PHI words. Following the goal of HIPAA to protect the privacy of patient’s health care data, achieving a high recall is quintessential, whereas a high precision is desired in order to preserve the integrity and readability of the text. The same applies for the false positive rate.

Notice that the overall correctness in labeling text as either being a personal identifier or not is not a good measure of performance: Most character sequences are not personal identifiers, and will be correctly treated as such by almost any algorithm. Yet, these will far overshadow any differences in the performance in detecting actual identifiers, which is what we ultimately care about. The cost of making a mistake is different depending on whether we misclassify an observation as a PHI when it is in fact not (i.e., a false positive), which is a minor issue, compared to a false negative, i.e. missing a PHI and thus putting patient’s privacy in jeopardy.

First, we have evaluated our method by measuring precision and recall on a gold standard corpus of nursing notes made available on PhysioNet.10 The data set was originally collected by Neamatullah et al.11 as part of an evaluation of their deid software for PHI de-identification. The data set contains 2,434 records, all of which have been de-identified manually by several clinicians. The authors manually labeled instances of PHIs and replaced them by realistic surrogate data. The records are nursing notes, which were manually inserted by the nurses at the end of their shift and thus contain typos, incorrect punctuation, and false orthography.

Here, we employ the term gold standard to solely signify that we treat the PhysioNet corpus of nursing notes as correctly annotated and of high quality. It is one of the few publicly available data sets of medical notes and has a history of being used a benchmark to evaluate new approaches for de-identifying textual medical records. On the other hand,
many data sets have characteristic features that distinguish them from each other, so that a method attaining the best performance on the PhysioNet corpus may not generally be superior.

However, coupled with the usage of technical terminology and non-standard abbreviations, it is apparent that this the PhysioNet data set represents a challenging corpus for any de-identification software. The following is an excerpt of one of the nursing notes. It shows the use of abbreviations and is furthermore written entirely in capital letters, a fact which significantly complicates the detection of proper nouns in the English language:

58 YO FEMALE READMITTED TO CCU TODAY S/P CATH WITH PA LINE ON MILRINONE. PT WITH PMH MI ‘92, CABG X3 ‘92, REDO ‘95, DDD PACER ‘95, AFLUTTER S/P ABLATION ‘96, AFIB S/P CARDIOVERSION

The motivation behind developing deidentify was the need to deidentify physician’s notes as part of a project on the development of Chronic Kidney Disease. We have used a sample of forty-eight records to evaluate the de-identification performance, in addition to the gold-standard data set from PhysioNet. Our records are doctor’s notes written by CKD specialists to the referring physician after each patient visit. Hence, they have – as can be seen from the following example – the form of a letter:

Dear XXX, I saw your patient, Mrs. YYY, in the office today for follow up of her transplantation. As you will recall she had problems with acute interstitial nephritis which ended up being endstage renal disease in early 2000’s and has subsequently undergone transplantation from one of her family members. She was seen in September at which point her creatinine was stable at 1.1. Her Prograf was somewhat at low limits of acceptable levels (…)

As these short excerpts illustrate, a major challenge of providing an off-the-shelf de-identification software is given by the vast differences between the nature of free-text documents to be stripped of identifying information, which often requires manual intervention to ensure good results.

5 Results

On the gold-standard corpus, our method successfully de-identified 774 out of 822 names belonging either to doctors, patients or relatives. Locations were a bit harder to detect: Out of 367 locations in the test set, 95 were missed by deidentify. Highly structured identifiers such as phone records were correctly identified 100% of the time. Overall, this translates to a recall of 0.919. In total, we identified 1,600 of the 1,724 labeled identifiers in the nursing note corpus. deidentify flagged 2,674 PHIs in the corpus in total. This gives rise to a precision value of 0.645. The results for the different PHI types are summarized in Table 3.

The gold-standard test set by Neamatullah et al.11 poses significant challenges for any de-identification procedure because the nursing notes are not written in complete sentences and not consistently capitalized. From our experience, deidentify performed significantly better on other data, which does not share these characteristics. This is to be expected, as a clear sentence structure and capitalization of words affect the performance of the used NER system.

Nonetheless, the results are comparable to other methodologies proposed in the literature. In terms of recall, we improve upon the work by Neamatullah et al., who obtained a recall of 0.834 on the gold standard corpus for the case where no customized dictionary was used. This is the appropriate comparison since the performance attained by our procedure is also achieved without tailoring its dictionaries to the data set. On the other hand, the precision attained by our method is a bit lower, which might reflect the inherent trade-off between the goals of achieving a high precision and a high recall.

The low precision value achieved for dates might look a bit worrisome. The reason for the large number of false positives is that many lab values referenced in the notes are incorrectly classified as dates and substituted. Correctly disambiguating these from each other is a surprisingly challenging problem, which cannot completely be solved via pattern matching. However, since lab values are usually available in the structured data fields of the EHR, this seems
Table 3: Summary of de-identification performance for different PHI types. For each type and sub-type, the table displays the total number of identifiers from the respective category in the gold standard, and the number of false negatives as well as precision and recall of the deidentify procedure. See Section 4 for a discussion on why these are useful metrics to evaluate the performance of a de-identification procedure.

<table>
<thead>
<tr>
<th>PHI Type</th>
<th>PHI Sub-Type</th>
<th>Count</th>
<th># FNs</th>
<th>Recall</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>Patient Name</td>
<td>54</td>
<td>3</td>
<td>0.944</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Patient Initial</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Clinician Name</td>
<td>593</td>
<td>41</td>
<td>0.925</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Relative/Proxy Name</td>
<td>175</td>
<td>2</td>
<td>0.989</td>
<td></td>
</tr>
<tr>
<td>Name (overall)</td>
<td></td>
<td>822</td>
<td>48</td>
<td>0.95</td>
<td>0.734</td>
</tr>
<tr>
<td>Date</td>
<td></td>
<td>482</td>
<td>4</td>
<td>0.992</td>
<td>0.256</td>
</tr>
<tr>
<td>Location</td>
<td></td>
<td>367</td>
<td>95</td>
<td>0.741</td>
<td>0.922</td>
</tr>
<tr>
<td>Phone</td>
<td></td>
<td>53</td>
<td>0</td>
<td>1.00</td>
<td>0.899</td>
</tr>
<tr>
<td>Overall</td>
<td></td>
<td>1724</td>
<td>124</td>
<td>0.919</td>
<td>0.645</td>
</tr>
</tbody>
</table>

Table 4: Summary of de-identification performance for names in the sample of doctor’s notes. The table displays the total number of identifiers from the respective category, the number of false negatives as well as precision and recall of the deidentify procedure.

<table>
<thead>
<tr>
<th>PHI Type</th>
<th>PHI Sub-Type</th>
<th>Count</th>
<th># FNs</th>
<th>Recall</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>Patient Name</td>
<td>24</td>
<td>0</td>
<td>1.0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Clinician Name</td>
<td>112</td>
<td>6</td>
<td>0.946</td>
<td></td>
</tr>
<tr>
<td>Name (overall)</td>
<td></td>
<td>136</td>
<td>6</td>
<td>0.956</td>
<td>0.738</td>
</tr>
</tbody>
</table>

6 Discussion

The performance of our de-identification strategy is comparable with the results of the deid software by Neamatullah et al.11 for the case without a customized dictionary. This is a fair comparison, given that our results were obtained without the use of a custom dictionary as well. It might well be that there is an upper limit in terms of accuracy that can not be crossed without relying on either a custom-built model or alternatively a curated set of dictionaries containing identifiers to be removed. Using a custom dictionary allows one to add terms that need to be replaced, which any detection algorithm might otherwise find it impossible to flag. One example would be physician initials on notes that should be removed to ensure proper de-identification, but which most algorithms would not be able to correctly flag
as identifying information.

The Medical De-identification System (MeDS) by Friedlin et al. achieved impressive results (a recall of more than 99% of all HIPAA-specified identifiers and an equally impressive precision of more than 90%) using a solely rule-based system.

Uzuner et al. report on the results of a de-identification challenge that was part of the Informatics for Integrating Biology and the Bedside (i2b2) project. The participating teams achieved recalls between 0.80 and 0.96, and precision values between 0.74 to 0.97.

Several of the systems submitted as part of the challenge “took advantage of the specific organization of discharge summaries that are characteristic of the institution from which these were drawn” and did worse when applied to different data. MeDS was able to utilize patient information in the header section of documents as well.

When using deidentify as an off-the-shelf tool, performance is as expected lower than what is achieved by those systems. Yet, the differences are not very large, and one should keep in mind that deidentify can be run with a custom-built model and allows users to manually expand the dictionaries extracted from the texts. However, a direct performance comparison with the tools of Friedlin et al. and Uzuner et al. was not possible since their tools are not publicly available. They share this feature with the vast majority of de-identification tools. In their literature review, Meystre et al. found that just five out of eighteen investigated tools were made publicly available.

In several studies, human annotators performed comparatively well to the studied software systems. Given the laborious nature of manual annotation, de-identification software thus provides huge savings in terms of effort, while at the same time leading to similar less-than-optimal performance.

To preserve patient privacy and to fulfill all legal requirements, we deem it necessary to still refine the results of our procedure by manually expanding the built dictionaries and/or retraining the used NLP models on data similar to those to be de-identified. With the graphical user interface of deidentify, this is an easy task. After an initial run on a selection of notes, one should review the results and collect all instances of missed identifiers. From our experience, the same issues crop up over and over again, and hence even only a small custom dictionary can yield major benefits. For example, we were able to improve the performance of our tool by supplying a list of foreign names that the algorithm missed out on.

7 Conclusion

The need for de-identification of medical records is growing given the proliferation of EHR and other clinical information systems. To comply with the regulations imposed by the HIPAA legislation, protected health information (PHI) has to be removed from medical records so that they can be shared among researchers. Yet, there are not many freely available and easy-to-use tools to perform this task.

We have developed a new de-identification software built using state-of-the-art web technologies for removing personal identifiers from free-text medical records. Our tool uses a combination of pattern matching and machine learning algorithms, leveraging the power of the Stanford Named Entity Recognition (NER) system.

Besides offering good performance, we have implemented several secondary characteristics that a de-identification tool should ideally possess:

- It should provide a graphical user interface (GUI) so that it can be used by non-experts.
- It should be platform independent and easy to setup, without requiring lots of third-party software to be installed manually.
- It should be applicable even in absence of training data.
- It should be open-source and not proprietary software, as to ensure that the source code can be studied by everyone and changes can be made to accommodate specific use cases.

deidentify was designed with these goals in mind, and we hope that it offers a convenient and efficient way to scrub medical data of personal identifiers. This could significantly lessen the burden placed on individuals who wish to conduct research on free-text medical data.
The field of NLP has made considerable progress in recent years, including the rise of deep learning. Extrapolating this trend, it is highly likely that deidentify may be improved in the future due to improvements in Named Entity Recognition (NER) and related algorithms. In the near term, the tool could be improved by a better handling of dates going beyond regular expressions, for example by taking the context into account.

Further improvements could come from the embedding of structured information related to the texts to be identified (laboratory results, medical staff lists, patient information etc.). For example, structured laboratory results could prevent that these values are mistaken for dates when they appear in the free-text documents. The upcoming release of deidentify allows users to supply a selection of whitelisted words for each record, which are not replaced under any circumstances.

The usage of the tool in a variety of settings could provide further insights into its current flaws and how they could be mitigated. This potential emphasizes the benefits of releasing medical informatics software under open-source licenses, as well as providing a platform where potential improvements, feedback, and bug reports can be shared.

References


Mobile apps for mood tracking: an analysis of features and user reviews

Clara Caldeira, BS¹, Yu Chen, PhD¹, Lesley Chan, BS¹, Vivian Pham, BS¹, Yunan Chen, PhD¹, Kai Zheng, PhD¹
¹University of California, Irvine, Irvine, CA, USA

Abstract

Many mood tracking apps are available on smartphone app stores, but little is known about their features and their users’ experiences. To investigate commercially available mood tracking apps, we conducted an in-depth feature analysis of 32 apps, and performed a qualitative analysis of a set of user reviews. Informed by a widely adopted personal informatics framework, we conducted a feature analysis to investigate how these apps support four stages of self-tracking: preparation, collection, reflection, and action; and found that mood tracking apps offer many features for the collection and reflection stages, but lack adequate support for the preparation and action stages. Through the qualitative analysis of user reviews, we found that users utilize mood tracking to learn about their mood patterns, improve their mood, and self-manage their mental illnesses. In this paper, we present our findings and discuss implications for mobile apps designed to enhance emotional wellness.

Introduction

The World Health Organization (WHO) has defined mental health as “a state of well-being in which every individual realizes his or her own potential, can cope with the normal stresses of life, can work productively and fruitfully, and is able to make a contribution to his or her community.” Mental disorders such as stress and depression can be burdensome and disruptive, and also increase the risk of becoming ill from other diseases such as cardiovascular disease and diabetes. Despite the importance of mental well-being, initiatives to increase mental health awareness and services have been hindered by the absence of mental health in public health agendas and the lack of integrating mental health with primary care. Creating technology to support self-management of mental health may benefit consumers by empowering them to have a more active role over their own well-being, and making mental health resources more accessible. Along this line, both researchers and industry are increasing their efforts to develop technological solutions that enable individuals to manage their mental health. For example, Google has recently launched Verily, a healthcare lab that aims to digitalize mental health with the help of brain and behavioral sciences.

Mood tracking has been investigated as an approach to help healthy individuals stay in healthy emotional states, and assist individuals with mental diseases, such as bipolar disorder and depression, in their health management. Tracking health related data helps users to gain knowledge about their health, establish the correlation between the tracked data and their health conditions, and proactively engage in healthcare management. Self-tracking mood helps users increase their awareness and proactive self-regulation of their emotional well-being. The ubiquity of smartphones and their apps has enabled their use for a myriad of health purposes. According to a report from Statista, “health and lifestyle” is one of the most popular categories in smartphone apps. Self-tracking is among the most common strategies used by health apps, including tracking preventative behavior such as exercise, as well as monitoring health indicators, such as blood pressure and mood. The latest versions of smartphones provide built-in self-tracking functions for health.

Researchers have argued that many apps lack scientific- and clinical-based evidence in the feature design, and also lack evidence of consumers’ empirical usage. A few studies have investigated Android and iOS apps for specific health conditions, such as mental diseases and weight loss. To the best of our knowledge, little research has investigated the features, consumers’ experiences, and challenges of using mobile apps that are designed for mood tracking. Despite emerging sensing technologies and apps that could potentially detect users’ emotions, since mood and emotions are subjective, the accuracy of such sensing technologies remain unclear. Therefore, we investigate apps that allow users to self-report their mood, and we use self-tracking to refer to keeping track of mood via self-report.
Background

Emotional well-being is an essential part of mental health\textsuperscript{13}. Positive mood can enhance cardiovascular, hormonal and immune functions, promote healthy behaviors such as better sleep and exercise\textsuperscript{14}, and lead to more open-minded thinking and effective problem solving\textsuperscript{15}. Self-tracking mood could facilitate gaining knowledge and awareness of one’s mood patterns and thus help maintain emotional well-being. Researchers have employed various technologies to help individuals to record, analyze, and regulate mood\textsuperscript{16,17}. For example, MobiMood\textsuperscript{18} allows users to register their mood by selecting colors, and Aurora\textsuperscript{19} supports self-tracking by allowing users to select photos that best represent their mood. CopeSmart\textsuperscript{20} users log their mood by selecting a corresponding value on a scale. Other mobile apps have elicited users’ mood using emojis\textsuperscript{21}. Besides self-tracking, apps such as Aurora\textsuperscript{19} and MobiMood\textsuperscript{18} also encourage users to share self-tracked mood data in their social networks to seek emotional support\textsuperscript{22}. Study results suggest that users are willing to share their mood with their friends in social networks\textsuperscript{18,19}. In addition to designing self-tracking technologies for the general population, researchers have also investigated the use of mood tracking to help patients with mental diseases self-manage their conditions\textsuperscript{23}. For example, MoodZoom\textsuperscript{24} helps participants with bipolar disorder to rate anxiety, elation, sadness, anger, irritability, and energy on a 7-point Likert scale. The above studies mainly investigate one particular app in research settings.

Since the use of mobile apps for promoting health has grown exponentially in recent years, researchers have started to systematically review apps to identify features, opportunities, and challenges for healthcare practices\textsuperscript{8}. For example, Payne et al.\textsuperscript{25} have conducted a review on the behavioral functionality of mobile apps in health interventions, and suggested that mobile apps are a feasible and acceptable means of health intervention. However, they have also called for more rigorous research and design guidelines for the functionalities and interfaces of apps available for smartphone. Other researchers have provided comprehensive analyses of the literature on mobile apps for specific health conditions, such as mental illnesses\textsuperscript{9} and weight loss\textsuperscript{10}. These studies pointed out the limitations of evaluating mobile apps for health intervention within a limited number of participants using one specific app. Systematic reviews of health apps have identified a gap between the standards in scientific and clinical research, and the actual features in consumer products. For example, Nicholas et al.\textsuperscript{5} surveyed apps designed for bipolar disorder, and found that the available apps failed to provide essential information to help users assess their conditions, and argued that a new framework for mobile mental health research was essential to provide evidence-based health management. To the best of our knowledge, there has been no systematic review of mobile apps for mood tracking, how features of such apps align with healthcare research findings, and consumers’ experiences with using such apps in their daily lives.

Methods

To investigate features of apps currently available for smartphones, and their users’ experiences, we conducted a review on mood tracking apps, evaluated their features, and analyzed a set of consumer reviews.

App selection

In order to find a set of currently popular apps, a systematic search was conducted in the Android Play Store and the Apple App Store in January 2017. We searched for apps using the following keywords: (Mood OR emotions OR feelings OR affective OR happiness) AND (track OR log OR journal OR record OR diary). Apps were deemed eligible if they met all of the following criteria: 1) mood tracking is the main focus of the app, 2) mood is self-reported (not detected by sensors), 3) the app is in English, 4) it was last updated no earlier than 2014, and 5) it has at least 25 ratings. While consumers might provide ratings and reviews for an app, the number of ratings and reviews might differ, since users might rate an app (i.e., give it a score between 1 and 5 starts) without writing a review (i.e., a free text comment). We chose ratings as a measure of popularity because it was a comparable measure provided by both Android and iOS app stores. Criteria measuring popularity and recent updates were used to exclude apps of poor quality, which are not maintained or heavily used currently. The first two authors independently screened each app by reading its name and description, and examining any screenshots available in the app stores to decide on a set of apps for an in-depth feature and review analysis. Disagreement was dissolved through discussion.

App feature analysis

In this study, we conducted a feature analysis of mood tracking apps. For each app, two of the first four authors down-
loaded the app, extracted the main features, and then coded the features based on the stages of personal informatics tools each apps supports. To conduct the feature analysis, we used a set of key measurements based on the stage-based model of personal informatics systems11 – systems that help users collect, reflect on, and gain knowledge on their personal data11. We employed this model because it is a widely-adopted conceptual framework that can be used to evaluate self-tracking tools for each stage of the self-tracking process. The personal informatics model describes five stages: 1) the Preparation stage consists of planning that occurs before users start collecting personal data, 2) the Collection stage refers to when users record their data, 3) the Integration stage includes formatting and combining data, 4) the Reflection stage involves making sense of and learning from their personal data, and 5) the Action stage, when users act based on the insights gained through reflection. Self-tracking systems should be designed to support each stage11. While there are more recent models of self-tracking practices (e.g., Epstein et al.12), we choose the original model for our analysis because it is more simple, and because the model by Epstein et al.12 includes stages that do not need to be supported by self-tracking tools (e.g., deciding to track and selecting a tool).

App review analysis

In addition to reviewing system features, we analyzed reviews that users left in the app stores for the apps selected. Analyzing user reviews allows us to gain insights from the perspectives of the consumers of mood tracking apps. To do so, we downloaded all reviews from each app that were written prior to January 2017. We used two criteria to screen each review: 1) it is written in English, and 2) it is longer than 20 characters. Due to the large number of reviews, we randomly selected a subset of reviews for qualitative coding. Two independent coders read each app review and noted their themes. After reaching data saturation, we verified the remaining reviews to validate our results. Then, we used axial coding to extract themes present in the data26. Because it is not possible to obtain background information on who wrote each review, we were not able to differentiate between users with different characteristics (e.g., different genders or ages).

Results

App selection

The initial search returned 249 hits on the Android Play Store and 493 on the iOS App Store. After screening and discussion between the first two authors, 32 apps were selected, including 19 Android apps and 13 iOS apps (Table 2). Most of the selected apps are designed for the general population. Only one application – eMoods Bipolar – focuses on a particular health condition. There are two explanations for this. First, apps focused on specific populations, such as those with mental illnesses, tend to place mood tracking as a secondary focus, with other functions such as medication management as the main emphasis. Since in this study we are interested in apps designed specifically for mood tracking, many apps that aim for specific mental health management were excluded. Second, mood tracking apps that focus on a particular population tend to have fewer users, which makes them less likely to have over 25 ratings than mood tracking apps for the general population. Thus, all but one of the apps included in our review are designed for a general population. For several of the apps, we included both Android and iOS versions (e.g., iMoodJournal). However, for five of them, one of the two versions did not meet the criterion of over 25 ratings, and thus they were not included in the analysis. These apps are marked with an * in Table 2.

App feature analysis

After iteratively coding features using the personal informatics stage model, we categorized the features into four stages in Table 1. For the Preparation stage, we included features that provide information to prepare users for tracking their moods, such as instructions, explanations, and other resources about the app itself or the type of mood tracking introduced in the app. The Collection stage consists of app features that support recording mood using different methods, and setting reminders for logging mood in the future. For the Reflection stage, the most common app feature is the data visualization function, which comes in various styles. Finally, the Action stage mainly includes 1) a recommendation feature in which an app provides any recommendations or resources for further actions based on the mood recorded by users, and 2) a sharing feature that allows users to share or export their mood data. We did not include the Integration stage in this analysis because all of the selected apps handled this process automatically.

We analyzed the features of each application, including their functionalities and user interfaces. In Table 2, we present
<table>
<thead>
<tr>
<th>Stage</th>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preparation</td>
<td>instructions &amp; explanation</td>
<td>information about how to conduct mood tracking</td>
</tr>
<tr>
<td>Collection</td>
<td>interface for collecting mood data</td>
<td>text, pictures, colors, emoticons, etc</td>
</tr>
<tr>
<td>Reflection</td>
<td>visualizations</td>
<td>graphs, tables, lists</td>
</tr>
<tr>
<td>Action</td>
<td>recommendations</td>
<td>further actions based on data, e.g., how to improve mood</td>
</tr>
<tr>
<td></td>
<td>sharing</td>
<td>features for users to share or export data</td>
</tr>
</tbody>
</table>

The features in each app that correspond to four stages of mood tracking: Preparation, Collection, Reflection, and Action.

For the Preparation Stage, we found that only a few (N=8) applications provided substantial information about how to track mood, or about how tracking mood can help manage and promote well-being. Other apps provided little or no such information. However, even for apps that do provide features for the Preparation stage, such information is often very limited, and usually only appears when users first launch the app. Often, this information is not shown again after the first time the app is launched.

For the Collection Stage, all apps allow users to record their mood. We have found a diverse set of tracking interfaces, including 1) pre-defined text, i.e., text provided by the app that the users can select, such as ‘optimistic’ or ‘tired’, 2) free text, i.e., text that users can type themselves, 3) colors, i.e., choosing a color to represent the current mood, 4) pictures, either taking a picture or choosing one on the app to illustrate the current mood, 5) recorded audio, 6) emojis, i.e., picking an emoji or similar image to represent the current mood, and 7) ratings, e.g., 21 apps allow users to select the intensity of a particular mood. For instance, MoodPanda users could rate their happiness on a scale from 0 to 10. A few apps also support users to add notes to their mood. Notes could include contextual information, or a more detailed description of the mood. As for reminders, about one third of the apps do not include features for reminding users to enter data. Figure 1 presents screenshots of apps with different tracking interfaces.

**Figure 1:** Example screenshots of the Collection stage illustrating mood collection interfaces

For the Reflection Stage, we found that most apps (N = 29) provide mood visualization, such as bar and pie graphs. Other visualization modalities include lists, line graphs, calendar view, map view, and mood meter. For instance, pie charts and meters can help users get an overall view of their moods; calendar view and line graphs can facilitate users to find patterns in time, and map view can assist learning association between mood and physical locations. Many apps provide visualization support in one or multiple modalities for reflection.
<table>
<thead>
<tr>
<th>Android</th>
<th># reviews</th>
<th>information</th>
<th>words</th>
<th>custom words</th>
<th>notes</th>
<th>scale</th>
<th>color</th>
<th>audio</th>
<th>pictures</th>
<th>emoji</th>
<th>reminders</th>
<th>graphs or charts</th>
<th>calendar view</th>
<th>map view</th>
<th>recommendations</th>
<th>sharing features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diary - Mood Tracker*</td>
<td>6857</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Moodtrack Diary</td>
<td>921</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>T2 Mood Tracker</td>
<td>710</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>eMoods Bipolar</td>
<td>444</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>iMoodJournal</td>
<td>419</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Mood Log</td>
<td>318</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Moodlytics</td>
<td>295</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>iFeel Free*</td>
<td>215</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>MoodDiary BETA</td>
<td>168</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>MoodPanda</td>
<td>91</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Diary, mood tracker</td>
<td>40</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Mood Tracker</td>
<td>37</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>My Life My Voice</td>
<td>33</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Mood Meter</td>
<td>22</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Mood O'Scope*</td>
<td>19</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>WINKIA EMO*</td>
<td>17</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Mood diary tracker</td>
<td>12</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Schema Therapy</td>
<td>11</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Happy Meter</td>
<td>10</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>iOS</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>iMoodJournal</td>
<td>1493</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Moodnotes</td>
<td>864</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>T2 Mood Tracker</td>
<td>169</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Moody</td>
<td>102</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Moodtrack Diary</td>
<td>101</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>MoodPanda</td>
<td>93</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Moodlytics</td>
<td>62</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Happiness</td>
<td>59</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Moods</td>
<td>43</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Mood Meter</td>
<td>41</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>My Life My Voice</td>
<td>39</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Koko</td>
<td>33</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Easy Mood Diary*</td>
<td>25</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

For the Action Stage, only 7 of the 32 apps we analyzed provide features to support future actions, including recommendations for handling depression, contact information for a support hotline, strategies to avoid triggers, handle stress, and improve mood, and social features which help users look for support. Further, 25 apps have sharing features that allow users to send their mood data to their social networks for additional support, or to export the data in free text or spreadsheet format.
**App review analysis**

After the feature analysis, we investigated consumers’ reviews of the mobile apps. In total, we obtained an initial data set of 13,736 reviews. After excluding reviews not written in English or shorter than 20 characters, 8,584 reviews remained. We then randomly selected 1,000 reviews for analysis. From the app reviews extracted from the app stores, our analysis revealed consumer experience in the following aspects: consumers’ characteristics and goals in using the apps, their current app usage, and challenges they face. We summarize the findings in Table 3.

**Table 3:** Summary of main findings from consumer reviews

<table>
<thead>
<tr>
<th>Perspectives</th>
<th>Aspects</th>
<th>Stage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Motivation</td>
<td>Find patterns in mood data</td>
<td>Reflection</td>
</tr>
<tr>
<td></td>
<td>Learn about influencing factors</td>
<td>Collection</td>
</tr>
<tr>
<td></td>
<td>For users with mental conditions to monitor symptoms</td>
<td>Reflection</td>
</tr>
<tr>
<td></td>
<td>Make changes to improve mood based on insights</td>
<td>Action</td>
</tr>
<tr>
<td>Current usage</td>
<td>Log mood frequently</td>
<td>Collection, Reflection</td>
</tr>
<tr>
<td></td>
<td>Add explanation to mood to facilitate reflection</td>
<td>Collection, Reflection</td>
</tr>
<tr>
<td></td>
<td>Share data with healthcare providers</td>
<td>Action</td>
</tr>
<tr>
<td>Desired features</td>
<td>Features for specific mental illnesses</td>
<td>All stages</td>
</tr>
<tr>
<td></td>
<td>Add contextual information</td>
<td>Collection, Reflection, Action</td>
</tr>
<tr>
<td></td>
<td>Personalize mood input</td>
<td>Collection</td>
</tr>
</tbody>
</table>

**Motivations of using mood tracking apps.** Users reported that their goals for using mood tracking consist mainly of learning about their mood to find strategies to improve, cope with stress, and manage mental illnesses. For instance, users were often interested in finding patterns in their mood data to understand influencing factors, and those with mental illnesses utilized mood tracking to monitor their symptoms.

“This app is great for keeping me in check and knowing what things contribute to my emotional state. It also makes me stop and think about my day and how I’ve felt! Which is a great factor in being able to turn a bad day around. [...]” (Diary - Mood Tracker, Android).

“I adopted this app as a core component of my self-quantification and improvement effort. It has proven incredibly valuable in helping me analyze patterns and manage my approach to stress and increasing happiness.” (Happiness, iOS).

By monitoring and learning about their own emotions, users reported feeling more in control of their mood, making informed decisions with the goal to become happier, and becoming more confident and positive in their emotional well-being.

Besides the general population, we found that users with mental illnesses also used mood tracking apps, even though the apps we selected were mainly designed for the general population. Users reported having a mental illness in approximately 10% of the analyzed reviews. The most common conditions mentioned in the app reviews were depression, bipolar disorder, and anxiety. In addition to the common mental illnesses, borderline personality disorder, obsessive compulsive disorder, attention-deficit/hyperactivity disorder, posttraumatic stress disorder, psychosis, and schizophrenia, were also cited as diseases that made users feel the need to track their mood.

“I have major depressive disorder and sometimes it is hard for me to keep track of how I feel one day to the next, especially when going to therapy. I actually heard about this app through some people on tumblr and I’m glad I tried it out. Wonderful app. [...]” (Diary - Mood Tracker, Android).

“As someone who suffers chronic physical and mental illnesses, this app is perfect to find trends in what is making me ill. [...]” (Diary - Mood Tracker, Android).

In the above review, users with mental illness also want to use mood monitoring apps because they might have other issues to track at the same time. In such cases, they might choose the app designed for the general population instead of those specifically designed for mental diseases.

Thus, user reviews indicate that the general population mainly uses the apps to learn about their mood patterns, better
manage their emotions, and make informed decisions to enhance their emotional well-being. Meanwhile, some users with mental illnesses also use the apps to monitor their mood. Some of them chose these apps instead of disease-specific apps for more flexibility and monitoring other chronic conditions together.

**Current usage.** We found that many users value the ability to register their mood frequently and add explanations to support reflection. Some users also share their mood data with therapists and psychiatrists. Most reviews that mentioned tracking frequency reported tracking daily, or multiple times per day. The reviews also showed that users appreciated the ability to present and share the visualized mood data with physicians, counselors, or therapists so they could monitor symptoms, or to support a possible diagnosis. A few reviews also mentioned that the app had been recommended to them by a healthcare provider. In a few cases, users with mental diseases tracked their mood alongside a family member, so both could monitor the occurrence of symptoms. These findings suggest that users collaborate with healthcare providers, caregivers, or relatives during the entire mood tracking process, or during the Reflection stage.

“[...] I have a mood disorder and the graphs make a great visual aid for sharing mood fluctuations with my doctor. [...]” (iMoodJournal, iOS).

“[...] [The app] was recommended by my therapist and I will recommend it to everyone I know. [...]” (iMoodJournal, iOS).

Based on the review, it seems that consumers could extend the logged mood data for individual usage to sharing with others. They reported leveraging mood tracking data to communicate with their healthcare providers, family members, and caregivers.

**Desired features of mood tracking apps.** In many reviews, users asked for additional features. The most commonly requested features were the capability to add personalized mood options, process logged data (e.g., order, search by mood/location/time), and update privacy settings (e.g., use a password to limit other people’s access to the data).

Also, while those who reported having a mental illness predominantly reported having positive experiences with the apps, they would often ask for specific features for their particular conditions. For instance, apps may assume that very positive emotions are desirable, but that might not be the case for people with bipolar disorder. For them, being too happy or optimistic might signify a manic episode. Users with mental illness also expected features such as tracking other symptoms or tracking medicine.

“[...] I’d hope to soon be able to sort my emotions by physical location. [...]” (Moodnotes, iOS).

From the consumer review above, it seems that users desire features that help them associate mood data with contextual information, such as location, to better manage their emotional well-being and their health conditions. For users who suffer from mental illnesses, sometimes they hope to have more options for mood and personalize the input.

“[...] The ability to leave notes is great and makes up a little for the lack of variety in mood choices. I also suffer from bipolar and add and find the mood choices way too simplistic when I can feel many different shades of very irritable, beyond happy into excitable and beyond that into a bit manic which might be the top end of the mood spectrum but is completely not nice. [...]” (iMoodJournal, iOS).

“[...] I suffer from a condition that makes me feel things rather intensely and only having five possible emotions to track seems lacking. However, you can add tons of activities and skills.” (Diary - Mood Tracker, Android)

**Discussion**

We found that mood tracking apps have many features supporting the Collection and Reflection stages of self-tracking, such as diverse forms of data collection and visualization. But the analysis revealed a lack of support for the Preparation and Action stages. Through analyzing app reviews, we found that users seek to learn about their mood patterns and improve them, and to manage emotional distress and mental illnesses. Many users pursue these goals by collecting data frequently, registering contextual information through notes, and sharing their data with healthcare providers. Still, they wish for additional features, such as the ability to personalize mood options, process their data for analysis, secure data to preserve privacy, and features specific for particular conditions.
Design app features to support Preparation, Collection, Reflection, and Action. While we found that most apps offer several different kinds of tracking options (Collection) and data visualization (Reflection), support for Action is mainly focused on sharing with social networks or exporting data. Most apps did not provide appropriate features for supporting the Preparation stage. The review analysis also indicates that apps could better support different types of users (i.e., the general population and users with mental conditions) and help them decide whether the mood tracking apps are well suited for them. Because being knowledgeable about self-tracking could help users achieve better results, it is likely that the lack of support for Preparation and Action of tracking constitutes a barrier to the effective use of mood tracking apps.

Our findings indicate that users’ experiences and results could be improved by extending the support for the Preparation and Action stages of mood tracking. Providing users with more information about tracking could help them to better prepare by setting concrete goals, better plan their data collection, and gain more insights about their mood patterns. Presenting them with advice or guidelines about how to handle crises and effectively improve their overall emotional wellness based on their mood data could also improve outcomes by helping users to take appropriate action for their situation and goals. We did not observe a substantial number of requests for these features from the user reviews, which might partially explain why they are seldom present. But supporting the Preparation and Action stages better is likely to help users to become more knowledgeable about the benefits and limitations of self-tracking, and maximize how much they benefit from this practice.

Consider both the general population and users with mental conditions. The majority of the apps analyzed were designed for the general population. There is limited previous work on the benefits of mood tracking for those who do not have mental conditions, but our results show that there is substantial interest in mood tracking from the general public. Our study shows users desire to manage and improve their mood despite not having mental illnesses. Given the fast-paced and busy lifestyles in industrialized societies, mood management is an important area that could be further supported by technology. Our study points to the motivations, current preferred features and future desired features of mood tracking apps, and provides valuable insights for enhancing wellness interventions and features supporting self-knowledge and awareness.

While these popular mood tracking apps may not support the integrated needs of disease management, we found users with many different mental illnesses utilized these apps for their self-management. The diversity of conditions users reported illustrates that mood tracking could assist in the self-management of several mental illnesses. While a recent analysis of available health apps found that many conditions have few or no specific apps available, our results indicate that individuals with these conditions are utilizing apps designed for other populations. Since individuals with mental conditions might have particular needs, even those who have other options (i.e., disease-specific apps) may choose to use apps intended for the general public, due to preferring their design or features. In order to use non-disease specific apps, this population requires more flexible features to be able to track specific aspects related to their illness, such as symptoms related to manic episodes. For these users, receiving support for the Preparation and Action stages is crucial, since such support could help them to track their mood more efficiently and effectively, and to decide on how to act based on their symptoms. Based on this finding, we believe that mood tracking should be investigated as a strategy to support a wide range of mental illnesses through offering flexible and personalized mood tracking options. More effort towards investigating this population’s needs and how effective mood tracking can be for individuals with each illness, could help with the development of more appropriate tools for supporting the health management needs of people with mental illnesses.

Support both individual reflection and data sharing. The app feature analysis suggests that most apps allow users to share their mood data with others. We found from consumer reviews that users share their mood data with their physicians, therapists, and counselors. Users who share the reviews with healthcare professionals are likely to be suffering from mental conditions or mental distress. Sharing mood tracking data with healthcare professionals indicates that these apps are being used not only for self-management, but also for communicating symptoms and collaborating with healthcare providers. The apps provide ways to export data and thus make it easier to share. Similar to other kinds of patient-generated data, it is likely that the kind of data visualization that is appropriate for patients might not be preferred by healthcare providers. Thus, more research is needed to investigate mood tracking apps that can facilitate data sharing between patients and healthcare providers. Further, it is important to investigate how to support
sharing data with family members, caregivers, and friends, while protecting users’ sensitive information. Since family members, caregivers, and healthcare providers might have different motivations and information needs, it might be worthwhile to consider different options in different kinds of data and visualization for shared data.

Implications for clinical practice. The outcomes of this study indicate that mood tracking tools are being used by many people, including people who have mental illnesses. Our data indicate that apps should improve their support of preparation and action activities, but they also can offer benefits to users by helping them to increase awareness of their mood patterns. Health care providers may assist patients in choosing adequate mood tracking apps and using them effectively, if they are informed about which options would match patients’ needs. Providing patients with information on mood tracking could compensate for the limited support for preparation and action activities from the apps.

Analyzing app reviews can reveal valuable insights about real users’ experience, but there are associated limitations. First, since not every user is expected to write reviews, there might be selection bias of the user opinions we can access. Similarly, we are not able to estimate the distribution of user groups, e.g., the general population and those with mental conditions. Since not every user who wrote reviews choose to disclose their mental conditions, we cannot assume that they do not suffer from any conditions if illnesses are not mentioned in the reviews. Additionally, even though we assume that the reviews could reflect consumers’ real experiences, it is difficult to determine how long they have used the app and whether they have abandoned the app.

Conclusion

A qualitative analysis on consumers’ reviews of mood tracking apps showed that both the general population and consumers with mental conditions used the apps. The general population mainly used the apps to learn about their mood patterns and cope with stress, while users with various mental conditions also used the apps to monitor their symptoms. They not only tracked the data for themselves, but also used the data to communicate with their healthcare providers and share with family members.

Acknowledgements

We thank CAPES and the NSF (Grant No. 1219197) for providing funding for this work. Thanks to Brittney Katayama, Christopher Le, Fadumo Ali, Jimmy Chun-Yi, Julie Oh, Oksana Lubyanaya, Xiaoyu Wen, Mary Angeline Ferrer, Ngoc Tran Le, Sean Lee, Tiffany Lee, Carmela Monis, Micah Nacpil, and Kelly Pham for their contributions.

References


26. Corbin J, Strauss A. Basics of qualitative research 3e.


Combining rules, background knowledge and change patterns to maintain semantic annotations

Silvio Domingos Cardoso¹,², Chantal Reynaud-Delaître, PhD², Marcos Da Silveira, PhD¹, Cédric Pruski, PhD¹
¹LIST, Luxembourg Institute of Science and Technology, 5, avenue des Hauts-Fourneaux, L-4362 Esch-sur-Alzette, Luxembourg; ²LRI, University of Paris-Sud XI, France

Abstract

Knowledge Organization Systems (KOS) play a key role in enriching biomedical information in order to make it machine-understandable and shareable. This is done by annotating medical documents, or more specifically, associating concept labels from KOS with pieces of digital information, e.g., images or texts. However, the dynamic nature of KOS may impact the annotations, thus creating a mismatch between the evolved concept and the associated information. To solve this problem, methods to maintain the quality of the annotations are required. In this paper, we define a framework based on rules, background knowledge and change patterns to drive the annotation adaption process. We evaluate experimentally the proposed approach in realistic cases-studies and demonstrate the overall performance of our approach in different KOS considering the precision, recall, F1-score and AUC value of the system.

Introduction

In order to support various tasks in medical information systems such as retrieving, sharing and exchanging information, the data stored within these systems are usually annotated with concept codes from standard Knowledge Organization Systems¹ (KOS). These annotations make the semantics of the data explicit for machines in order to automatize the above-mentioned tasks. However, KOS evolve over time and their elements may be modified, which, in turn, may affect the dependent annotations as shown in our previous work² and drastically impact the reliability of applications exploiting them. Methods and tools are therefore required to keep semantic annotations updated as KOS evolves. The idea is to avoid a complete re-annotation of the document, which is both time consuming and requires human intervention for validation purpose.

Over the past decade, several approaches have been developed to tackle this problem. For sake of clarity, we split this complex problem into two sub-problems: (1) how to detect the annotations impacted by the evolution of the KOS; and (2) how to update impacted annotations to keep them consistent with the new version of the KOS. Many existing approaches focus on the first sub-problem, proposing several different techniques such as logic rules³–⁵ or NLP processes to detect changes⁶,⁷ or even database versioning⁸,⁹. The second sub-problem is mainly addressed by removing the outdated annotations from the corpus⁷, or by adapting existing annotations. By applying the former approach, the user accepts loosing part of the existing knowledge. To avoid this, the latter approach can be adopted. Different adaptation methods are proposed and can be classified according to the techniques used. In⁹, the authors proposed a migration algorithm. Frost et. al.¹⁰ proposed a novel algorithm for optimizing gene set annotations using entropy minimization over variable clusters (EMVC). This filters annotations for each gene set to remove inconsistencies. In their work, Traverso Ribon et al. proposed the AnnEvol¹¹ framework to describe datasets of ontology-based annotated entities. These approaches usually focus on one kind of KOS (e.g. Gene Ontology), or on very expressive ontologies to use reasoning techniques³, and do not exploit information that can be acquired from the analysis of KOS evolution. In our previous work, we designed the DyKOSMap framework for maintaining valid the semantic mappings between evolving KOS¹². However, unlike mappings that formally connect concepts, semantic annotations refer to information of a different nature (e.g. text, images, video ...) that is less rich from a semantic point of view. In consequence, the approach that we proposed for mappings needs to be redesigned to address the semantic annotation maintenance problem.

To overcome the limitations of existing approaches mentioned above, our contribution consists of: (i) the definition of a set of rules following a rigorous analysis of the evolution and adaptation of a set of annotations over a 10-year period² (ii) the combination of these rules with two other techniques to improve the quality of the maintenance process: The first technique relies on the use of background knowledge (BK)¹³ while the second exploits semantic change patterns
Table 1: Example of an evolving annotation, extracted from our silver standard.

(SCP)\textsuperscript{14} that have shown great capabilities for the maintenance of ontology mappings\textsuperscript{12}. Moreover, we provide an experimental assessment of our framework by maintaining a corpus of documents annotated with ICD-9-CM, MeSH, NCIt and SNOMEDCT.

The remainder of the paper is structured as follows: The following section presents our method for adapting semantic annotations. Then, we introduce the experimental evaluation of our approach and the results we obtained at validation time. Finally, we discuss the results and conclude our paper.

Semantic Annotation Adaptation

In this section we describe our rule-based approach for adapting outdated semantic annotations and the material to evaluate it. We represent annotations following the model of the W3C Web Annotation Data Model\textsuperscript{1}. We included some modifications, e.g., a property to link the evolved annotation to its past version, the KOS change which impacted the annotation and other properties\textsuperscript{2}. To illustrate how an annotation is represented in our system, Table 1 contains the original annotation (from 2009) in the first line and the evolved annotation (from 2010) in the second line.

In our analysis, we used the following features: the name and version of the KOS, the reference of the document used as a resource, the concept code, the annotated text followed by the start and end offset, i.e., the position in the document where this annotation can be found, and the prefix and suffix, i.e., the information that comes before and after the annotation. However, if available, more metadata can be included, respecting our annotation form\textsuperscript{2}. The illustrative example shows one annotation produced with the 2009AA MeSH version using the PubMed document 232\textsuperscript{2} and the concept D019684. The annotated text is “Magnoliophyta”, and this can be found in the position [4587,4600]. We set up the system to have four words as a prefix “during the evolution of” and a suffix “(angiosperms) [5]. typical”. We can observe that the concept label used to annotate the text changed from 2009AA to 2010AA.

Our maintenance process relies on the combined use of seven rules, background knowledge and semantic change patterns. The rules proposed derive from our analysis of the evolution of annotations impacted by changes in the underlying terminology\textsuperscript{2}. The rules are part of the implementation of our global framework\textsuperscript{15}. The decision to execute a rule depends on several criteria, but instead of checking all annotations, a first filter is applied: only annotations associated to concepts that have changed (computed with COnto-Diff\textsuperscript{16}) are checked. Unlike rules for mapping maintenance\textsuperscript{12}, the proposed annotation maintenance rules consider many other features used to characterize semantic annotation\textsuperscript{2}. We based our approach on the guideline associated with semantic annotation\textsuperscript{17}. The criteria that we used to guide us during the annotation process also inspired us to define the sequence and conditions that our rules have to be applied.

The seven proposed rules are listed in Table 2. Each column represents one feature: the original concept code and the KOS version ($CP_{v0}$), annotated text ($Annot_{v0}$), a prefix ($Prefix_{v0}$), and a suffix ($Suffix_{v0}$). We also added one column to show the changes observed ($Changed_{v1}$) and another to indicate the rule executed for the presented situation (one rule by line). The proposed rules are:

1https://github.com/anno4j/anno4j
2https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2631504/
1. **MergeAnnot**: This rule will be applied when two parts of a document, annotated with different concepts, can be put together and annotated with one (more specific) new concept only. For instance, in 2004AA, the texts “pregnancy” and “hypertension” were annotated with the concept codes D011247 and D006973, respectively. In 2005AA, text containing both terms, i.e., “pregnancy-induced hypertension” was annotated with the new concept D046110, see Table 2.

2. **IncreaseAnnot**: This rule increases the amount of information that can be annotated after the evolution of the underlying KOS. To do this, we compare the new label or attribute values of the candidate concept with the information surrounding the initial annotation (i.e., we take into account the prefix and suffix of the annotated text). Concretely, this action modifies the offset value in the annotation model and (if needed) the concept ID, e.g., \{D002403, cathepsin\} ↔ \{D056668, cathepsin l\}, see the second example on Table 2.

3. **ResurrectAnnot**: In some cases, one concept can be temporarily deleted from a KOS, leading to the deletion of the associated annotation. For instance, the annotation *chemiluminescence* in Table 2 was removed by a change in MeSH 2005AA. This rule allows the annotation to be re-activated when the concept is re-integrated to the KOS, e.g., the concept D017083 in MeSH 2006AA.

4. **PluralAnnot**: This rule verifies whether the change in the underlying concept or attribute value is due to a plural or singular (agglutination ↔ agglutinations). In this case, the change in the terminology does not imply a change in the impacted annotations since the semantics of the concept are not altered. Note that plurals are language-dependent rules and we are evaluating only for English KOS.

5. **ChangeConceptAnnot**: This rule changes the concept ID of the annotation due to the evolution of the concept. This situation arises when the label or the attribute value of the concept, used to create the annotation, is moved to another concept or used to create a new concept. For instance, concept D003704 changed to D057174 (referring to *Semantic dementia*) in MeSH 2009AA/2010AA.

6. **SplitAnnot**: This rule splits an existing annotation if the evolution of the underlying concept leads to the creation of two more precise annotations. For instance, the text “diabetic foot ulcers”, annotated in 2005AA with the MeSH code D017719, was split into two other annotations in 2006AA: D017719 (diabetic foot) and D016523 (foot ulcers), see Table 2.

7. **SuperClassAnnot**: This rule changes the concept ID to the superClass ID since no concept can be found to precisely maintain the annotation. It will also change the relation (i.e., “Equivalent” → “Is A”) between the concept and the annotation. For instance, after checking whether any of the previous rules were executed with the annotation *infective agents*, the last example in Table 2, instead of deleting the annotation, we propose to using superClass to annotate the text. Thus, *infective agents* is a kind of *other organism groupings*. Note that it is only possible if the formalism used to annotate the text follows our proposed formalism².

The sequence in which the rules are executed is important to assure the quality of the modified annotations. Based on the propositions of the annotation guidelines¹⁷, we established the following sequence: MergeAnnot, IncreaseAnnot, ResurrectAnnot, PluralAnnot, ChangeConceptAnnot, SplitAnnot, SuperClassAnnot. We ranked first the rules that increase the information of an annotation (i.e., MergeAnnot and IncreaseAnnot), as suggested in the guideline “Annotate the most specific concept that correctly describes the disease mention”. The next rules (ResurrectAnnot, PluralAnnot and ChangeConceptAnnot) are mainly related to the structure of the KOS and text. We started by ResurrectAnnot because changing the concept ID (ChangeConceptAnnot) would increase the complexity in identifying the restoration of the concept. The PluralAnnot rule is an exception, because it does not affect the other rules and can be placed anywhere in the sequence. The SplitAnnot was placed close to the end of our process due to the rare cases where it occurs as mentioned by Doğan et. al.¹⁷. It respects the following recommendation: “Annotate a disease mention using multiple concepts to logically describe the disease mention, using the “+” concatenator”. Finally, the SuperClassAnnot was positioned at the end of our process as an alternative to the removal of the annotation.

The precision of rules has some limitations. We observed that some exceptions could sensibly increase the complexity of the rules (and the time required to execute them). Thus, we decided to evaluate other potentially complementary
Table 2: Example of annotations computed by our rules. \((CP_{v0})\) concept code in specific year, \((Annot_{v0})\) annotated text, \((Prefix_{v0})\) prefix, \((Suffix_{v0})\) suffix, \(Changed_{v1}\) result of applied rule.

<table>
<thead>
<tr>
<th>(CP_{v0})</th>
<th>(Annot_{v0})</th>
<th>(Prefix_{v0})</th>
<th>(Suffix_{v0})</th>
<th>(Changed_{v1})</th>
<th>Rule</th>
</tr>
</thead>
<tbody>
<tr>
<td>D011247 in 2004AA</td>
<td>pregnancy</td>
<td>diabetes mellitus and hypertension</td>
<td>-induced hypertension. Apgars were</td>
<td>{D046110, pregnancy-induced hypertension}, 2005AA</td>
<td>MergeAnnot</td>
</tr>
<tr>
<td>D006973 in 2004AA</td>
<td>hypertension</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D002403 in 2009AA</td>
<td>cathepsin</td>
<td>responses [67], a l-like gene (ee049537) has</td>
<td></td>
<td>{D056668, cathepsin l}, 2010AA</td>
<td>IncreaseAnnot</td>
</tr>
<tr>
<td>D017083 in 2004AA</td>
<td>chemiluminescence</td>
<td>of western blot</td>
<td>were acquired</td>
<td>{D017083, chemiluminescence}, 2006AA</td>
<td>ResurrectAnnot</td>
</tr>
<tr>
<td>D000371 in 2009AA</td>
<td>agglutination</td>
<td>of antibodies. weakly reactive</td>
<td>required an adequate light</td>
<td>{D000371, agglutinations}, 2010AA</td>
<td>PluralAnnot</td>
</tr>
<tr>
<td>D003704 in 2009AA</td>
<td>Semantic dementia</td>
<td>frontotemporal dementia pnfa</td>
<td>?prion</td>
<td>{D057174, semantic dementia}, 2010AA</td>
<td>ChangeConceptAnnot</td>
</tr>
<tr>
<td>D017719 in 2005AA</td>
<td>diabetic foot ulcer</td>
<td>associated with</td>
<td>are recommended</td>
<td>{D017719, diabetic foot}{D016523, foot ulcers}, 2006AA</td>
<td>SplitAnnot</td>
</tr>
<tr>
<td>C50922 in 2009AA</td>
<td>infective agents</td>
<td>the most common</td>
<td>, the necrotic base of</td>
<td>{C14376, other organism groupings}, 2010AA</td>
<td>SuperClassAnnot</td>
</tr>
</tbody>
</table>
methods in order to improve the quality of our outcomes. In this sense, we selected two other methods: background knowledge and semantic change patterns. Details of each method can be found in \cite{13} and \cite{15}, respectively. We briefly introduce the main aspects of these two methods below:

**Background Knowledge (BK)**

The main idea behind BK is to use information inferred from external ontologies in order to discover the semantic relation between two successive versions of a concept \cite{13}. For instance, the evolution of the label “Magnoliophyta” to “Angiosperms” cannot be characterized only by considering the syntactic aspect. However, external “sources of knowledge” can tell us that these two terms are synonyms. We used the mappings between the concepts from different terminologies to determine the relation between these concepts. In our case, we used the mappings contained in Biportal \cite{18}. However, this approach can be extended to any external resource.

The BK algorithm (see algorithm 1) presents an overview of the whole process. Fig. 1 helps to illustrate how the algorithm works. The input of the algorithm is the concept ID ($C_s$), label ($L_s$), KOS target ($KOS_t$), and KOS source ($KOS_s$), e.g., (D019684, Magnoliophyta, MESH, \{SNOMEDCT, ICD-9-CM, NCIT\}). After initializing the variables (lines 1 to 2), our method queries external sources using the impacted annotation (line 3) label and stores the resulting concepts ($Request$). For instance, the concept 420928000, from SNOMEDCT, is one candidate. Only concepts candidates belonging to the source KOS ($KOS_s$) are kept (lines 4-5). Then, for each concept from $Request$, the mappings are collected (line 6). Only mappings to the target KOS ($KOS_t$) are kept (lines 7-10), they are the gray boxes in Fig. 1. From all candidates that satisfy the previous conditions, only the best candidate is selected to maintain our annotation (lines 11-13). The selection criteria is based on the semantic similarity (i.e., Tversky similarity) between the concept source and the concept target, i.e., ($C_s$ and $C_t$) from MESH.

**Algorithm 1:** Background Knowledge to find the link between 2 concepts

| Input: | Concept source $C_s$; Label $L_s$; Ontology source $KOS_s$; Ontology Target $KOS_t$ |
| Output: | Concept Target $C_t$ |
| MappingSet ← $\emptyset$ |
| Result ← $\emptyset$ |
| $Request ← getConceptsFromBK(L_s)$ |
| forall $cp ∈ Request$ do |
| | if ($cp ∈ KOS_s$) == TRUE then |
| | | MappingSet ← getMappings($cp$) |
| forall mapping ∈ MappingSet do |
| | target ← getConceptTarget($mapping$) |
| | if ($target ∈ KOS_t$) == TRUE then |
| | | Result ← target |
| forall obj ∈ Result do |
| | calSemanticDistance($C_s$, obj) |
| $C_t ← getHighSimilarity(Result)$ |
| return $C_t$ |

**Semantic Change Pattern**

Change Patterns are morphosyntactic modifications observed in attribute values of a concept, using linguistic-based features to identify the correlation between concepts over time. This technique has already been explored in the context of ontology mapping adaptation \cite{14} and we adapted it to the maintenance of annotations \cite{15}. It consists of a set of rules that allow the identification of the impact of evolving an ontology on the semantics of the concept.

For instance, the annotation, “Physiologic processes” produced using MeSH 2008AA, was impacted in 2009AA. This
Figure 1: Use of BioPortal as Background Knowledge

is due to a change in the attribute value in the definition of concept D010829 leading to “Physiological Phenomena”. The rules proposed in\textsuperscript{14} will determine which class of change it belongs to: \textit{total copy}, \textit{total transfer}, \textit{partial copy}, \textit{partial transfer}. We can conclude, for instance, that if changes in the ontology lead to a total transfer of information from one concept to a new one, than the annotation will probably change the source (or target) concept to this new one too. This is a simplification of inferences that can be used. To better understand the context and the complexity behind SCP, we advise reading\textsuperscript{14}.

Experimental Assessment

In this section, we introduce the method and material we have used to evaluate our approach. The experiments we have conducted consist of applying our approach to a set of annotated documents and comparing it to a corpus of reference representing an evolved version of the initial set of documents.

Material

Our annotation maintenance process takes as its input:

- a set of outdated annotations,
- the old and new OWL versions of the KOS used to generate the annotation.

In our experiments we have used: International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM), Medical Subject Headings (MeSH), National Cancer Institute Thesaurus (NCIt) and Systematized Nomenclature of Medicine - Clinical Terms (SNOMEDCT). We used versions 2009AA and 2010AA downloaded from the UMLS and transformed into OWL files. These KOS follow the formality expressed in the UMLS metathesaurus\textsuperscript{3}, e.g., \textit{(Concept A rdfs:superClassOf Concept B)} ; \textit{(Concept A skos:prefLabel term 1)}, etc. We used COnto-Diff\textsuperscript{10} to identify the changes in the new version of the KOS since these changes are strongly correlated with the validity of annotations\textsuperscript{2}.

Silver Standard

Since no annotation baseline generated with sequential ontologies versions exists, we had to build our own corpus of reference using the annotations produced in\textsuperscript{2} as a basic resource. To do this, we randomly selected 500 annotations generated with the 2009AA version of the four KOS (125 annotations from each KOS) and asked three experts to manually validate/correct the evolution of the 500 selected annotations, according to the 2010AA version of the corresponding KOS. Each expert validated 1/3 of the annotations and no discussions between them was organized. The consolidated

\textsuperscript{3}https://www.ncbi.nlm.nih.gov/books/NBK9685/
outcomes compose the silver standard, which can be downloaded from [http://www.elisa-project.lu/](http://www.elisa-project.lu/), look into menu publications/downloads. We adopt the term "silver" to indicate that our reference is based on only one viewpoint.

To measure the effectiveness of the proposed approach, we used classic well-known metrics from the literature, such as, precision, recall, F1-score, ROC curve, accuracy, false/true positives/negatives. For the sake of readability, we will present only three metrics in the table. But, we used all six metrics to investigate/understand two characteristics of our method: i) the capacity of our framework to detect impacted annotations after changing a KOS concept; and ii) the ability to correctly evolve those impacted annotations into consistent ones. In this case, consistency means equivalency with the silver standard. We measured the efficiency of the rules alone, the BK alone, and the SCP alone and the efficiency of the combined techniques in order to determine whether they complement each other.

### Results

When applying the three annotation maintenance methods (BK, SCP, Rules) to our dataset we can observe a significant difference in the results (based on the six criteria used). For instance, as shown in Table 3, all three methods can provide good precision, but there is a significant variation [0, 0.98] regarding the recall. In the first line of Table 3, we present the precision, recall, and F1-Score resulting from applying the BK method to four different subsets of our initial dataset (ICD-9-CM, MeSH, NCIt, and SNOMEDCT). We also evaluate the consequence of combining the methods (2-by-2, and all together). For instance, the fourth line of the table presents the results of combining BK and Rules methods, while the seventh line shows the results of combining all three methods.

The goal of this first set of experiments was to evaluate whether the methods (or a combination of them) provides satisfactory quality (in terms of F1-Score) to determine whether an impacted annotation will evolve or not. Note that we are not yet evaluating whether the annotation evolved correctly (this is part of the second evaluation step). A quick analysis shows that all methods can accurately identify some of the evolving annotation, but not all. From a practical point of view, whether an error of 2% is acceptable for the domain, then KOS engineers can trust our best method to identify annotations that will change (i.e., the minimal observed precision was 98%). However, the recall can be significantly different according to the dataset and the method adopted. We detail the reasons for this in next section. We would like to highlight that our Rules method had an impressive performance, obtaining in some cases a F1-Score of 99%.

The second evaluation process consists of applying methods to select which adaptation actions can make the annotation evolve correctly, and compare the outcomes with the silver standard. The goal is to measure how precise our recommendations are. Table 4 describes the performance of each method regarding the four different datasets. Each experiment is represented by an Area Under the Curve (AUC) value, giving the probability that a randomly selected instance will correctly be adapted by our method19.

The AUC values of the analyzed methods vary according to the dataset. A quick analysis in Table 4 shows that combining all methods provides slightly better results than applying only one of them. Furthermore, SCP shows the

### Table 3: Precision (P), Recall (R) and F1-Score (F1) of impacted annotations computed using three different methods (BK, SCP, Rules) and the combination of them. The red and orange colors indicate low and medium recall, respectively.

<table>
<thead>
<tr>
<th>Method</th>
<th>ICD9CM</th>
<th>MeSH</th>
<th>NCIT</th>
<th>SNOMEDCT</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P</td>
<td>R</td>
<td>F1</td>
<td>P</td>
</tr>
<tr>
<td>BK</td>
<td>1</td>
<td>0.16</td>
<td>0.28</td>
<td>1</td>
</tr>
<tr>
<td>Rules</td>
<td>1</td>
<td>0.98</td>
<td>0.99</td>
<td>0.98</td>
</tr>
<tr>
<td>SCP</td>
<td>1</td>
<td>0.08</td>
<td>0.15</td>
<td>1</td>
</tr>
<tr>
<td>BK &amp; Rules</td>
<td>1</td>
<td>0.98</td>
<td>0.99</td>
<td>0.98</td>
</tr>
<tr>
<td>BK &amp; SCP</td>
<td>1</td>
<td>0.19</td>
<td>0.32</td>
<td>1</td>
</tr>
<tr>
<td>Rules &amp; SCP</td>
<td>1</td>
<td>0.98</td>
<td>0.99</td>
<td>0.98</td>
</tr>
<tr>
<td>CombineAll</td>
<td>1</td>
<td>0.98</td>
<td>0.99</td>
<td>0.98</td>
</tr>
</tbody>
</table>
Table 4: AUC values of developed heuristics used to maintain annotations. The red and blue color highlight the lower and higher values for each dataset, respectively.

<table>
<thead>
<tr>
<th>Method</th>
<th>ICD9CM AUC</th>
<th>MeSH AUC</th>
<th>NCIT AUC</th>
<th>SNOMEDCT AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>BK</td>
<td>0.613</td>
<td>0.554</td>
<td>0.663</td>
<td>0.708</td>
</tr>
<tr>
<td>Rules</td>
<td>0.899</td>
<td>0.850</td>
<td>0.721</td>
<td>0.833</td>
</tr>
<tr>
<td>SCP</td>
<td>0.593</td>
<td>0.550</td>
<td>0.606</td>
<td>0.500</td>
</tr>
<tr>
<td>BK &amp; Rules</td>
<td>0.915</td>
<td>0.863</td>
<td>0.721</td>
<td>0.833</td>
</tr>
<tr>
<td>BK &amp; SCP</td>
<td>0.601</td>
<td>0.554</td>
<td>0.663</td>
<td>0.708</td>
</tr>
<tr>
<td>Rules &amp; SCP</td>
<td>0.895</td>
<td>0.838</td>
<td>0.731</td>
<td>0.833</td>
</tr>
<tr>
<td>CombineAll</td>
<td>0.923</td>
<td>0.863</td>
<td>0.731</td>
<td>0.833</td>
</tr>
</tbody>
</table>

lowest AUC values of all heuristics. We also verified that the AUC has significant differences between the KOS, like those between ICD-9-CM and NCIt. Detailed explanations on these observations are provided in the next section.

Discussions

The analysis of annotation evolution in the healthcare domain is an understudied topic. As explained in the introduction section, several works propose the automatic detection of inconsistent annotations, but few of them address the automatic correction of inconsistent annotations. The work presented in this paper shows that some annotations can be preserved/adapted after the evolution of the KOS used to generate the annotations. Three methods were proposed: Background Knowledge, Semantic Change Patterns, and Domain Specific Rules. The outcomes presented in the results section demonstrate that we can obtain high AUC by applying these methods together in the automatic maintenance of annotations or to support domain experts in this activity.

When analyzing each method, we observed that BK contributes to the precision of the annotation changes. The main characteristic of BK is that it depends on the richness of information in other sources (e.g., ontologies with overlapping concepts). Furthermore, BK can provide unaligned mappings to past KOS versions leading to another phase to filter inconsistent results. This limitation of the method generates a low recall (but good precision) of results. We estimate that it can be an interesting complementary method for the maintenance process. Another aspect that can be deduced from the experiments is the dependency of the BK method on the expressivity and consistency of the KOS. For example, MESH D002544 has as a synonym the concept labels that are considered siblings in other KOS (e.g., “Cerebral infarct left hemisphere” SNOMEDCT 362323007 and “Cerebral infarct right hemisphere” SNOMEDCT 362322002), leading to loose information when the system follows the KOS mappings that cross MeSH. We also observed that SNOMEDCT and ICD-9-CM allow equal “labels of concepts” (e.g., in SNOMEDCT the concept “diverticulosis” has the following codes 31113003, 397881000, and 68047000), which can lead the system to select the wrong concept to replace the impacted annotations, and necessitating an additional disambiguation phase. Another limitation of BK is that we can only find the last version of KOS in the Bioportal (i.e., from 2016), but our experiments use documents annotated with the version 2009AA and 2010AA. Versionning is an aspect that is not yet integrated to Bioportal, but it deserves to be considered in future.

The analysis of SCP also shows a good precision and low recall. The reason here is that SCP considers only change between concepts that are in the same neighbourhood (i.e., siblings, super-, and sub-concepts). Thus, changes that move the concept to other branches of the KOS are not included, leading to an increased number of false negatives. For SNOMEDCT, Table 4, we did not observe any cases with SCP in our dataset. Since, the data used was randomly selected, we consider that it was a coincidence. We did our analysis based on the results coming from the other KOS. However, this heuristic is able to cover cases where the Rules or BK do not work. For instance, the annotation “ubiquitin carboxyl-terminal hydrolase” NCIt C21490 correctly evolved to “ubiquitin carboxyl-terminal hydrolase BAP1”. Thus, the AUC value to NCIt in Table 4 is higher in Rules & SCP than that compared to BK and its combinations (lines 1, 4, and 5).

Domain-specific rules are defined to describe frequent patterns of changes, and it is expected to generate outcomes from them with good precision and recall. This was the case for the rules proposed in our experiments. However,
our rules do not cover all annotation evolution cases perfectly; some ambiguities are still observed. The proposed rules are not very precise when the annotated text and the concept label (or synonyms) do not have an exact match. For instance, the text “dysarthria” was annotated in 2009AA with the concept 784.5 (Other speech disturbance); in 2010AA, a new sub-concept 784.51 (“dysarthria”) was created. The rules were unable to determine that the specialized concept was more appropriated to the annotation. This evolution was correctly proposed by the BK method. Such cases of misclassification are more frequent in NCIt because we observed fewer annotations with an exact match and a lower number of mappings from/to this KOS in Biportal.

Another aspect to highlight, is that depending on the context in which the maintenance methods were used (e.g., high expressive KOS), there are considerable differences in the sets of results. We also observed that a combination of methods can be used for a more complete set of evolution situations, as in following:

- SCP and BK methods show low complementary results to identify whether the KOS evolution impacts the annotations, but an improvement was observed by combining the methods to identify the correct evolution of the annotation.
- On one hand, Rules increase the amount of corrected annotations of all BK and SCP analyzed cases. On the other hand, BK and SCP have a minimum effect on the rules for the first set of experiments and demonstrate the ability to improve the identification of correct evolution of the annotation (second set of experiments).
- Compared to the pairwise combination of the methods (BK & Rules, SCP & Rules), applying the three methods together improves (or at least keeps the same) the AUC of the correct evolution of the impacted annotations (second set of experiments).

Conclusion

The work presented in this paper shows the possibility of having annotation maintenance tools that can keep tracking KOS evolution and measure the impact that it will have on the set of annotated documents. Moreover, it also demonstrates that automatic correction/adaptation of annotations can reach a reasonable reliability rate. But, it is important to highlight that the role of human beings is still determinant to assure the quality of the annotations in critical scenarios, as observed in the biomedical domain. Our approach contributes to a more generic objective that intends to define methods and formalism to improve the annotation task in order to better support the annotation maintenance. In previous work, we evaluated the annotation formalism and proposed some extensions to close the identified gaps. In future work, we will evaluate the performance of our approach in a larger temporal window and search for methods that can generate additional rules from the data.

References


4http://www.elisa-project.lu/


Detecting Evidence of Intra-abdominal Surgical Site Infections from Radiology Reports Using Natural Language Processing

Alec B. Chapman, BM¹, Danielle L. Mowery, PhD¹,², Douglas S. Swords, MD³,
Wendy. W. Chapman, PhD¹,², Brian T. Bucher, MD¹,²
¹Biomedical Informatics, University of Utah School of Medicine, Salt Lake City, UT;
²IDEAS Center, George E. Wahlen Veterans Affairs Medical Center, Salt Lake City, UT;
³Pediatric Surgery, University of Utah School of Medicine, Salt Lake City, UT

Abstract

Free-text reports in electronic health records (EHRs) contain medically significant information - signs, symptoms, findings, diagnoses - recorded by clinicians during patient encounters. These reports contain rich clinical information which can be leveraged for surveillance of disease and occurrence of adverse events. In order to gain meaningful knowledge from these text reports to support surveillance efforts, information must first be converted into a structured, computable format. Traditional methods rely on manual review of charts, which can be costly and inefficient. Natural language processing (NLP) methods offer an efficient, alternative approach to extracting the information and can achieve a similar level of accuracy. We developed an NLP system to automatically identify mentions of surgical site infections in radiology reports and classify reports containing evidence of surgical site infections leveraging these mentions. We evaluated our system using a reference standard of reports annotated by domain experts, administrative data generated for each patient encounter, and a machine learning-based approach.

Introduction

Health care-associated infections (HAIs), such as surgical site infections (SSIs), affect one in every twenty hospitalized patients and account for $10 billion dollars in potentially preventable health care expenditures annually¹. Identification and reporting of HAI after selected procedures is a required hospital quality measure by several federal agencies. In addition, detection of HAIs in a timely manner may alter treatment courses, reducing hospital costs and improving patient care.

Information indicating an HAI may be recorded in a variety of locations in the electronic health record (EHR): physician notes, radiology reports, and microbiology reports. One common method of documenting HAIs is using administrative data such as International Classification of Diseases (ICD). Due to the lack of specificity in administrative billing codes and timing of their assignment, these codes cannot reliably be used to identify HAIs in an accurate and timely manner. Furthermore, administrative data are often insufficient and incomplete due to the underutilization of structured data fields, lack of standardization, and unknown quality of clinical data². Relevant information may instead be documented in free-text reports and text fields of the EHR. However, these reports are unstructured and the information contained within them must be extracted into a structured, computable format in order to utilize them. The traditional method of extraction is manual chart review, such as in the National Surgical Quality Improvement Program (SQIP). Specifically, these programs leverage trained surgical case reviewers to manually extract data from EHRs to detect and to report the development of HAIs³. However, manual abstraction is expensive, labor-intensive, and time-consuming. More efficient methods must be considered to extract this information.

Natural language processing (NLP) systems can be developed to automatically extract HAI information from textual data as an efficient and effective method of HAI detection. Melton et al. trained the NLP system MedLEE to identify adverse events (AEs) from discharge summaries⁴. MedLEE uses grammatical rules to identify concepts and their semantic context, which are then mapped to controlled vocabularies⁵. MedLEE outperformed manual review for AEs, but their study included only one postoperative HAI (wound infection), which the system achieved a low precision (0.34) for detecting wound infections. Their overall recall (0.28) and precision (0.45) were fairly low, while their specificity was quite high (0.98). Penz et al. combined MedLEE and phrase-matching techniques that utilize regular
expressions to identify AEs related to the placement of central venous catheters. By combining these two methods, they achieved good recall (0.72), specificity (0.80), and precision (0.64). False positives occurred due to indications of risk, in which the doctors discussed the hypothetical risks of a catheter placement. Some mentions were also missed due to misspellings in the text. The authors compared the NLP performance to that of ICD-9 and CPT codes. They concluded that NLP techniques are far more sensitive than using administrative data, which captured less than 11.5% of all central line placements.

Several studies have specifically focused on the effectiveness of NLP at detecting surgical AEs. FitzHenry et al. applied NLP methods to surveil for the presence multiple surgical complications, including pneumonia, wound infection, and urinary tract infections. They achieved moderate to high performance for detecting each type of complication measured: pneumonia (recall: 0.90 and specificity: 0.80); wound infection (recall: 0.63 and specificity: 0.77); and urinary tract infections (recall: 0.80 and specificity: 0.95). Similarly, Murff et al. compared the performance of NLP techniques in the detection of surgical AEs to the performance of using patient safety indicators. The NLP system performance ranged from moderate to high for detecting pneumonia (recall: 0.64 and specificity: 0.95) and sepsis (recall: 0.89 and specificity: 0.94). Their system had a higher recall and lower specificity than administrative data. Both of these studies were conducted in Veteran Affairs (VA) centers and used a gold standard of manual review performed by a VA-SQIP nurse.

The main limitation of these studies is the lack of additional clinically relevant information (such as anatomic location) in development of these detection systems. The goal of this study was to evaluate the performance of an NLP algorithm in the extraction of evidence of intra-abdominal surgical site infections that specifies the anatomic locations of these SSIs. The overall hypothesis is that NLP methods will achieve a level of accuracy equal to or greater than that of manual chart review and greater than that of administrative data. Our objective was (1) to develop a knowledge base that supports deep semantic extraction of evidence of intra-abdominal surgical site infection mentions and their anatomic locations and (2) to test a knowledge base-powered NLP algorithm's ability to make report-level predictions of whether evidence of an intra-abdominal surgical site infection is present. We evaluated our NLP algorithm's performance using a reference standard of expert-annotated text and compared its performance against other baseline approaches such as administrative data and machine learning.

Methods

The overall approach for this study included four steps: (1) identifying surgical patients and collecting their associated EHR data; (2) annotating text data and developing a knowledge base; (3) detecting mentions of SSIs and classifying reports for the presence of a SSI; and (4) evaluating automated methods against the reference standard annotations for SSIs. Institutional review board approval (IRB) was obtained.

Patient Identification and Dataset Extraction

We identified a cohort of patients undergoing gastrointestinal surgery from MIMIC-III Critical Care Database and collected their de-identified EHR data. Specifically, we collected administrative information (admission date, procedure date, ICD-9 diagnosis codes, and Current Procedural Terminology (CPT) codes). We limited the study to patients undergoing gastrointestinal surgery covering topics of esophageal, gastric, small intestine, large intestine, liver, pancreas, abdominal wall, ovarian, uterus, kidney procedures based on 28 CPT codes (e.g., CPT 49000=Under Incision Procedures on the Abdomen, Peritoneum, and Omentum). We limited our review and process to only include computed tomography (CT) reports due to their standardized structure and reliable reporting of findings. From this cohort, all the CT reports within 30 days after the surgical procedure were collected for our dataset.
Annotation Study

We conducted an annotation study using our development dataset to generate a reference standard of mention-level and report-level classifications using the full dataset. Two domain experts, a surgeon (author BB) and a surgical resident (author DS), annotated these reports using the annotation tool, the extensible Human Oracle Suite of Tools (eHOST)10. In these reports, evidence for a SSI is commonly described by phrases such as “fluid collection” or “abscess”51. To confirm the mentions (subsequently referred to as “fluid collections”) we were referencing findings related to the gastrointestinal tract surgery, we limited our search to detecting mentions co-occurring in the same sentence with descriptions of a gastrointestinal anatomic location such as “liver” or “abdomen”. The annotators annotated sentences containing mentions of fluid collection as one of the following classes: positive evidence of fluid collection; negated evidence of fluid collection; indication for exam; and annotated mentions of anatomic location within these sentences marked as positive evidence of fluid collection (Table 1). After annotation was complete, we then randomly split our dataset into development set (n=565 documents for 409 patients) and test set (n=100 documents for 96 patients).

Table 1. Mention-level annotation schema; Bolded=terms indicating fluid collection; italics=contextual terms indicating anatomic location, negation, uncertainty, historicity, and indication.

<table>
<thead>
<tr>
<th>Class Name</th>
<th>Definition</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive Evidence of Fluid Collection</td>
<td>A positive mention of fluid collection mentioned as potentially occurring at some point in time in the same sentence as a relevant anatomic location.</td>
<td>“Fluid collection is seen in the abdomen.”</td>
</tr>
<tr>
<td>Negated Evidence of Fluid Collection</td>
<td>A clear statement ruling out the possibility of fluid collection.</td>
<td>“There is no evidence of fluid collection in the abdomen.”</td>
</tr>
<tr>
<td>Indication for Exam</td>
<td>A phrase indicating that the purpose of the exam is to check for fluid collection.</td>
<td>“PURPOSE OF EXAM: Rule out abscess.”</td>
</tr>
<tr>
<td>Anatomy</td>
<td>A part of the body in the GI tract that occurs in the same sentence as a mention of fluid collection.</td>
<td>“Hematomas are seen around in the right lower quadrant.”</td>
</tr>
</tbody>
</table>

Mention Detection

To automatically identify and classify evidence of surgical site infection from each report according to our annotation schema, we developed an NLP system called SSI-Detect. We created a module called Fluid Collection Finder (fcFinder) that targets and classifies fluid collections*. fcFinder leverages the pyConText library, an adaptation of the ConText algorithm. ConText utilizes trigger terms and termination points to extract and associate contextual features such as negation, temporality and experiencer to targeted terms12. Specifically, pyConText extends the ConText algorithm by leveraging NetworkX digraphs to relate targets with modifiers and supports report-level classification13. pyConText has been adapted to support several use cases: to identify pulmonary embolisms, to classify cancer history, and to flag reports with significant carotid stenosis findings13-15. Using the development set, we created a knowledge base lexicon of 46 fluid collection targets and 128 anatomic location modifiers*. fcFinder utilizes this knowledge base to implement the pyConText algorithm and identify relationships between contextual and anatomical modifiers and fluid collection targets. After creating these relationships, fcFinder applies rules to classify each sentence containing a mention of a fluid collection finding as being positive evidence of fluid collection, negated evidence of fluid collection, or indication for exam. Figure 1 depicts an example pyConText markup sentence with a target and its associated modifiers of negation and anatomic location.
Figure 1. An example of negated evidence of fluid collection. The target, “abscess”, is modified by “no evidence”, a forward-direction negation modifier, and “abdomen”, a bidirectional anatomic modifier.

Evaluation

We computed the performance of fcFinder by comparing the mention-level findings of positive evidence, negative evidence, and indication with the annotated reference standard. Definitive, historical, and probable evidence were all considered positive evidence mentions and could be matched with one another. Two findings were considered a match if their spans overlapped and they had the same class. We defined a true positive (TP) as a correctly identified mention, a false positive (FP) as a spuriously flagged mention, and a false negative (FN) as a missed mention. We evaluated fcFinder’s performance overall and within each class using F1-score (Eq. 1), recall (Eq. 2), and precision (Eq. 3).

\[
\text{(Eq. 1) } F1 = \frac{2TP}{2TP + FP + FN} \quad \text{(Eq. 2) } \text{Precision} = \frac{TP}{TP + FP} \quad \text{(Eq. 3) } \text{Recall} = \frac{TP}{TP + FN}
\]

Report Classification

Using the mention-level annotations created by fcFinder, we classified a report as either fluid collection-present or fluid collection-not present using a python module called fcClassifier. If a report contained at least one positive mention-level annotation, the report was labeled as fluid collection-present. If it contained only negative or indication findings, or no findings at all, then it was labeled as fluid collection-not present. Additionally, we assessed fcClassifier’s performance using three configurations: 1) without context; 2) without anatomy; and 3) with all modifiers. First, we implemented fcClassifier without context using only the targets without any linguistic (negation, historicity, indication) or semantic (anatomy) modifiers. This could essentially be considered a keyword search. We then tested fcClassifier using linguistic modifiers, but no semantic modifiers. This allowed us to evaluate the importance of excluding negated mentions of fluid collection and of requiring the use of anatomic location in order to accurately classify reports. Finally, we tested fcClassifier using both linguistic and semantic modifiers.

Baselines

We generated two baseline approaches for classifying each report comparison against fcClassifier: 1) ICD-9 codes and 2) n-grams. First, we developed a simple ICD-9 classifier that predicts fluid collection-present for each report from a patient encounter encoded with ICD-9 code 998.53. For any report from a patient without this ICD-9 code, it predicts fluid collection-not present. We then experimented with several machine learning models utilizing n-grams (1-4 window word features) from the development set, excluding words occurring in an English stopwords list and word features occurring less than twice. To reduce the likelihood of overfitting, we applied the following feature selection strategy. We determined an upper bound threshold of n features to consider approximated using the square root of total training reports as 25 features. We ranked the informativeness of these 25 features by p-value using Chi-square with 5-fold cross-validation. We evaluated each classifier’s performance at intervals of 5 features (5, 10, 15, 20, 25 features) and selected the highest performing classifier according to F1-score with the fewest number of features. We experimented using several machine learning classification models, including a naive bayes, random forest classifier, and linear support vector machine (SVM). For this study, we applied the most predictive classifier,
the linear SVM, leveraging the 15 highest-ranked word features to the test set. In Table 2, we compared these n-gram features to fcFinder’s knowledge base and determined some terms are shared with our rule-based approach, giving us confidence this could be a reasonable baseline approach.

Table 2. 15 top-ranked features for n-gram classifier. **Bolded**=shared terms with fcFinder’s knowledge base.

| Collection, fluid, fluid collection, hematoma, cm, drain, collections, drainage, fluid collections, subcapsular, pigtail, catheter, rim, noted, extravasation |

**Evaluation**

We compared the performance of each baseline approach as well as fcClassifier and its various configurations against the testing set reference standard using accuracy (Eq. 4), F1-score, recall, and precision. We defined true positives (TP) as correctly classified reports as fluid collection-present and true negatives (TN) as correctly classified reports as fluid collection-not present. Specifically, we aimed to determine how well each approach could predict a report as fluid collection-present or fluid collection-not present in the test set. We assessed whether the differences in performance between the fcClassifier using all modifiers and all other approaches were statistically different using McNemar’s test.

(Eq. 4) \[ \text{Accuracy} = \frac{TP + TN}{\text{All reports}} \]

**Results**

**Annotation Study**

We calculated the inter-annotator agreement between annotators over 4 batches. There were a total of 249 mention-level annotations in the test set (Table 3). The most prevalent mention-level annotations were positive evidence of fluid collection (57%; 142 of 249 annotations) followed by indication for exam (32%; 80 of 240 annotations). The most prevalent report-level classifications were fluid collection-not present (69%). F1-scores were high and consistent for mention-level classes ranging from 0.88 to 0.89 and near excellent for report-level classification (0.96).

Table 3. Inter-annotator Agreement in Test Set using F1-score (surrogate for Cohen’s Kappa).

<table>
<thead>
<tr>
<th>Mention-level</th>
<th>Number of Annotations (%)</th>
<th>F1-scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>249 (100%)</td>
<td>0.91</td>
</tr>
<tr>
<td>Positive evidence of fluid collection</td>
<td>142 (57%)</td>
<td>0.89</td>
</tr>
<tr>
<td>Negated evidence of fluid collection</td>
<td>27 (11%)</td>
<td>0.88</td>
</tr>
<tr>
<td>Indication for exam</td>
<td>80 (32%)</td>
<td>0.88</td>
</tr>
</tbody>
</table>

**Report-level**

| Fluid collection status         | 100 (100%)                | 0.96      |
**Mention Detection**

We evaluated fcFinder’s ability to identify and classify mentions of fluid collection from a reference standard using F1-score, recall, and precision (Table 4). Results were comparable between the development and test datasets (not shown). The algorithm performed with high F1-scores ranging from 0.87 to 0.90 for each category. Recall and precision was comparable (+/-2 points) for overall and for evidence of fluid collection. Recall was notably lower (-18 points) than precision for indication for exam. In contrast, recall was higher (+9 points) than precision for negated evidence of fluid collection.

**Table 4.** Mention-level comparison of fcFinder against the reference standard in the test test. **Bolded**=highest score for each metric.

<table>
<thead>
<tr>
<th></th>
<th>F1-score</th>
<th>Recall</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>0.90</td>
<td>0.89</td>
<td>0.91</td>
</tr>
<tr>
<td>Positive evidence</td>
<td>0.91</td>
<td>0.92</td>
<td>0.90</td>
</tr>
<tr>
<td>Negated evidence</td>
<td>0.91</td>
<td>0.96</td>
<td>0.87</td>
</tr>
<tr>
<td>Indication for exam</td>
<td>0.87</td>
<td>0.79</td>
<td>0.97</td>
</tr>
</tbody>
</table>

**Report Classification**

The prevalence of fluid collections in the development set (fluid collection-present: 42% and not present: 58%) was slightly skewed to positive cases compared to the test set (fluid collection-present: 31% and not present: 69%). We evaluated fcClassifier's ability to classify reports containing positive mentions of fluid collection compared to two baseline approaches: ICD-9 codes and n-grams (Table 4). ICD-9 codes achieved moderate performance (F1: 0.48; recall: 0.45). The use of targets without context for report classification resulted in improvement in classification (F1: 0.66; recall: 1.0). However, this improvement in recall results in a reduction of precision (0.49). fcClassifier without anatomic references improves precision (0.64) without sacrificing recall (1.0). fcClassifier with anatomic references performs with the highest F1 (0.88) and precision (0.82) with high recall (0.93). A comparison of each baseline to fcClassifier using McNemar’s test suggests this improvement is not by chance alone.

**Table 5.** Report-level classification of fcClassifier against the test set reference standard. **Bolded**=best performance for each metric.

<table>
<thead>
<tr>
<th>Classification Approach</th>
<th>Accuracy</th>
<th>F1-score</th>
<th>Recall</th>
<th>Precision</th>
<th>P-value vs. fcClassifier</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICD-9 codes</td>
<td>0.72</td>
<td>0.48</td>
<td>0.45</td>
<td>0.52</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>SVM n-grams</td>
<td>0.86</td>
<td>0.76</td>
<td>0.71</td>
<td>0.81</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>fcClassifier w/o context</td>
<td>0.70</td>
<td>0.66</td>
<td>1.0</td>
<td>0.49</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>fcClassifier w/o anatomy</td>
<td>0.88</td>
<td>0.79</td>
<td>1.0</td>
<td>0.64</td>
<td>0.002</td>
</tr>
<tr>
<td>fcClassifier</td>
<td><strong>0.95</strong></td>
<td><strong>0.88</strong></td>
<td>0.93</td>
<td><strong>0.82</strong></td>
<td>–</td>
</tr>
</tbody>
</table>
Discussion

We developed an NLP algorithm called fcFinder to automatically detect fluid collections from radiology reports. Our system utilized lexical and semantic features to determine the context of mentions of fluid collections. Based on the mention-level findings, we then classified reports as either fluid collection-present or not present. Our system performed well at both the mention and report levels, outperforming all compared methods.

Error Analysis

We examined instances where incorrect mention level-annotations caused misclassification of reports. Most instances of misclassification were caused by terms that were missing in our lexicon. The errors made by the system can be grouped into the following categories:

- **False positive target**: Our lexicon included the literal “collectin”, which was meant to match misspellings of “collection”, but also matched “fluid collecting”, which is not equivalent to a fluid collection. This resulted in an incorrect positive evidence of fluid collection annotation, which caused the document to be incorrectly classified as fluid collection-present.

- **Missing anatomic location modifier**: We did not have the literal “postoperative site” in our lexicon, resulting in a false classification of fluid collection-not present. We also deliberately excluded the literal “subcutaneous” as an anatomic modifier, due to its lack of specificity, but it was once used by the annotators, leading to another false negative classification.

- **Missing pseudoliteral**: Our lexicon included a number of “pseudoliteral” terms that would explicitly exclude phrases that could lead to false positives, such as “collection of gas” or “subpleural collection”. One report was incorrectly classified because “collection of intraluminal gas” was not included in our lexicon of pseudoliterals. A similar error was caused by the phrase “pockets of gas”.

- **Abbreviation**: A number of mistakes were made due to clinicians' use of a question mark, “?”, as an abbreviation for indication for exam. This was not included in our lexicon and several phrases such as “?absecess” were not annotated as indication for exam. This is an example of one of the biggest challenges facing clinical NLP, which is the informal and inconsistent nature of clinical text.

- **Reference standard problem**: A small number of mismatches were due to mistakes made by the annotators in the reference standard.

Key Findings

**Contextual Features**

Overall, fcFinder performed well with each mention-level class: positive evidence (F1: 0.91), negated evidence (F1: 0.91), and indication for exam (F1: 0.87). This allowed fcClassifier to achieve high results with report classification (accuracy: 0.95; F1: 0.88). fcClassifier's performance demonstrates the importance of linguistic and semantic context. ICD-9 codes, which provide information at a patient level and do not consider any information that is contained within a report, performed much worse at report classification (F1: 0.48). Searching for target concepts without taking into account any context resulted in a perfect recall (1.0), but a very low precision (0.49). When utilizing linguistic and semantic context, fcClassifier maintained a high recall (0.93) while achieving a much higher precision (0.82). This high precision could enable a more accurate selection of reports for review, greatly reducing the amount of work needed to evaluate and detect surgical site infections. This demonstrates the value of including contextual features when extracting information from clinical reports.
Anatomic Location Modifiers

Our intention was to identify only reports that contained fluid collections in the site of recent gastrointestinal surgeries. To do this, we developed a knowledge base of 128 anatomic location modifiers. The use of these modifiers enabled us to exclude reports that contained mentions of other collections, such as pleural or sternal collections, but no gastrointestinal fluid collections. The importance of anatomical designation is shown by the decrease in overall classification performance (~9 points) when anatomic modifiers were excluded from the criteria for mention-level findings. The significant increase in both recall and precision using anatomic modifiers is a valuable contribution towards the effective and efficient identification of surgical site infections. This knowledge base should be expanded to other anatomic regions if utilized in other settings.

N-gram Classifier

The best-performing baseline model was the n-gram SVM classifier (F1: 0.76). Despite not using any contextual features, this model achieved fairly high accuracy (0.86) and precision (0.81), and reasonable F1 (0.76) and recall (0.71). While the SVM's precision was comparable to fcClassifier's (0.81 vs. 0.82), its recall was much lower (0.71 vs. 0.93). Given the top 15 word features used by the n-gram model, we can conclude that the model does not take into account any linguistic (negation, historicity) or semantic (anatomic locations) context. However, the presence of procedural terms (“drainage”, “catheter”, “drain”) suggests the model includes situational context from the narrative, namely the treatment of fluid collection, which might be equally useful for asserting the presence of fluid collection. Future work could combine situational with linguistic and semantic context. Additionally, a system that utilizes structured data, such as ICD-9 codes, in combination with NLP could achieve even better results than those we report here.

Value of Rule-based Methods

This study demonstrates the potential of rule-based NLP to offer a semantically rich and detailed narrative of patient care. Techniques such as administrative data and machine learning offer very little detail beyond binary classification of a report. Rule-based NLP methods, such as those utilized by fcFinder, can extract much more detailed information. Here we showed the value of leveraging both linguistic and anatomic location modifiers to restrict SSI findings to a particular anatomic region. Similar methods can also model the administration and effectiveness of treatment in conjunction with the change of severity and course of a disease over time, without requiring much extra effort in developing new models to address each new question. This can offer a much more detailed and specific review of patient care, and can also enable a generalized approach to evaluating quality of treatment.

Comparison to Other Works

By conventional standards, our system achieved acceptable results. No other study has specifically aimed to extract mentions of fluid collection as evidence of SSI, but previously studies have successfully extracted other evidence of SSIs. Melton et al.’s algorithm, which utilized MedLEE, achieved a detected postoperative wound infection with a precision of 0.34. The algorithm implemented by FitzHenry et al. to detect wound infection achieved a recall of 0.63 and specificity of 0.77. Murff et al. report a recall of 0.89 and specificity of 0.94 in the detection of sepsis. We achieved a recall of 0.92 and precision of 0.90 in the detection of positive mentions of fluid collection. Our algorithm introduces the novel technique of connecting targets with anatomic location modifiers. This technique contributed to our overall performance and was a unique method of using NLP to detect a surgical AE. While some studies have evaluated the extraction of anatomic site of a condition, to our knowledge, no studies have done so in connection to surgical site infections or other surgical adverse events.

Limitations and Future Work

There are several limitations to our study. We used reports from the MIMIC-III Critical Care Database. These reports were written for surgical patients in intensive care. Because of the severity the patient’s condition, the prevalence of surgical site infections is higher than what we would expect to see in the general population. Reports
were then more likely to contain mentions of our targets, fluid collection. This may not directly generalize to the general population, where we would expect the prevalence of surgical site infection to be ~8-10% of patients. In this testing sample, 31% of notes were found to have fluid collections-present. However, the MIMIC-III dataset is openly available and the work that we have done with it can be reproduced and compared against. Furthermore, in the future, we will apply these methods to a more general population which could more broadly validate the results of this study.

Another limitation is the focused scope of this project. We looked only at radiology reports, specifically including only CT reports because of their relatively consistent structure and content. We focused on one specific anatomic region, and our targets included a relatively small lexicon of a single clinical concept, fluid collection, as evidence of surgical site infection. The CDC lists three additional criteria for defining an organ/space surgical site infection: purulent drainage from a drain, organisms found in a culture, and diagnosis of a surgical site infection by a surgeon or attending physician18.

Despite the relatively narrow focus of this project, it has proven the utility of a number of system features that can be applied in future studies and future implementations of SSI-Detect. In particular, this study has shown how a knowledge base of anatomic modifiers combined with NLP contextual tools results in a much higher precision and much richer semantic detail. This study is the first step in a larger framework of applying NLP methods to identify surgical adverse events. Future work should focus on expanding the knowledge base which we developed for this study, improving the performance of our system on non-enriched text, conducting a data analysis across various report types to test the generalizability of our methods, and expanding these methods to address other anatomic regions and clinical problems.

Conclusions

The automated detection of adverse surgical event has the potential to revolutionize how providers and hospitals detect and report quality measures. Currently, these efforts rely on manual chart review which limit the scalability and generalizability of quality measurement activities. In addition, leveraging automated detection into clinical decision support services may lead to better surveillance and the potential to improve patient care. We executed automated detection of evidence of intra-abdominal surgical site infections using a rule-based NLP system that achieved accuracy similar to manual chart review. In addition, this system outperformed other approaches using administrative data or SVM machine learning techniques. Future work in this field will seek to broaden our NLP approach and leverage addition types of structured and unstructured healthcare data to the detection of postoperative adverse events.

Acknowledgements

We thank our anonymous reviewers for their invaluable feedback on this manuscript.

* The source code and lexicon for fcFinder is openly available at https://github.com/abchapman93/fcFinder.

References

Deep Learning Solutions for Classifying Patients on Opioid Use

Zhengping Che\(^1\), Jennifer St. Sauver, PhD\(^2\), Hongfang Liu, PhD\(^2\), Yan Liu, PhD\(^1\)

\(^1\)Department of Computer Science, University of Southern California, Los Angeles, CA
\(^2\)Department of Health Sciences Research, Mayo Clinic, Rochester, MN

Abstract

Opioid analgesics, as commonly prescribed medications used for relieving pain in patients, are especially prevalent in US these years. However, an increasing amount of opioid misuse and abuse have caused lots of consequences. Researchers and clinicians have attempted to discover the factors leading to opioid long-term use, dependence, and abuse, but only limited incidents are understood from previous works. Motivated by recent successes of deep learning and the abundant amount of electronic health records, we apply state-of-the-art deep and recurrent neural network models on a dataset of more than one hundred thousand opioid users. Our models are shown to achieve robust and superior results on classifying opioid users, and are able to extract key factors for different opioid user groups. This work is also a good demonstration on adopting novel deep learning methods for real-world health care problems.

1 Introduction

Opioid analgesics are effective, commonly prescribed medications used for management of both acute and chronic pain in patients with many different medical conditions and following many medical procedures\(^1,2\). Prescription of opioids in the United States is high, and between 2011 and 2012, nearly 7% of the adult population was estimated to have taken an opioid in the last thirty days\(^3,4\). However, as reported in several previous studies, these medications do not effectively control pain in all patients\(^5,6,7\), and many patients are at high risk of adverse effects due to these medications\(^7,8\). A meta-analysis of randomized trials found that 80% of patients treated with opioids for chronic, non-cancer pain experienced at least one adverse event, with symptoms ranging from mild nausea to life-threatening respiratory depression\(^8\). In addition, the US is experiencing an opioid epidemic. Specifically, opioids are increasingly misused and diverted from their intended recipients, and abuse and overdoses have risen alarmingly in the last ten years\(^9\). The rate for drug overdose deaths, driven largely by opioid overdose, increased approximately 140% from 2000 to 2014\(^10\). In 2017, one of the largest pharmaceutical distributors in US was fined a record $150 million for failing to report suspicious orders linked to the opioid addiction epidemic\(^11\). Indeed, prompt and proper actions need to be taken to achieve a balanced opioid usage strategies, stem the tide of this public health epidemic, and prevent further devastating consequences.

The factors that contribute to opioid use – particularly the patient factors that contribute to long-term, chronic use of these medications and/or dependence or abuse of these drugs are poorly understood. Previous work found significant increases in incident opioid prescriptions for chronic, non-malignant pain between 1997 and 2005 in the Kaiser Permanente and Group Health populations\(^12\). Additionally, the proportion of the population receiving long-term therapy nearly doubled in the same time frame. The most common indications for long-term use in this study were chronic back pain, extremity pain, and osteoarthritis. Apart from these data, however, little is known about who receives opioid analgesic prescriptions in an average community. Additionally, with the exception of a few studies exploring the role of mental illness, depression, or previous patterns of substance abuse\(^13,14,15\), patient characteristics that might contribute to these adverse outcomes have not been described.

The rapid growth in electronic health record (EHR) adoption provides a wealth of patient information that could help identify patients at high risk of long-term opioid use or dependence. If one predictive or classifying model can leverage such data for analysing opioid usage and/or dependence, that is, the model has the ability to identify patients likely to benefit from or get addicted to these medications and target therapy more appropriately to them, we can expect those models to be able to extract the knowledge of the clinical characteristics associated with the progression of a short-term to an episodic or long-term opioid prescribing pattern and aid in the identification of at-risk patients and provide the basis for developing targeted clinical interventions. In the era of data explosion, however, more powerful data-driven learning models are in urgent demand in order to fully utilize the large amount of EHR data, identify meaningful features for opioid dependence or abuse, provide precise information for clinicians to make early decisions, and ultimately contribute to better personalized health care quality.
In this paper, we utilized state-of-the-art deep learning models on a much larger data set for opioid usage prediction and factor investigation tasks. Deep learning models have brought lots of significant successes including but not limited to recognizing and distinguishing thousands of human faces at a time\textsuperscript{16,17}, understanding, translating, and generating human languages\textsuperscript{18,19}, and mastering games and beating top human professional players\textsuperscript{20}. Deep learning is also revolutionizing the health care domain with the focuses on a variety of important and challenging tasks, such as computational phenotyping\textsuperscript{21,22}, predictive modeling\textsuperscript{23,24} and medical imaging analyzing\textsuperscript{25,26}. It is well known that deep learning solutions equipped with ample computational resources and large-scale datasets are able to go far beyond traditional statistical methods and shed light on intriguing real-world applications in health care. In this paper, we demonstrated our proposed deep learning solutions for identifying opioid user groups and showed that they provided superior classification results and outperformed other widely used learning baseline methods. We validated important factors and risk factors identified by deep learning models with previous clinical studies. Our work also provided a practical example on properly adopting novel deep learning methods for real-world health care problems leveraging large-scale EHR data.

2 Data and Task Descriptions

In this work, we took a cohort of 142,377 patients from the Rochester Epidemiology Project (REP)\textsuperscript{27}. The total number of people identified by REP, as shown in previous work\textsuperscript{28}, cover about 98.7% of the population that reside in Olmsted County by the US Census. Thus, this large-scale dataset is well-representative for this population-based study and suitable for powerful and complex deep learning models.

Cohort Selection

First, all outpatient drug prescriptions were obtained from Mayo Clinic and the Olmsted Medical Center from January 1, 2003 through March 31, 2016 for patients who authorized the use of their medical records for research purposes. The drug prescriptions were standardized using the 2016 version of RxNorm\textsuperscript{29}. We kept the records for all patients who received at least one opioid analgesic prescription between July 1, 2013 and March 31, 2016 and did not have any opioid prescriptions 6 months prior to their first prescriptions within the study period. The analgesic prescriptions were determined by the RxNorm Code, with either National Drug File Reference Terminology (NDF-RT) code C8834 (Opioid Analgesics) or ingredient code 10689 (Tramadol) and 352362 (Acetaminophen/Tramadol). In order to remove incorrectly duplicated and modified prescription records, only the last prescription would be kept if same drug prescriptions were made for one patient within 30 minutes. A cohort of $N = 102,166$ patients was created after these data cleaning and selection steps.

Group Identification

All patients were classified into three groups, namely \textit{short-term} users (ST), \textit{long-term} users (LT), and \textit{opioid-dependent} users (OD). ST and LT groups were defined by the CONSORT study\textsuperscript{30} and the same as in our previous work\textsuperscript{31}. Episodes of opioid prescription lasting longer than 90 days and with 120 or more total days supply or 10 or more prescriptions were classified as long-term ($N_{LT} = 21,570$), while others were classified as short-term ($N_{ST} = 80,596$). $N_{OD} = 749$ opioid-dependent patients were further identified by the diagnosis of “opioid dependence” from their problem lists. It is noting that the relatively low identification rate might be due to the fact that only part of dependent patients got explicit diagnosis in the problem lists by doctors. All identified dependent users were validated by clinicians. Table 1 shows detailed data characteristics of each patient group, which also match the finding in our previous related study on a smaller dataset\textsuperscript{31}. Two classification tasks were considered in our experiments: 1) whether the patient will become a long-term opioid user or just a short-term opioid user (Task \textit{ST-LT}), and 2) whether a long-term opioid user is an opioid-dependent patient or not (Task \textit{LT-OD}).

3 Methodology

In this section, we first describe our feature extraction and temporal data processing steps. Next, two deep learning models deployed in our study are presented: A deep feed-forward neural network model with multiple hidden layers, and a recurrent neural network model with Long Short-Term Memory which can better model time series data. Several ways used to improve the model performance and obtain important features are discussed, followed by the descriptions of some machine learning baselines. In the following part, we use bold capital letter (e.g., $W$) to refer to matrix variable, bold lowercase letter (e.g., $b$) for vector variable, and unbold letter (e.g., $l$, $D$) for scalar, if not specified.
Table 1: Data characteristics of different patient groups.

<table>
<thead>
<tr>
<th></th>
<th>Short-Term/ST</th>
<th>Long-Term/LT*</th>
<th>Opioid-Dependent/OD</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Count</td>
<td>Percentage</td>
<td>Count</td>
<td>Percentage</td>
</tr>
<tr>
<td>Total Number of Patients</td>
<td>80596</td>
<td>78.89%</td>
<td>21570</td>
<td>21.11%</td>
</tr>
<tr>
<td>Sex</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Men</td>
<td>37981</td>
<td>47.13%</td>
<td>8417</td>
<td>39.16%</td>
</tr>
<tr>
<td>Women</td>
<td>42453</td>
<td>52.67%</td>
<td>13075</td>
<td>60.62%</td>
</tr>
<tr>
<td>Other/Unknown</td>
<td>162</td>
<td>0.20%</td>
<td>48</td>
<td>0.22%</td>
</tr>
<tr>
<td>Age</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>≤ 18</td>
<td>5900</td>
<td>7.32%</td>
<td>447</td>
<td>2.07%</td>
</tr>
<tr>
<td>19 – 29</td>
<td>13701</td>
<td>17.00%</td>
<td>1311</td>
<td>6.08%</td>
</tr>
<tr>
<td>30 – 49</td>
<td>27896</td>
<td>34.36%</td>
<td>5416</td>
<td>25.11%</td>
</tr>
<tr>
<td>50 – 64</td>
<td>18027</td>
<td>22.37%</td>
<td>5570</td>
<td>25.82%</td>
</tr>
<tr>
<td>≥ 65</td>
<td>15272</td>
<td>18.95%</td>
<td>8826</td>
<td>40.92%</td>
</tr>
<tr>
<td>Race</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>66184</td>
<td>82.12%</td>
<td>19297</td>
<td>89.46%</td>
</tr>
<tr>
<td>Hispanic</td>
<td>4151</td>
<td>5.15%</td>
<td>697</td>
<td>3.23%</td>
</tr>
<tr>
<td>African American</td>
<td>4131</td>
<td>5.13%</td>
<td>898</td>
<td>4.16%</td>
</tr>
<tr>
<td>Asian</td>
<td>3225</td>
<td>4.00%</td>
<td>361</td>
<td>1.67%</td>
</tr>
<tr>
<td>Other/Unknown</td>
<td>2905</td>
<td>3.60%</td>
<td>317</td>
<td>1.47%</td>
</tr>
<tr>
<td>Mortality</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dead</td>
<td>4481</td>
<td>5.56%</td>
<td>17075</td>
<td>79.16%</td>
</tr>
<tr>
<td>Alive/Unknown</td>
<td>76115</td>
<td>94.44%</td>
<td>4495</td>
<td>20.84%</td>
</tr>
<tr>
<td>Tobacco Use</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Never/Unknown</td>
<td>46264</td>
<td>57.40%</td>
<td>7159</td>
<td>33.19%</td>
</tr>
<tr>
<td>Secondhand Only</td>
<td>746</td>
<td>0.93%</td>
<td>750</td>
<td>3.48%</td>
</tr>
<tr>
<td>Past/Current</td>
<td>33586</td>
<td>41.67%</td>
<td>13661</td>
<td>63.33%</td>
</tr>
<tr>
<td>First Time of Anxiety or Depression</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Never</td>
<td>58322</td>
<td>72.36%</td>
<td>11002</td>
<td>51.01%</td>
</tr>
<tr>
<td>Before FOT†</td>
<td>10431</td>
<td>12.94%</td>
<td>3230</td>
<td>14.97%</td>
</tr>
<tr>
<td>After FOT</td>
<td>11843</td>
<td>14.69%</td>
<td>7338</td>
<td>34.02%</td>
</tr>
<tr>
<td>First Time of Substance Abuse</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Never</td>
<td>70039</td>
<td>86.90%</td>
<td>15283</td>
<td>70.85%</td>
</tr>
<tr>
<td>Before FOT</td>
<td>4315</td>
<td>5.35%</td>
<td>1730</td>
<td>8.02%</td>
</tr>
<tr>
<td>After FOT</td>
<td>6242</td>
<td>7.74%</td>
<td>4557</td>
<td>21.13%</td>
</tr>
<tr>
<td>First Time of Other Psychological Diagnosis</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Never</td>
<td>35716</td>
<td>43.31%</td>
<td>3482</td>
<td>16.14%</td>
</tr>
<tr>
<td>Before FOT</td>
<td>27627</td>
<td>34.28%</td>
<td>9253</td>
<td>42.90%</td>
</tr>
<tr>
<td>After FOT</td>
<td>17253</td>
<td>21.41%</td>
<td>8835</td>
<td>40.96%</td>
</tr>
</tbody>
</table>

Feature Extraction We retrieved structured EHR data of the chosen patients from REP historic sources, Olmsted Medical Clinic and Hospital, Mayo and Mayo Clinic Health System between January 1, 2003 and March 31, 2016. We extracted code records with time stamps and other information from three chart tables: diagnoses (DX), procedures (PR), and prescriptions (RX). The details are shown in Table 2. Instead of taking raw records in these tables, we mapped all the codes to a higher level code space, for the following two reasons: First, coding systems used in Mayo were different and often change from time to time. For example, three different coding systems, ICD-9, ICD-10, and HICDA (Hospital International Classification of Diseases Adapted) were used for disease records in DX table. HICDA codes were used only before 2011, ICD-10 codes have not been in use until the year of 2015. This prevented us from taking one single raw code system and thus a consistent mapping of these conceptually-overlaid codes was required. Second, since there were tens of thousands of different raw codes in each table, the raw data tables were quite sparse and difficult to be examined in the feature level. Therefore, we mapped all DX and PR codes into categories in Clinical Classifications Software (CCS) and all RX codes into NDF-RT class. In PR table, we also recorded the corresponding quantity together alone with the code.

† Notice all patients included in OD group were also included in LT group.
† FOT refers to the time of the first opioid prescription for each patient.
Table 2: Record table descriptions and statistics.

<table>
<thead>
<tr>
<th>Table Name</th>
<th>Descriptions</th>
<th># of Records</th>
<th>Raw Code Coding Systems</th>
<th>Count</th>
<th>Mapped Code Coding System</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>DX</td>
<td>Diagnosis records</td>
<td>56,229,157</td>
<td>ICD-9, ICD-10, HICDA</td>
<td>43,438</td>
<td>CCS</td>
<td>284</td>
</tr>
<tr>
<td>PR</td>
<td>Procedure, service, and surgical index</td>
<td>46,386,740</td>
<td>ICD-9, ICD-10, CPT/HCPCS</td>
<td>18,984</td>
<td>CCS</td>
<td>245</td>
</tr>
<tr>
<td>RX</td>
<td>Prescription records</td>
<td>8,102,477</td>
<td>Ingredient RxNorm Code</td>
<td>2,460</td>
<td>NDF-RT Class RxNorm Code</td>
<td>307</td>
</tr>
</tbody>
</table>

Temporal Data Processing  We applied 1-of-K (one-hot) encoding\(^35\) on the extracted features and used either the temporal sum-pooling or segmentations of the encoded features to get numerical features from sparse categorical features simply yet effectively. The 1-of-K encoding converts each record line to a single binary vector of the same length, and temporal sum-pooling and segmentation step (i.e., sum-pooling in each temporal segment) further aggregates features along the temporal direction. For RX table, we used the length of days instead of 1 when applying 1-of-K encoding on prescription records to utilize the important quantitative effective length information for each prescription. For example, a prescription record with length of 4 days was converted into a vector like \([0,...,4,...,0]\). Our recurrent neural network models were able to handle time series data directly and capture temporal information. We computed the sum-pooling vector\(^36\) in each year and stacked them into a matrix, which we referred as yearly temporal segmentations. Since the medical records for patients might be of different length, the resulted matrices also had different length \(T_{\text{seg}}\), thus these matrices could not be directly used in other models including non-recurrent deep networks and other machine learning baselines. For those model we chose the sum-pooling along all time steps and thus obtained a vector with fixed length. The length was the sum of the numbers of all mapped features from three tables \((D = D_{DX} + D_{PR} + D_{RX} = 836)\) in our dataset. The prediction models took the obtained data as the input and provided predictive results. The entire pipeline is illustrated in Figure 1.

Figure 1: An illustration of the proposed pipeline from raw cohort data to final prediction. Left: DNN prediction model for data with temporal sum-pooling; Right: RNN prediction model for data with temporal segmentation.
Deep Feed-Forward Neural Network (DNN) Model A deep feed-forward neural network model (DNN) is composed with multiple non-linear transformation layers. The output of each layer is fed to the next layer as input. For a DNN model with $L$ layers (i.e., $L - 1$ hidden layers and one final output layer), the input vector for the $l$th layer is denoted as $x^{[l]} \in \mathbb{R}^{D^{[l]}}$, and rectified linear unit (ReLU) function is used as the non-linear transformation function for each hidden layer. The output of layer $l$ is

$$h^{[l]} = ReLU(W^{[l]}x^{[l]} + b^{[l]}),$$

which is also the input to the next layer $x^{[l+1]}$. Here $W^{[l]}$ and $b^{[l]}$ are parameters which can be learned via back-propagation during training. Notice that in our case we had input dimension $D^{[1]} = D$. We chose ReLU function because 1) it is shown to not suffer from gradient vanishing problem during training compared with other non-linear transformations such as sigmoid and tanh functions, and 2) our model investigation step requires it. To conduct binary classification tasks, we applied sigmoid function $\sigma(x) = \frac{1}{1+\exp(-x)}$ in the last layer and set the output dimension to be 1. In other words, we had $W^{[L]} \in \mathbb{R}^{L^{[L]} \times 1}$ and $h^{[L]} = \sigma(W^{[L]}x^{[L]} + b^{[L]}) \in [0, 1]$. With this proposed DNN model structure, we could learn the weights by optimizing binary cross-entropy loss function during training, which is

$$\ell_{loss} = -\sum_{n=1}^{N}(y_n \log h^{[L]} + (1 - y_n) \log(1 - h^{[L]})),$$

where $y_n$ is the binary label for $n$th patient. In ST-LT prediction task, $y_n = 1/0$ indicates long-term/short-term opioid user. In LT-OD prediction task, $y_n = 1$ and $y_n = 0$ are for opioid-dependent and other patients, respectively.

Recurrent Neural Network (RNN) Model In order to handle sequential or temporal data of arbitrary length and capture temporal information from the data, recurrent neural network (RNN) models are widely used. Unlike feed-forward neural network models, RNN models perform the same operation at each step of the time sequence and feed the output to the next time step as part of the input. Thus, RNN models are able to memorize what they have seen before and benefit from shared model weights (parameters) for all time steps. In order to capture complex long temporal dependency and avoid vanishing gradient problems, some modified RNN models such as Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) have been proposed with state-of-the-art performance. Assume the input is a matrix $X \in \mathbb{R}^{T \times D}$, where $T$ is the number of temporal segments and varies for different patients, and $D$ is the feature dimension. The $t$th row $x_t \in \mathbb{R}^{D^{[l]}}$ of the matrix represents the encoding vector at time step $t$. We used LSTM in our RNN prediction model. At each time step $t$, LSTM takes the input at that time step $x_t$ and output at previous time step $h_{t-1}$ to update its inner cell state $c_t$ and produce the current output $h_t$, as follows:

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f) \quad i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i) \quad o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tanh(W_c x_t + U_c h_{t-1} + b_c) \quad h_t = o_t \odot \tanh(c_t)$$

Here $W_f, U_f, b_f$ are all model parameters, $\sigma()$ is the sigmoid function, $\tanh()$ is the tangent function, and $\odot$ refers to element-wise multiplication between two vectors. We set the initial values of $h_0$ and $c_0$ to be 0. For binary classification tasks we applied another sigmoid layer of dimension 1 on top of the last output $h_T$, i.e., $h^{top} = \sigma(W h_T + b) \in [0, 1]$, and trained the RNN model by optimizing the binary cross-entropy loss on the training dataset.

Implementation and Training Techniques We implemented our deep learning models with Python Theano and Keras libraries and all the models were reproducible. For both DNN and RNN models, we set the dimension of each hidden layer to be 256, which was chosen to have proper size and good performance. Several training techniques were designed or used to better handle our data. First, we applied an L-1 regularizer with coefficient $1 \cdot e^{-4}$ to make the model robust and able to select important features. Our preliminary experiments showed that L-1 provided more compact models with better or similar performance as L-2 or no regularizers. Second, dropout technique with rate $p_{dr} = 0.5$ was used for all layers to reduce overfitting and avoid harmful weight co-adaptations. This was implemented by randomly dropping out units by probability of $p_{dr}$ in the neural networks at training time and re-scaling all the weights by $W_{test} = p_{dr} W_{train}$ at test time. Third, we applied novel batch normalization on all non-recurrent layers. The basic idea is to normalize the activations of the previous layer such that the outputs keep mean of 0 and standard deviation of 1 in each mini-batch during training. The running averages computed on training dataset are used to normalize the outputs at test time. This strategy speeded up training process and improved
Investigating Important Features Deep learning models are often argued to be difficult to interpret and investigate, especially because of their complex structures and thousands of or even millions of parameters. Furthermore, carelessly attempting to check and visualize individual units in neural networks might lead to misleading conclusions. However, by checking the overall model weights and structures, it is still possible to identify important features extracted from the deep learning models and obtain rough quantitative evaluations. We designed the feature importance score \( I \) for such purpose. We first take the weight matrix in the first layer \( W^{[1]} \in \mathbb{R}^{D \times D} \) of a DNN model as an example, where each column \( u^{[1]}_d \) of \( W^{[1]} \) corresponds to the \( d \)th input feature. A simple way to quantify the feature importance is to take the summation of each column. The first importance score of \( d \)th input feature is formally defined as \( I_1(d) = \sum_{i=1}^{D} W^{[1]}[i, d] \), where \( W[i, d] \) is the number in \( i \)th row and \( d \)th column of \( W \). However, we only consider the first layer in this score, which is definitely insufficient for a deep models. To overcome this issue, we need to take weights in higher layers into consideration. Since ReLU function was used as transformation function in all the hidden layers, we multiplied weights in all layers and took the value at the corresponding index as the importance score. We also need to take care of the impact of batch normalization since it introduces different scales on parameters, so we apply batch normalization operation before we multiply the weight matrix for each layer. Thus, the second importance score can be defined formally as

\[
I_2 = W^{[1]} BN^{[1]} \cdot W^{[2]} BN^{[2]} \cdot \ldots \cdot W^{[L]} BN^{[L]} (1) \in \mathbb{R}^{1 \times D}
\]

where \( BN^{[l]} \) denotes the batch normalization operation for layer \( l \). This process can also be viewed as a simplified version of the original deep neural networks without non-linear transformations or bias vectors. In our experiments, \( I_2 \) was used for our DNN models. In order to validate the way of investigating important features and verify the selected features, we checked previous clinical studies and compared with features from our baseline models.

Other Machine Learning Baselines In order to evaluate the proposed deep models and validate the findings, we also compared some commonly used machine learning baselines in clinical research, including Logistic regression (LR), linear support vector machine with hinge loss (SVM), and random forest (RF). All the baselines are implemented in Python Scikit-learn package. We kept most of the default settings and hyperparameters which are shown to be effective in practice, but made several specific changes to better fit our tasks. In order to distinguish important features we introduce sparsity into the model coefficients, we also used L-1 penalty in LR and SVM, tuned the regularization strength \( C \) by searching from \( 1 \cdot e^{-4} \) to 10 and finally chose \( C = 0.1 \) in our experiments since it usually provided best prediction results. In RF, using more trees usually leads to better results, but also possibly makes the model computationally inefficient and overfitted to training samples, and the model size also will grow linearly to the number of trees. Since using more trees brought negligible performance improvement but drastically increased the model size in our preliminary experiments, we took the default setting (10 trees) so that the RF model had moderate size as others. As shown in Table 3, all the tested models had comparable sizes and thus the performance comparison was fair.

4 Results and Discussions

Classification Result Comparison As mentioned before, we conducted two classification tasks (ST-LT and LT-OD). All 102,166 patients were included in the 5-fold cross validation for ST-LT task. Only 3.47% long-term users

Table 3: Model size comparison when saved into binary files in disk. All deep learning models are serialized and saved in HDF5 files, and other models are saved in cPickle files.
Table 4: Long-term opioid patient prediction (ST-LT) results (mean ± 95% conﬁdence interval). In Setting A, we take
all the medical records before the date when the patient is marked as long-term user or Mar, 31, 2016, whichever is
earlier; Setting B is the same as Setting A except that we exclude all the opioid and non-opioid analgesics prescriptions;
In Setting C we take records made before the patient’s ﬁrst opioid prescription. Best results shown in bold.

Setting A

Setting B

Setting C

LR

Baseline Models
SVM

RF

DNN-1hl‡

Deep Models
DNN-2hl
DNN-3hl

Acc.

0.8946 ± 0.002

0.8938 ± 0.002

0.8666 ± 0.004

0.8960 ± 0.002

0.8954 ± 0.001

0.8975 ± 0.002

0.8961 ± 0.002

AUC

0.9074 ± 0.002

0.9038 ± 0.002

0.8747 ± 0.003

0.9086 ± 0.002

0.9082 ± 0.002

0.9091 ± 0.002

0.9094 ± 0.002

Prec.

0.8483 ± 0.007

0.8671 ± 0.006

0.8213 ± 0.009

0.8539 ± 0.013

0.8546 ± 0.009

0.8567 ± 0.009

0.8719 ± 0.008

Rec.

0.6099 ± 0.006

0.5868 ± 0.007

0.4702 ± 0.018

0.6122 ± 0.009

0.6082 ± 0.006

0.6178 ± 0.006

0.5957 ± 0.007

κ

0.6473 ± 0.007

0.6383 ± 0.007

0.5249 ± 0.018

0.6516 ± 0.007

0.6489 ± 0.004

0.6571 ± 0.006

0.6472 ± 0.006

Acc.

0.8385 ± 0.002

0.8372 ± 0.002

0.8162 ± 0.002

0.8371 ± 0.002

0.8340 ± 0.002

0.8352 ± 0.002

0.8371 ± 0.002

AUC

0.8369 ± 0.002

0.8366 ± 0.002

0.8044 ± 0.002

0.8412 ± 0.002

0.8362 ± 0.002

0.8362 ± 0.003

0.8466 ± 0.002

Prec.

0.7161 ± 0.010

0.7309 ± 0.011

0.6590 ± 0.011

0.7319 ± 0.013

0.6999 ± 0.010

0.7121 ± 0.022

0.6889 ± 0.012

Rec.

0.3892 ± 0.005

0.3623 ± 0.006

0.2683 ± 0.005

0.3612 ± 0.008

0.3749 ± 0.016

0.3712 ± 0.018

0.4207 ± 0.020

RNN

κ

0.4177 ± 0.007

0.4005 ± 0.009

0.2952 ± 0.006

0.3998 ± 0.008

0.3996 ± 0.013

0.4005 ± 0.009

0.4297 ± 0.010

Acc.

0.7917 ± 0.001

0.7908 ± 0.001

0.7890 ± 0.001

0.7919 ± 0.001

0.7920 ± 0.001

0.7915 ± 0.001

0.7989 ± 0.001

AUC

0.7323 ± 0.003

0.7327 ± 0.003

0.6936 ± 0.003

0.7220 ± 0.004

0.7340 ± 0.004

0.7218 ± 0.004

0.7536 ± 0.003

Prec.

0.5366 ± 0.019

0.5303 ± 0.021

0.5007 ± 0.010

0.5670 ± 0.031

0.5943 ± 0.012

0.5774 ± 0.027

0.5692 ± 0.028

Rec.

0.0996 ± 0.003

0.0800 ± 0.003

0.1279 ± 0.004

0.0646 ± 0.005

0.0672 ± 0.011

0.0490 ± 0.015

0.1991 ± 0.002

κ

0.1090 ± 0.005

0.0885 ± 0.005

0.1289 ± 0.006

0.0756 ± 0.004

0.0658 ± 0.013

0.0587 ± 0.016

0.2076 ± 0.007

Table 5: Opioid-dependent patient prediction (LT-OD) results (mean ± 95% conﬁdence interval). Settings A, B, C are
deﬁned the same as those in Table 4. Best results shown in bold.

Setting A

Setting B

Setting C

LR

Baseline Models
SVM

RF

DNN-1hl‡

Deep Models
DNN-3hl
DNN-2hl

Acc.

0.6929 ± 0.010

0.6805 ± 0.007

0.7417 ± 0.007

0.7441 ± 0.010

0.7550 ± 0.008

0.7547 ± 0.009

0.7607 ± 0.009

AUC

0.7119 ± 0.010

0.6985 ± 0.010

0.7773 ± 0.011

0.7853 ± 0.012

0.7975 ± 0.010

0.8044 ± 0.011

0.8060 ± 0.010

Prec.

0.5385 ± 0.017

0.5212 ± 0.010

0.7049 ± 0.022

0.6214 ± 0.021

0.6323 ± 0.012

0.6328 ± 0.018

0.6896 ± 0.020

Rec.

0.5924 ± 0.022

0.5748 ± 0.019

0.3986 ± 0.016

0.6233 ± 0.028

0.6457 ± 0.031

0.6471 ± 0.024

0.5205 ± 0.021

κ

0.3262 ± 0.017

0.2966 ± 0.016

0.3555 ± 0.015

0.4273 ± 0.019

0.4520 ± 0.021

0.4571 ± 0.017

0.4505 ± 0.019

Acc.

0.6763 ± 0.007

0.6669 ± 0.009

0.7331 ± 0.010

0.7376 ± 0.011

0.7406 ± 0.009

0.7427 ± 0.006

0.7417 ± 0.006

AUC

0.6968 ± 0.008

0.6898 ± 0.010

0.7659 ± 0.013

0.7720 ± 0.009

0.7821 ± 0.012

0.7829 ± 0.008

0.8010 ± 0.007

Prec.

0.5156 ± 0.013

0.5029 ± 0.012

0.6784 ± 0.022

0.6214 ± 0.023

0.6146 ± 0.017

0.6289 ± 0.017

0.7107 ± 0.019

Rec.

0.5743 ± 0.022

0.5600 ± 0.018

0.3867 ± 0.024

0.5733 ± 0.026

0.6162 ± 0.025

0.5810 ± 0.039

0.3976 ± 0.026

κ

0.2951 ± 0.020

0.2734 ± 0.020

0.3301 ± 0.028

0.4046 ± 0.020

0.4201 ± 0.019

0.4098 ± 0.018

0.3787 ± 0.021

Acc.

0.6404 ± 0.007

0.6332 ± 0.012

0.6994 ± 0.007

0.6870 ± 0.009

0.6911 ± 0.009

0.7065 ± 0.008

0.6956 ± 0.008

AUC

0.6512 ± 0.009

0.6429 ± 0.010

0.6999 ± 0.011

0.7130 ± 0.014

0.7216 ± 0.014

0.7279 ± 0.014

0.7144 ± 0.011

Prec.

0.4639 ± 0.030

0.4554 ± 0.017

0.6019 ± 0.021

0.5491 ± 0.020

0.5485 ± 0.023

0.6193 ± 0.024

0.5975 ± 0.018

Rec.

0.4605 ± 0.020

0.4629 ± 0.017

0.3067 ± 0.018

0.4590 ± 0.075

0.5338 ± 0.065

0.3305 ± 0.030

0.2895 ± 0.028

κ

0.1906 ± 0.023

0.1821 ± 0.022

0.2342 ± 0.020

0.2702 ± 0.029

0.3006 ± 0.026

0.2542 ± 0.024

0.2542 ± 0.032

RNN

are opioid dependent and thus the labels are quite imbalanced for LT-OD task. To get robust prediction and features,
we randomly generated 14 datasets with class ratio of 13 by downsampling the non-opioid-dependent patients. Each
generated dataset had records from 2 237 patients. We further introduced three different settings (A, B, C) to test
model performances in different simulated situations. The deﬁnitions of the settings are described in the caption
of Table 4. Setting A was the ideal case and the best prediction results could be achieved in this setting since all
possible information was taken into consideration. After we found the analgesics usage can be good indicators for our
prediction tasks and might hide other indicators, we designed Setting B which might impair the prediction performance
but help us ﬁnd some hidden but useful features. Setting C was the most practical case among the three and we took
it to demonstrate the early prediction capacity of our methods. For all the settings and tasks, classiﬁcation accuracy
‡ DNN-khl

refers to DNN models with k hidden layers and one output layer, k ∈ {1, 2, 3}.

531


(Acc.), area under the receiver operating characteristic curve score (AUC), precision (Prec.), recall (Rec.), and Cohen’s kappa coefficient (κ) are reported. Results for ST-LT and LT-OD tasks are shown in Table 4 and Table 5, respectively. First, deep models provided the best performance in terms of most evaluation metrics. The improvements on LT-OD were larger than ST-LT. Second, the RNN models which captured temporal information usually but not always beat standard DNN models. It obtained the best AUC score in 5 out of 6 settings. This implies that even loosely segmented time series contains useful temporal information. However, the superiority of RNN was shown to be less on LT-OD than ST-LT, and one possible reason is the lack of training samples on LT-OD.

Feature Analysis It is useful to get to know which features are more related to opioid use, or played more important roles in the prediction models. We take the DNN-3hl models in Setting A and show the top ten most important features ordered by the absolute value of importance score $Z_2$ in Table 6. Basically, features with positive/negative score can be interpreted as positively/negatively correlated to the prediction target (long-term use in ST-LT, and opioid dependence in LT-OD). The score should only be compared within the same model and the same task. For both tasks, “Opioid Analgesics” prescription is selected as the most important indicators. “Non-opioid Analgesics” is also an important factor for long-term opioid use but not very useful to distinguish opioid-dependent user from long-term user. Several disorders diagnoses, such as “substance-related disorders”, “anxiety disorders”, and “other mental health disorders” (e.g., interview, evaluation, and consultation), are all highly related to opioid dependence. These findings are consistent with previous studies and most of the top features are also selected by LR and RF baselines. In addition, the scores for top features in LT-OD task are closer than those in ST-LT. This indicates that in Setting A identifying opioid-dependent users is a more challenging task which requires the exploit of more different features. The fact that all models had higher evaluation score on ST-LT than LT-OD in Setting A (Table 4 and 5) also supported the same claim. As we only did preliminary investigations, more details and validations will be discovered in the following work.

Table 6: Most Important features for long-term opioid patient (ST-LT, left) and opioid-dependent patient (LT-OD, right) identified from DNN-3hl model.

<table>
<thead>
<tr>
<th>ST-LT Prediction</th>
<th>LT-OD Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table</td>
<td>Code</td>
</tr>
<tr>
<td>RX</td>
<td>C8834</td>
</tr>
<tr>
<td>RX</td>
<td>C8890</td>
</tr>
<tr>
<td>RX</td>
<td>C8838</td>
</tr>
<tr>
<td>PR</td>
<td>CCS 227</td>
</tr>
<tr>
<td>DX</td>
<td>CCS 258</td>
</tr>
<tr>
<td>RX</td>
<td>C4859</td>
</tr>
<tr>
<td>DX</td>
<td>CCS 203</td>
</tr>
<tr>
<td>DX</td>
<td>CCS 205</td>
</tr>
<tr>
<td>DX</td>
<td>CCS 98</td>
</tr>
<tr>
<td>DX</td>
<td>C8834</td>
</tr>
</tbody>
</table>

5 Summary

In this paper, we applied deep learning models for opioid user group predictions on a large-scale real-world EHR dataset. The deep learning models were able to achieve superior classification performance and identify useful feature indicators for opioid-dependent and long-term users. Our work demonstrated how novel deep learning models can be utilized to obtain state-of-the-art performance in practical clinical studies. In our future work, we plan to further investigate important features extracted from deep models, and incorporate numerical and unstructured EHR data along with code records into deep learning prediction models. We also plan to explore more fancy deep learning models to capture the temporal dependencies and evolutions for medical records of opioid users.

References


43. Chollet F. Keras. GitHub; 2015.


Automated Metabolic Phenotyping of Cytochrome Polymorphisms Using PubMed Abstract Mining

Luoxin Chen, MA1, Carol Friedman, PhD1, Joseph Finkelstein, MD, PhD1
1Department of Biomedical Informatics, Columbia University, New York, NY, US

Abstract

Pharmacogenetics-related publications, which are increasing rapidly, provide important new pharmacogenetics knowledge. Automated approaches to extract information of new alleles and to identify their impact on metabolic phenotypes from publications are urgently needed to facilitate personalized medicine and improve clinical outcomes. Cytochrome polymorphisms, responsible for a wide variation of drug pharmacodynamics, individual efficacy and adverse effects, have significant potential for optimizing drug therapy. A few studies have addressed specialized efforts to automatically extract cytochrome polymorphisms and their characterizations regarding metabolic phenotypes from the literature. In this paper, we present a novel rule-based text-mining system to extract metabolic phenotypes of polymorphisms from PubMed abstracts with a focus on cytochrome P450. This system is promising as it achieved a precision of 85.71% in a preliminary proof-of-concept evaluation and is expected to automatically provide up-to-date metabolic information for cytochrome polymorphisms, which is critical to advance personalized medicine and improve clinical care.

Introduction

An overwhelming majority of medications are characterized by xenobiotics that are foreign chemical substances not normally produced by or expected to be present within human organisms. Hepatic enzymes are responsible for the metabolism of xenobiotics by first activating them, followed by conjugation of the active secondary metabolite and subsequent excretion in bile or urine. The most important group of enzymes involved in xenobiotic metabolism is hepatic microsomal cytochrome P450 (CYP or P450) since the majority of drugs are the substrates for these enzymes. CYP forms a family of about 20 cytochrome enzymes that have been confirmed to participate in xenobiotic metabolism in humans. Among these enzymes, CYP2D6, CYP2C9, CYP2C19, CYP2B6, CYP3A4 and CYP1A2 play the most critical role since they account for more than 90% of drugs metabolized by CYP.

Recent studies demonstrated that CYP enzymes are highly polymorphic and that cytochrome polymorphisms are responsible for a wide variation of drug pharmacodynamics, individual efficacy and the occurrence of adverse effects. Specifically, single nucleotide polymorphisms (SNP) of genes encoding CYP enzymes result in a spectrum of metabolic phenotypes responsible for individual drug responses. CYP phenotypes are divided into the extensive (or normal) metabolizers (EM) that demonstrate regular metabolic activity and the poor metabolizers (PM) that are characterized by low metabolic activity. Additional phenotype categorization includes intermediate metabolizers (IM) that have metabolic activity ranging between PM and EM, and rapid metabolizers (RM) that demonstrate higher metabolic activity than the EM.

Broad availability of relatively inexpensive testing of cytochrome polymorphisms can facilitate use of pharmacogenetics in clinical practice. Multiple studies demonstrated a significant potential of pharmacogenetics in optimizing drug therapy, reducing adverse effects and improving clinical outcomes. For example, it was shown that people with CYP2D6 genes having a rapid metabolizer phenotype should avoid codeine use because of increased toxicity, which in some cases may have lethal consequences. Lack of effectiveness of codeine was demonstrated in poor metabolizers of CYP2D6 for whom codeine had to be replaced by other medications not metabolized by CYP2D6. As polymorphisms of CYP2C19 affects the antiplatelet agent clopidogrel, patients undergoing percutaneous coronary intervention who are known to be poor metabolizers of CYP2C19 are required to consider alternate antiplatelet therapy to avoid the risk of bleeding complications. Cytochrome polymorphism in older adults with polypharmacy was shown to be an independent risk factor of frequent hospital readmissions.

During recent decades, we witnessed an explosion of publications in the area of pharmacogenetics. Information on newly discovered alleles and their potential impact on drug metabolism is constantly expanding. The majority of published information is available in PubMed, however optimal extraction and summarization of this information still represents a significant challenge. With PubMed growing at a double-exponential rate, more than two citations are added each minute. Since 2000, there was nearly a 600% increase in pharmacogenetics-related publications. Automated approaches to extract information concerning new alleles and to identify their impact on
metabolic phenotypes are urgently needed to further advance evidence-based personalized medicine. Extraction of pharmacogenetic knowledge from PubMed and other sources has been discussed in the recent literature. Various tools, including text mining and natural language processing (NLP) for broad aggregation of information related to genetic polymorphisms have been described. For example, Rubin et al developed an automatic system to obtain PubMed abstracts containing some pharmacogenetics knowledge, including gene products, drug effects, genetic variants, and the relationships among them, while Garten et al established Pharmpresso, a text-mining tool used for extracting pharmacogenomic concepts from the literature and determining gene-drug interactions. However, limited attention has been given to development of a specialized tool for targeted extraction of cytochrome polymorphisms and their characterization along the metabolic phenotype continuum. Introduction of such a tool can significantly enhance pharmacogenomics-driven clinical decision support tools for personalized medication optimization, reduce adverse effects and improve patient care.

The goal of this project is (1) to develop a system for automated metabolic phenotyping of cytochrome polymorphisms based on currently available information in PubMed abstracts; and (2) to estimate its precision in a preliminary proof-of-concept assessment.

Materials and Methods

Materials
This study was conducted using articles from PubMed and a genetic polymorphism reference released by EBM Consult. The articles used in this study consisted of the titles and abstracts associated with the variants of five cytochromes (CYP2C9, CYP1A2, CYP2C19, CYP2D6 and CYP3A4), which was obtained using our corpus generator. The genetic polymorphism reference tables from EBM Consult, which contains verified information about metabolic functions of the five cytochrome polymorphisms, was employed as a gold standard to evaluate the efficacy of our system. The article set was divided into a training set and test set. The training set was used for development only, and contained articles relevant to CYP2C9, while the articles from the remaining four cytochromes constituted the test set, and was used for evaluation.

Method

Overview

An overview of our system is shown in Figure 1, and a more detailed explanation is provided below. The system consists of four components: 1) the corpus generator, which performs PubMed queries via the Entrez Programming Utilities (E-utilities) to obtain titles and abstracts of articles that potentially contain gene variants, 2) the gene variant extractor, which applies regular expressions to recognize and extract gene variants from the articles obtained by the corpus generator, 3) the signal word fetcher, which performs stemming on the words in the articles obtained by the corpus generator and matches the stemmed terms with a predefined dictionary to fetch a set of words that denote the metabolic phenotypes (i.e. rapid metabolizer, intermediate metabolizer, poor metabolizer and normal metabolizer), and 4) the scorer, which determines the closest gene variant-signal word pairs as evidence of normal or aberrant metabolic functions, and then specifies the type of metabolic function with the highest score as the predicted result.

Figure 1. Overview
Corpus Generator

The first component of this system is the corpus generator, which aims to obtain a relevant corpus of PubMed abstracts related to polymorphisms. A query was designed for each cytochrome to select related articles via the Application Program Interface (API) of National Library of Medicine, Entrez Programming Utilities \(^{19}\). Basically, to acquire relevant evidence for variants of a specific gene, articles discussing the specific gene are initially retrieved, where the articles are restricted using the criteria that they are written in English, concern humans, and are not review articles. Review articles are excluded for two reasons. First, review articles are not original articles but collections of information that were reported in other original articles. Hence, review articles would not provide any new information even if they were included. Moreover, using review articles would introduce bias in our study since the information in those articles would be counted twice.

Therefore, this query contains 5 features, consisting of the gene name, the MeSH tags English, human, the negation of the MeSH tag animal (to ensure that the article was mainly about humans), and the negation of the publication tag review. An example query for CYP2C9 is shown below:

\[
\text{CYP2C9 AND Human NOT Animal AND Eng[LA] NOT review[PT]}
\]

Once this type of query returns the PubMed identifiers (PMIDs) of relevant articles, the E-utilities library is then used to download the citations of those articles from the MEDLINE database.

However, the articles returned from the above query were published by a very broad set of journals, where some journals are less relevant to pharmacology or pharmacogenomics. To determine a list of the more relevant journals, we used E-utilities to search for the broad subject term Pharmacology or Genetics from the NLM Catalog \(^{20}\) and then filtered out articles published in journals that were not on this list. The simple query used to obtain this list from the Catalog was:

\[
\text{pharmacology[Broad Subject Term] OR genetics[Broad Subject Term]}
\]

Gene Variant Extractor

After obtaining articles relevant to this task, the gene variant extractor recognizes and extracts gene variant mentions from their abstracts and titles. The gene names and their variants follow specific conventions most of the time. For example, the cytochrome p-450 family genes are usually written as CYP+number+letter+number, and then use ‘*’ at the end of the name followed by an integer to indicate a variant. This convention enables the use of regular expressions to identify the gene variants. By reviewing articles and analyzing features in our training set, we developed a regular expression to specify gene variants of the CYP gene family, as follows:

\[
\W?(\text{CYP}).*(\d+[a-zA-z]\d+)\D?.*(\*\d+).\D?
\]

Although the regular expressions above could recognize valid gene variants in text, a practical problem arose when processing the text. For example, gene variants in articles could be written in an abbreviated way as “… CYP2C9*3 (I359L), *13 (L90P), *26 (T130R), *28 (Q214L), *30 (A477T) …” or “CYP2C9 *33 and *34”. Although the regular expression works for an individual variant, in these cases it could not recognize all the variants of the same gene because of the repetition of the different variants of the gene. In order to solve this problem, we noted that the gene name and variants form a tree structure in text, where the gene name is the parent and all variants are its children, as shown in Figure 2. Each child with its parent could constitute of the comprehensive name of a gene variant (see Figure 2 below).

![Figure 2. The tree structure of gene variant names](image-url)
Using this property, we designed two regular expressions to match the name of gene and indication of variant respectively, drew the tree structure for each gene by tracking the index of each matched gene name or variant, and acquired all the gene variants mentioned in the text from the tree structure alternatively.

The two regular expressions for this is shown below:

Gene name: \W?(CYP).*(\d+[a-zA-z]\d+).D?

Variants: (\*d+).D?

**Signal Word Fetcher**

The metabolic functions of gene variants are usually indicated through a limited set of special key words, such as *poor* or *rapid*, which indicate the metabolic phenotypes specified as signal words in the articles. A signal word fetcher was designed to obtain signal words based on a predefined dictionary of signal words, which was established using the training set in the following steps. First, a Porter stemmer was used to stem the words in text (abstracts and titles). After getting the stemmed text, the Complex Sentiment Analysis tool of TAPoR was applied on the stemmed text to analyze the frequency of words which could show positive indication or negative indication in a predefined vocabulary within the tool. The most frequent words in the predefined vocabulary were manually reviewed and selected as signal words, which were included in the dictionary. Specifically, those signal words were manually divided into four categories, in order to meet the general classification of metabolic functions, consisting of “normal”, “rapid”, “intermediate” and “poor”. Some signal words for each category are shown in the table below:

<table>
<thead>
<tr>
<th>Category</th>
<th>Some signal words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>wild, normal, extensive</td>
</tr>
<tr>
<td>Rapid</td>
<td>rapid, ultrarapid, increase, accelerate, induce, prompt</td>
</tr>
<tr>
<td>Intermediate</td>
<td>impair, inhibit, decrease, reduce, slow, inactive</td>
</tr>
<tr>
<td>Poor</td>
<td>poor, absent</td>
</tr>
</tbody>
</table>

Using the pre-defined dictionary that was established, the system uses the same Porter stemmer to preprocess the text in the test set and identify signal words by a string matching method.

**Scorer**

The last component, the scorer, uses the gene variants and signal words extracted by the gene variant extractor and signal word fetcher to predict the metabolic phenotypes for each variant. The criteria use the closest gene variant-signal word pairs in the text as evidence for scoring. Assume \( P_{t,j} \) is the score representing that gene variant \( i \) belongs to category \( j \), where \( j \) can be “normal”, “rapid”, “intermediate” and “poor” as mentioned above, where the scores could be computed using the formula below:

\[
P_{t,j} = \sum_{k \text{ is a signal word in category } j} \text{ number of pairs } (i, k)
\]

Once the scores are calculated, they are organized into a table with scores for specifying each variant to each category of metabolism, where the category with the highest score is chosen as the predicted result for each gene variant. For example, CYP2C19 *5, which had 0 points for “normal”, 3 points for “rapid”, 8 points for “intermediate” and 9 points for “poor”, was assigned to “poor” category because of the highest score. If there is a tie among some categories of a variant, all these categories are assigned to it as potential candidates. However, when evaluating the results, a gene variant with multiple categories is considered to be an incorrect prediction.

**Evaluation Design**

The evaluation test set for this task consisted of 4 cytochromes (CYP1A2, CYP2C19, CYP2D6 and CYP3A4), whose variants have been well studied and the types of metabolic functions have been established by EBM Consult.
For each cytochrome, we applied the system we developed to obtain relevant articles related to the cytochrome and to acquire the predicted metabolic functioning for all its variants. The predicted results were compared and evaluated using the reference released by EBM Consult. Note that the reference table is incomplete, and only the variants recorded in the reference table were used for evaluation. The precision of the predictions was calculated using the formula below:

\[
\text{Precision} = \frac{\text{number of correct predictions}}{\text{number of total predictions}}
\]

As stated in the Scorer subsection above, each tie involving multiple categories was counted as a false positive. Sensitivity was not directly computed or reported in the results section for several reasons, which are described in the discussion section. In addition, we did not consider variants which we obtained that were not in the reference table as false positives. The reasons are also explained below.

Results

The entire system was developed using the Anaconda of Python 2.7 version with useful packages/tools including regular expressions (“re”), Natural Language Toolkit (“NLTK”) and Complex Sentiment Analysis tool of TAPoR. In addition, the web API, E-utilities was used to obtain the articles for the corpus generator.

Table 2 shows the evaluation results for our system broken down by the 4 cytochromes. In total, 35 variants with verified information were recognized and predictions were generated by our system, where 30 of the 35 were correctly predicted. The total precision is 85.71%, which proved the efficacy of our algorithm. Note: the wild type (*1) is not counted here since it is always normal.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Correct predictions</th>
<th>Total predictions</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>CYP1A2</td>
<td>4</td>
<td>4</td>
<td>100.00%</td>
</tr>
<tr>
<td>CYP3A4</td>
<td>15</td>
<td>16</td>
<td>93.75%</td>
</tr>
<tr>
<td>CYP2D6</td>
<td>7</td>
<td>10</td>
<td>70.00%</td>
</tr>
<tr>
<td>CYP2C19</td>
<td>4</td>
<td>5</td>
<td>80.00%</td>
</tr>
<tr>
<td>Total</td>
<td>30</td>
<td>35</td>
<td>85.71%</td>
</tr>
</tbody>
</table>

Table 3 shows the distribution of our predicted results. We can see that all the incorrect predictions came from the “poor” category.

<table>
<thead>
<tr>
<th>Category</th>
<th>Correct Predictions</th>
<th>Total Predictions</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>2</td>
<td>2</td>
<td>100.00%</td>
</tr>
<tr>
<td>Rapid</td>
<td>2</td>
<td>2</td>
<td>100.00%</td>
</tr>
<tr>
<td>Intermediate</td>
<td>18</td>
<td>18</td>
<td>100.00%</td>
</tr>
<tr>
<td>Poor</td>
<td>8</td>
<td>13</td>
<td>61.54%</td>
</tr>
<tr>
<td>Total</td>
<td>30</td>
<td>35</td>
<td>85.71%</td>
</tr>
</tbody>
</table>

The 5 prediction errors are shown in detail in table 4. The predication errors were due to predicting “poor” metabolizers as “intermediate” metabolizers. The error analysis is further explained in the discussion.
Table 4. Prediction errors

<table>
<thead>
<tr>
<th>Gene Variants</th>
<th>Annotated Category</th>
<th>Predicted Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>CYP3A4 *14</td>
<td>Poor</td>
<td>Intermediate</td>
</tr>
<tr>
<td>CYP2D6 *29</td>
<td>Poor</td>
<td>Intermediate</td>
</tr>
<tr>
<td>CYP2D6 *5</td>
<td>Poor</td>
<td>Intermediate</td>
</tr>
<tr>
<td>CYP2D6 *9</td>
<td>Poor</td>
<td>Intermediate</td>
</tr>
<tr>
<td>CYP2C19 *2</td>
<td>Poor</td>
<td>Intermediate</td>
</tr>
</tbody>
</table>

Discussion

In this paper, an automated system for metabolic phenotyping of cytochrome polymorphisms using PubMed abstract mining has been introduced. Initial evaluation based on a current reference standard for clinical care showed promising results with a precision of 85.71%. The main strength of our approach is its specific focus on cytochrome polymorphisms and its association with particular metabolic phenotypes. As this information is crucial for implementing pharmacogenomics-driven clinical decision support (CDS) at point of care, automated tools which may constantly update an underlying CDS knowledge repository based on the most recent evidence have significant potential in promoting evidence-based precision medicine.

Our results conform to previous reports which demonstrated an ability to query the “bibliome” (compendium of biomedical textual resources) in order to answer questions about genes and their associations. Successful efforts in mining pharmacogenomics literature resulted in automatic recognition of various relevant named entities including genes, gene variants and corresponding proteins, associated diseases and pathological syndromes, medications and other factors pertinent to clinical care. “Bibliome” mining was shown to automatically convert massive volumes of unstructured knowledge into a structured actionable form supporting scientific discovery and clinical care. These approaches were used for interpretation of findings in genome-wide association studies as well as identification of candidate gene lists. Pharmpresso is an example of one of successful tools for text analysis aimed at extracting pharmacogenomic concepts from the literature and determining gene-drug interactions. Approaches for mining pharmacogenic literature are widely used to support the Pharmacogenomics Knowledgebase (PharmGKB) and work of the Clinical Pharmacogenetics Implementation Consortium (CPIC).

In recent years, a number of NLP systems have been developed to retrieve information from biomedical text, especially to extract biomedical information from the literature, demonstrating that NLP methods could efficiently generate information through abstracts mining. There has been very limited work regarding automated metabolic phenotyping of cytochrome polymorphisms using natural language processing. Rubin et al developed an automatic system to obtain PubMed abstracts containing some pharmacogenetics knowledge, including gene products, drug effects, genetic variants, and the relationships among them, by using the Porter Stemmer to preprocess the articles and then implementing supervised machine learning algorithms to classify the most relevant articles based on relevant pharmacogenetics terms; they achieved the best F1-score of 0.82. Yeniterzi et al implemented an algorithm based on regular expressions and supervised machine learning models to automatically identify the PubMed abstracts containing information on protein level mutations which have impact on a given enzyme, and achieved greater than 85% accuracy for all categories. Coulet et al, started from a preliminary pharmacogenic entities lexicon and implemented a drug-gene interactions extractor by detecting the syntactic structure of the relevant entities within the statements of PubMed abstracts. Based on Coulet’s work, Percha et al designed a supervised machine learning model to infer drug-drug interactions from the drug-gene interactions. Mallory et al used DeepDive, a supervised information extraction tool, to extract gene interactions from over 100 000 full-text articles.

While the studies above achieved desirable efficacy for text mining tasks, most of them used either large sets of references (e.g. lexicon for entities) or supervised machine learning methods to identify relations such as drug-gene interactions, or to classify the focus of articles based on a set of annotated data. Such methods are not applicable in our task. A significant challenge for our task is that, to the best of our knowledge, there is no annotated corpora for cytochrome polymorphisms and that reference standards to help validate results are usually incomplete. Hence, approaches using large external reference sets or supervised machine learning algorithms could not be used for our task. Therefore, we used rule-based NLP methods and scoring based on a development set, and the results based on
the text set showed that the metabolic phenotypes of cytochrome variants were successfully extracted. Although  
regular expressions are simple NLP methods, it is practical to study performance using simple methods before using  
more complex ones, especially if they require an annotated corpus that has not yet been generated. If machine  
learning methods are developed in the future, the simple regular expression method could be used as a baseline.  

There were 46 gene variants obtained by the system that were not in the reference table, as shown in Table 5. They  
were not considered as false positives because the reference table was incomplete. The reference table from EBM  
Consult was used for a clinical decision support application, and was established as being 100% correct to avoid  
errors in the system. Only gene variants whose metabolic functions had been generally accepted by physicians and  
researchers were included in the reference table, and consequently some variants whose metabolic phenotypes were  
not generally accepted or had not yet been discovered by those generating the reference might not be included in it.  
For example, CYP3A4*22 which was predicted as an “intermediate” metabolizer by our system was not included in  
the reference standard we used, but was listed as the same metabolic phenotype in the Human Cytochrome P450  
(CYP) Allele Nomenclature Database, which was not chosen as our reference standard because this database is  
maintained by an individual webmaster and the reliability of this database is unknown. In addition, for gene variants  
whose metabolic functions were not yet recorded in any reference set, they might have been discovered more  
recently by researchers, or they may not have been included because they still needed more evidence to confirm the  
functioning. Our system could point to potentially new reference material since it finds related articles and predicts a  
possible result which could be further studied and judged by experts upon manually reading and assessing the  
information that was generated by the system. Hence, gene variants obtained by our program but had no  
corresponding metabolic functioning in the reference table might also be correct and should not be treated as false  
positives.

Table 5. 46 Variants beyond the reference standard

<table>
<thead>
<tr>
<th>Gene Variants</th>
<th>Metabolic Functions</th>
<th>Gene Variants</th>
<th>Metabolic Functions</th>
</tr>
</thead>
<tbody>
<tr>
<td>CYP1A2*2</td>
<td>intermediate</td>
<td>CYP2D6*20</td>
<td>intermediate</td>
</tr>
<tr>
<td>CYP1A2*5</td>
<td>intermediate</td>
<td>CYP2D6*21</td>
<td>intermediate</td>
</tr>
<tr>
<td>CYP1A2*10</td>
<td>intermediate</td>
<td>CYP2D6*25</td>
<td>intermediate</td>
</tr>
<tr>
<td>CYP1A2*17</td>
<td>rapid or intermediate</td>
<td>CYP2D6*26</td>
<td>rapid or intermediate</td>
</tr>
<tr>
<td>CYP1A2*28</td>
<td>intermediate</td>
<td>CYP2D6*28</td>
<td>intermediate</td>
</tr>
<tr>
<td>CYP2C19*6</td>
<td>intermediate</td>
<td>CYP2D6*30</td>
<td>intermediate</td>
</tr>
<tr>
<td>CYP2C19*7</td>
<td>intermediate</td>
<td>CYP2D6*31</td>
<td>poor</td>
</tr>
<tr>
<td>CYP2C19*8</td>
<td>intermediate</td>
<td>CYP2D6*33</td>
<td>normal</td>
</tr>
<tr>
<td>CYP2C19*9</td>
<td>intermediate</td>
<td>CYP2D6*35</td>
<td>rapid</td>
</tr>
<tr>
<td>CYP2C19*10</td>
<td>intermediate</td>
<td>CYP2D6*36</td>
<td>poor</td>
</tr>
<tr>
<td>CYP2C19*15</td>
<td>intermediate</td>
<td>CYP2D6*37</td>
<td>intermediate</td>
</tr>
<tr>
<td>CYP2C19*27</td>
<td>intermediate</td>
<td>CYP2D6*39</td>
<td>normal</td>
</tr>
<tr>
<td>CYP2C19*28</td>
<td>intermediate</td>
<td>CYP2D6*40</td>
<td>poor</td>
</tr>
<tr>
<td>CYP2D6*2</td>
<td>rapid</td>
<td>CYP2D6*43</td>
<td>intermediate</td>
</tr>
<tr>
<td>CYP2D6*7</td>
<td>poor</td>
<td>CYP2D6*49</td>
<td>intermediate</td>
</tr>
<tr>
<td>CYP2D6*8</td>
<td>poor</td>
<td>CYP2D6*53</td>
<td>rapid</td>
</tr>
<tr>
<td>CYP2D6*11</td>
<td>poor</td>
<td>CYP2D6*58</td>
<td>intermediate</td>
</tr>
<tr>
<td>CYP2D6*12</td>
<td>poor</td>
<td>CYP2D6*64</td>
<td>intermediate</td>
</tr>
<tr>
<td>CYP2D6*13</td>
<td>poor</td>
<td>CYP2D6*65</td>
<td>poor</td>
</tr>
</tbody>
</table>
Similarly, we did not compute sensitivity because some metabolic phenotypes recorded in the reference standard but not found by our system should not be considered as false negatives. First, the reference standard contained in-vitro or animal-related metabolic functions of some gene variants. Those were not determined by our system because we deliberately used articles discussing metabolic functions in humans and removed articles recording others, and therefore the in-vitro or animal articles should be excluded in the results. This was the major cause of missing the metabolic phenotypes of those variants. Additionally, some metabolic phenotypes in the reference were from review articles, or from articles published in journals considered not relevant, and those articles were excluded as we explained in the methods. Secondly, some articles did not mention the variants in the abstracts or titles, and we could not find them since PubMed citations only contain the titles and abstracts.

There were some errors in our system that did cause actual false negatives. Some were caused by articles (especially articles early stage) which mentioned the gene variants in irregular ways (i.e. used “dot+number” like “.7” to refer the variant, or directly used its abbreviation without mentioning its general name), and the regular expressions were not able to handle such cases. This could be improved by using more complicated regular expressions and introducing abbreviation vocabularies.

In total, our system missed 7 cytochrome polymorphisms, as shown in Table 6. If all of them were considered as false negatives, the sensitivity here would be 81.08%. However, five of them are in-vitro phenotypes and one was missing because there was no mention in the abstract. Those 6 variants should not be considered as false negatives as stated above. But for CYP3A4*7, the reasons for the error were more complicated. In total, there are 4 articles in PubMed discussing CYP3A4*7. Two of them were published in journals we excluded, and the remaining two used the term “CYP3A4.7” to name this variant. Thus, this variant should be counted as false negative. If we excluded the first 6 misses as false negatives, we would have 30 true positives and 1 false negative, and the sensitivity of our system would be 96.77%.

<table>
<thead>
<tr>
<th>Gene Variants</th>
<th>Reason(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CYP2D6*15</td>
<td>intermediate</td>
</tr>
<tr>
<td>CYP2D6*16</td>
<td>poor</td>
</tr>
<tr>
<td>CYP2D6*18</td>
<td>intermediate</td>
</tr>
<tr>
<td>CYP2D6*19</td>
<td>intermediate</td>
</tr>
<tr>
<td>CYP2D6*66</td>
<td>poor</td>
</tr>
<tr>
<td>CYP2D6*75</td>
<td>intermediate</td>
</tr>
<tr>
<td>CYP2D6*82</td>
<td>intermediate</td>
</tr>
<tr>
<td>CYP3A4*22</td>
<td>intermediate</td>
</tr>
</tbody>
</table>

Table 6. All missing cytochrome variants and reasons

As shown in Tables 3 and 4, all the problems in our prediction were incorrectly predicting “poor” metabolizers as “intermediate” metabolizers. There are several reasons for these errors. In general, “intermediate” indicates a variant which will inhibit the metabolic activity, while “poor” indicates inhibition of a higher level of inhibition. Thus, if a variant is a poor metabolizer, it could also qualify as an intermediate metabolizer, especially when the degree is difficult to capture. Consequently, only explicit signal words, “poor” and “absent”, indicate poor metabolizers, while other negative signal words that are more numerous can indicate intermediate metabolizers, and thus were assigned to the “intermediate” category. However, those negative signal words for “intermediate” category might also occur around “poor” metabolizers in the text to indicate negative effects on the metabolic activity. In addition, explicit signal words for poor metabolizers occurred much less frequently than other negative signal words associated with intermediate metabolizers, affecting scoring. For example, “CYP2D6*9 is an allelic variant conferring reduced enzymatic activity...” is a sentence from a paper in our corpus 29, which has a relation between “CYP2D6*9” and
“reduce”. However, “CYP2D6*9” is a poor metabolizer but “reduce” is a signal word for intermediate metabolizers. Another source of error occurred when pairing the variant and signal word because we simply chose the closest pair since it demonstrated a relatively good performance during development. However, such a pairing strategy was inaccurate for some situations, especially when multiple variants were mentioned. For instance, in the sentence “The proportions of non-functional alleles were 0.4, 10.6, 0.8, 2.1, and 0% for CYP2D6*3, *4, *4 x N, *5, and *6, respectively. Genotypically, only one of the subjects (0.9%) was homozygous for two inactive alleles and phenotypically classified as a poor metabolizer (PM)” 30, “CYP2D6*5” was paired to “inactive” rather than “poor” because “inactive” was closer to “CYP2D6*5”. Another source of error occurred for some variants which were associated with less than five articles. For these, the metabolic functions were more difficult to predict accurately due to the lack of evidence.

Future work will involve addressing some of the errors discussed above. We intend to establish a more sophisticated and accurate classification system based on use of contextual information in the abstracts to distinguish poor metabolizer and intermediate ones. This should be the most significant future work direction regarding NLP. However, since there is no annotated corpus available at this time, we will generate one. Also, the simple closest pairing method will be improved. The problem concerning a lack of articles cannot easily be solved, but this lack-of-evidence problem is expected to be improved as more research and articles are published. However, even though our predictions for those variants that are infrequently published may not be accurate regarding functionality, they still contain useful clinical information, and can be shown with a lower confidence.

In addition to the prediction system, the evaluation method also needs improvement. In the current study, only precision was computed due to the lack of a comprehensive reference set and annotated data. In the future, more comprehensive evaluation of our algorithm will be carried out based on manual review of the resulting articles. With constantly growing body of pharmacogenomics literature and new alleles and phenotypic associations reported in PubMed, tools to support decision making at the point of care should have ability to account for new evidence. The algorithm presented in this paper could be applied to automatically extract the metabolic phenotypes for cytochrome variants and assist in updating the relevant knowledge. It can summarize existing evidence by providing an estimate of phenotypic impact of new and current alleles based on the current state of literature even for alleles which metabolic phenotype is not yet reflected in authoritative sources. In addition, this algorithm could be adapted to other gene variants by simply changing the regular expressions, and is expected to have equally desirable results as this study. Overall, automated metabolic phenotyping of cytochrome gene variants can assist physicians in personalizing drug therapy and avoiding adverse drug reactions resulting in better clinical outcomes.

Conclusion

This paper presents a system to obtain currently available PubMed abstracts of a given cytochrome, extract gene variants from the retrieved articles, and predict the metabolic phenotypes based on the information in the retrieved abstracts. The preliminary evaluation resulted in a proof-of-concept assessment and demonstrated the efficacy and accuracy of this system, which was 85.71%. Our initial results are very promising and have shown that it would be feasible to use this system for automated metabolic phenotyping of cytochrome polymorphisms from the continually growing information from PubMed abstracts.

References

Large-scale Analysis of Opioid Poisoning Related Hospital Visits in New York State

Xin Chen, MS, Yu Wang, BS, Xiaxia Yu PhD, Elinor Schoenfeld, PhD, Mary Saltz, MD, Joel Saltz, MD, Fusheng Wang, PhD
Stony Brook University, Stony Brook, NY

Abstract
Opioid related deaths are increasing dramatically in recent years, and opioid epidemic is worsening in the United States. Combating opioid epidemic becomes a high priority for both the U.S. government and local governments such as New York State. Analyzing patient level opioid related hospital visits provides a data driven approach to discover both spatial and temporal patterns and identity potential causes of opioid related deaths, which provides essential knowledge for governments on decision making. In this paper, we analyzed opioid poisoning related hospital visits using New York State SPARCS data, which provides diagnoses of patients in hospital visits. We identified all patients with primary diagnosis as opioid poisoning from 2010-2014 for our main studies, and from 2003-2014 for temporal trend studies. We performed demographical based studies, and summarized the historical trends of opioid poisoning. We used frequent item mining to find co-occurrences of diagnoses for possible causes of poisoning or effects from poisoning. We provided zip code level spatial analysis to detect local spatial clusters, and studied potential correlations between opioid poisoning and demographic and social-economic factors.

Introduction
The United States is experiencing an epidemic of opioid related deaths. Since 2000, the rate of deaths from drug overdoses has increased 137%, including a 200% increase in the rate of overdose deaths involving opioids\(^1\). Given these alarming trends, combating opioid epidemic becomes a high priority for both the US government and local governments such as New York State. For example, the U.S. Department of Health and Human Services (HHS) has implemented evidence-based approaches to reduce opioid overdoses and the prevalence of opioid use disorder\(^2\). The MOON (Maximizing OpiOid safety with Naloxone) study tries to find out more information about the public's perception of opioid safety, naloxone distribution, and the use of the pharmacy as an integral site for public health intervention\(^3\).

With increased accessibility of health data driven by open data initiatives, large-scale patient level data analysis provides an opportunity of data driven approach to identity patterns of opioid epidemic and discover potential causes of opioid related deaths through studying large scale diagnoses from hospital visits. This will provide quantitative measurements and essential knowledge to support the governments for decision making on prevention, treatment, recovery and enforcement. Analyzing opioid data at fine spatial resolutions such as zip code can also reveal community level distribution patterns in terms of demography, regions and historical trends, which will provide crucial knowledge for residents, schools, businesses and health and human service professionals for seeking solutions.

As part of New York State’s open data initiative, New York State Statewide Planning and Research Cooperative System (SPARCS)\(^6\) collects patient level details on patient characteristics, diagnoses and treatments, services, and charges for each inpatient stay and outpatient visit (emergency department, ambulatory surgery, and outpatient services). It also includes locations of patients (street addresses). Researchers can benefit from SPARCS data by leveraging patients’ diagnosis histories, co-occurrences of diagnoses (primary and secondary ones), and location. Such unprecedented amounts of patient records make it possible for us to explore opioid poisoning in New York State with significant improvement of accuracy and coverage compared to prior studies on opioid abuse at global level\(^4\) or urban level\(^5\).

In this paper, we provide large scale analysis of opioid poisoning based on SPARCS data. We extract all patient records with opioid poisoning related primary diagnosis codes from all visits (inpatients, outpatients, emergency and ambulatory surgeries) from year 2010-2014 for our main studies, and from year 2003-2014 for temporal trend studies. We aim at four types of studies: 1) demographical based studies to explore disparities between population groups; 2) historical trends of opioid poisoning; 3) co-occurrence based studies to discover possible causes of opioid poisoning or effects from opioid poisoning; and 4) spatial analysis (zip code level) to explore potential local spatial clusters, and potential correlations between opioid poisoning and demographic and social-economic factors.
Methods

Data Sources

SPARCS Data. In this work, we used hospital discharge data from New York State SPARCS\(^6\). Any New York State healthcare facility certified to provide inpatient services, ambulatory surgery services, emergency department services or outpatient services is required to submit data to SPARCS. The purpose of SPARCS was to create a statewide data set to contribute to the goal of providing high quality medical care by serving as an information source\(^7\).

Opioid Poisoning Hospital Visits. While the SPARCS data contains a comprehensive list of diagnosis and treatment procedure codes for each discharge record, this paper focused on analyzing patient level patterns of opioid poisoning hospital visits.

In SPARCS data, each hospital discharge record contains only one primary diagnosis code. Each record also contains one or more optional secondary diagnosis codes that include all conditions that coexisted at the time of admission, or developed subsequently, which affected the treatment received and/or length of stay\(^7\).

In this work, we generated a subset of the SPARCS data using selected the International Classification of Diseases 9 (ICD-9) codes (also called billing codes) that pertain to opioid poisoning diagnoses, for both inpatient stays and outpatient visits (including emergency department visits, ambulatory surgery, and outpatient visits). We extracted opioid poisoning hospital visits by filtering the discharge records with their primary diagnosis code as the selected opioid poisoning ICD-9 codes.

The selected ICD-9 codes were a collection of poisonings by opiates, opium, heroin, methadone, and other related narcotics, including 9650 (Poisoning; Opiates and Related Narcotics), 96500 (Poisoning; Opium/alkaloids, unspecified), 96501 (Poisoning; Heroin), 96502 (Poisoning; Methadone), 96509 (Poisoning; Other opiates and related narcotics), E8500 (Accidental Poisoning; Heroin), E8501 (Accidental Poisoning; Methadone), and E8502 (Accidental Poisoning; Other Opiates and Related Narcotics).

<table>
<thead>
<tr>
<th>Table 1. Demographics of patients with opioid related hospital visits, New York State, 2010-2014.</th>
</tr>
</thead>
<tbody>
<tr>
<td>New York State Population*</td>
</tr>
<tr>
<td>-----------------------------------------------</td>
</tr>
<tr>
<td>19,594,330</td>
</tr>
<tr>
<td>Age and Sex</td>
</tr>
<tr>
<td>Under 5 years</td>
</tr>
<tr>
<td>5 to 14 years</td>
</tr>
<tr>
<td>15 to 44 years</td>
</tr>
<tr>
<td>65 years and over</td>
</tr>
<tr>
<td>Female persons</td>
</tr>
<tr>
<td>Race and Ethnic</td>
</tr>
<tr>
<td>White alone</td>
</tr>
<tr>
<td>African American alone</td>
</tr>
<tr>
<td>Asian alone</td>
</tr>
<tr>
<td>Hispanic or Latino</td>
</tr>
</tbody>
</table>

* 2010-2014 American Community Survey (ACS) 5-year estimates

Patients’ Residential ZIP Codes and Hospital Admission Years. We conducted analyses at ZIP code level with basic demographics for the year 2010-2014 given in Table 1. We approximated patients’ home location by combining the 5-digit ZIP code number with the patients’ home address and geographic data from TIGER/LINE data\(^8\). We then aggregated opioid poisoning hospital visits at ZIP code level (Figure 1) and generated opioid poisoning incidence rates for the following spatial and temporal analyses (Figure 4-5).

For temporal trend studies, we used the hospital admission year for the time range 2005-2014 for inpatient stays and 2003-2014 for emergency department visits (Figure 3).

Analysis Methods

Opioid Poisoning Incidence Rates
We counted opioid poisoning patients with at least one opioid poisoning hospital at ZIP code level as shown in Figure 1. The opioid poisoning incidence rates by ZIP code were calculated through dividing counts of patients with opioid poisoning hospital visits by ZIP level population counts from Census data. In this paper, we evaluated both statewide rate and rates at ZIP code level. The incidence rates provided useful information about how common a disease is when compared to other diseases, or how common a disease in a specific location is as compared to the global baseline.

**Frequent Co-Occurrence Patterns**

Disease co-occurrence, also known as comorbidity⁹⁻¹⁰ in public health studies, may imply the potential association across different types of diseases. The comorbidity is the presence of one or more additional diseases or disorders co-occurring with a primary disease or disorder¹⁸.

We explored frequent disease co-occurrence patterns to find the frequent disease diagnoses that co-exist with opioid poisoning diagnoses. We first ranked all the secondary diagnosis codes by their counts in all the opioid poisoning hospital visits (Table 2). We then use Apriori-like algorithm¹¹ to discover top co-occurrences of diseases for opioid poisoning diagnoses (Table 3).

Apriori Algorithm is a common data mining technique to identify co-occurrences or temporal patterns between diseases in clinical domain¹². This work tried to adopt Apriori algorithm to identify comorbidities among hospital visits. Apriori algorithm discovers frequent comorbidities by comparing their supports with a user-specified minimum support threshold.

For example, if the support of comorbidity pattern {96500 Poisoning by opium, 9670 Poisoning by barbiturates} (in Table 3) is 0.002, it means that 0.2% of all hospital visits have this comorbidity pattern. If the minimum support threshold is greater than 0.2%, this pattern will not be identified. However, if the minimum support threshold is set smaller than 0.2%, the comorbidity pattern will be extracted. In this work, we set the minimum support threshold as 0.1%.

**Spatial Clustering**

To test whether there is global spatial clustering tendency for opioid poisoning incidence rate or other spatial impact factors, we used Moran’s I (Tables 4)¹³. Moran’s I is a widely used global cluster test, which determines the degree of clustering or dispersion within a data set. The test result may range from 1 (perfect correlation), 0 (complete spatial randomness) to -1 (perfect dispersed). For the opioid poisoning incidence rate, a positive spatial autocorrelation means that the ZIP code areas with high rate are close to other areas with high rates.
In addition to the global cluster test (with Moran’s I index in Table 4) and visual analysis for mapping raw counts of opioid poisoning incidence (in Figure 1), we also took cluster and outlier analysis with Anselin Local Moran’s I statistics\textsuperscript{13} to quantitatively detect local clusters for opioid poisoning incidence rates (Figure 4). The Anselin Local Moran’s I statistics is a local cluster test that, given a set of weighted features, identifies statistically significant hot spots, cold spots, and spatial outliers.

*Spatial Regression*

To identify potential correlations between diseases and spatial impact indicators, we assessed both non-spatial and spatial correlation\textsuperscript{13}. Ordinary least square regression analyses (OLS) was used to determined non-spatial correlation between opioid poisoning and demographic/socio-economic factors. We then used Geographically Weighted Regression (GWR) to assess the spatially varying relationship at a local level.

OLS is a linear regression method that closely fits a function by minimizing the sum of squared errors. To determine potential candidate factors, we evaluated several possible factor combinations that form a properly specified OLS regression model. GWR is a local form of linear regression used to model spatially varying relationships. To evaluate the correlation between opioid poisoning and spatial impact factors accounting for data in surrounding areas, we then conducted GWR with a selected spatial impact factor based on OLS results. Specifically, a fixed kernel type function was used to calculate the GWR regression coefficients. The extent of the kernel is determined using the Akaike Information Criterion (AICc)\textsuperscript{14-15}.

*Results*

*Demographic-based Analysis*

Table 1 compared the demographics between New York State population and patients with opioid poisoning during 2010-2014. The overall opioid poisoning incidence rate in New York State was 12.33 per 10,000 persons per five years. The percentage of female in New York State were 51.5% compared with only 40.0% of female patients with opioid poisoning, suggesting that female was less likely to have opioid poisoning hospital visits.

![Figure 2. The age distribution of opioid poisoning in New York State, 2010-2014.](image)

Comparing different race and ethnicity groups, we can see significant disparities between racial-ethnic proportions in patient and general population. Whites made 72.2% of patients that was higher than the whites proportion in general
population. On the contrary, Asians had a very low percentage of patients that may require further investigation for the specific causes.

Figure 2 showed the age distribution of opioid poisoning in New York State, 2010-2014. We found a peak of opioid poisoning among young adults aged 21 to 25. Prior work has shown an increasing trend of adolescents and young adults with prescription drugs misuse. Our statistics on age for opioid poisoning hospital visits (Figure 2) was consistent with the prior national-wide survey.

Temporal Trends

Outpatient emergency departments (EDs) and inpatient stays play an important role in the treatment of drug poisoning. For the period of most recent year available 2010-2014, both opioid poisoning hospital EDs and inpatient stays continued to increase in the United States, as is true for New York State from our results. Statewide, there were 4,238 outpatient ED visits and 2,909 inpatient stays in 2014, a 116.4 percent increase and a 17.3 percent increase respectively from 2010.

Figure 3 showed temporal trends of the proportion of opioid poisoning hospital visits for the past decade. While we found a modest rising trend for inpatient stays, there was an increasingly rising trend for outpatient EDs. In prior studies, the rise in opioid consumption has also been associated with a sharp increase in emergency room visits for nonmedical opioid pain reliever (OPR) use. Such findings suggested potential misuse or abuse of the drugs as opiate overdoses may seek controlled substances during emergency department visits.

Frequent Co-Occurrence Patterns

We used frequent item mining to find co-occurrences of diagnoses for possible causes of poisoning or effects from poisoning. We used all ICD-9 diagnosis codes from opioid poisoning hospital discharge records. The diagnosis codes mark medical conditions, such as chronic ischemic heart disease, pure hypercholesterolemia, type 2 diabetes, and many other types of diseases.

Table 2 presented the top 20 diagnosis codes that co-existed with opioid poisoning diagnoses by their total counts in all opioid poisoning discharge records.

We found several diagnoses that were the well-established comorbidities of opioid dependence. For example, opioid use disorder is often associated with other substance use disorders, such as tobacco, alcohol and benzodiazepines, which are often taken to reduce symptoms of opioid withdrawal or craving for opioids, or to enhance the effects of administered opioids. Individuals with opioid use disorder are also at risk for the development of mild to moderate depression. Periods of depression are especially common during chronic intoxication or in association with physical or psychosocial stressors that are related to the opioid use disorder.
As we didn’t find any infection disease among the top co-existed diagnoses in Table 2, most of the opioid poisoning hospital visits may come from patients with prescription opioids. Usually, the most common medical conditions associated with opioid use disorder are viral (e.g., HIV, hepatitis C virus) and bacterial infections, particularly among users of opioids by injection. These infections are less common in opioid use disorder with prescription opioids.\textsuperscript{18}

The other co-existed diagnoses listed in Table 2 should be the effects from opioid poisoning, including acute respiratory failure, acute kidney failure, pneumonitis, hypopotassemia, and rhabdomyolysis.

**Table 2.** Top 20 Co-existed Diagnosis Codes of Opioid Poisoning Hospital Visits.

<table>
<thead>
<tr>
<th>Code</th>
<th>Description</th>
<th>Rank</th>
<th>Code</th>
<th>Description</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>3051</td>
<td>Tobacco Use Disorder</td>
<td>1</td>
<td>5849</td>
<td>Acute Kidney Failure, Unspecified</td>
<td>11</td>
</tr>
<tr>
<td>4019</td>
<td>Unspecified Essential Hypertension</td>
<td>2</td>
<td>25000</td>
<td>Diabetes Mellitus Without Mention Of Complication</td>
<td>12</td>
</tr>
<tr>
<td>311</td>
<td>Depressive Disorder, Not Elsewhere Classified</td>
<td>3</td>
<td>30000</td>
<td>Anxiety State, Unspecified</td>
<td>13</td>
</tr>
<tr>
<td>30550</td>
<td>Opioid Abuse, Unspecified</td>
<td>4</td>
<td>5070</td>
<td>Pneumonitis Due To Inhalation Of Food Or Vomitus</td>
<td>14</td>
</tr>
<tr>
<td>51881</td>
<td>Acute Respiratory Failure</td>
<td>5</td>
<td>30500</td>
<td>Alcohol Abuse, Unspecified</td>
<td>15</td>
</tr>
<tr>
<td>30401</td>
<td>Opioid Type Dependence, Continuous</td>
<td>6</td>
<td>30590</td>
<td>Other, Mixed, Or Unspecified Drug Abuse, Unspecified</td>
<td>16</td>
</tr>
<tr>
<td>9694</td>
<td>Poisoning By Benzodiazepine-Based Tranquilizers</td>
<td>7</td>
<td>2768</td>
<td>Hypopotassemia</td>
<td>17</td>
</tr>
<tr>
<td>78097</td>
<td>Altered Mental Status</td>
<td>8</td>
<td>33829</td>
<td>Other Chronic Pain</td>
<td>18</td>
</tr>
<tr>
<td>78009</td>
<td>Other Alteration Of Consciousness</td>
<td>9</td>
<td>49390</td>
<td>Asthma, Unspecified Type, Unspecified</td>
<td>19</td>
</tr>
<tr>
<td>30400</td>
<td>Opioid Type Dependence, Unspecified</td>
<td>10</td>
<td>72888</td>
<td>Rhabdomyolysis</td>
<td>20</td>
</tr>
</tbody>
</table>

**Table 3.** Top Co-existed diagnosis codes for different types of opiates and related narcotics.

<table>
<thead>
<tr>
<th>Opioid poisoning Diagnosis Code and Description</th>
<th>Code</th>
<th>Description</th>
<th>Co-existed Diagnosis</th>
<th>Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>96500 Poisoning; Opium (alkaloids)</td>
<td>9670</td>
<td>Poisoning By Barbiturates</td>
<td>E8501 Accidental Poisoning; Methadone</td>
<td>0.002</td>
</tr>
<tr>
<td></td>
<td>9679</td>
<td>Poisoning By Unspecified Sedative Or Hypnotic</td>
<td>33829 Other Chronic Pain</td>
<td>0.027</td>
</tr>
<tr>
<td>96501 Poisoning; Heroin</td>
<td>30550</td>
<td>Opioid Abuse, Unspecified</td>
<td>7245 Backache, Unspecified</td>
<td>0.019</td>
</tr>
<tr>
<td></td>
<td>30551</td>
<td>Opioid Abuse, Continuous</td>
<td>7242 Lumbago</td>
<td>0.013</td>
</tr>
<tr>
<td></td>
<td>97081</td>
<td>Poisoning By Cocaine</td>
<td>2720 Pure Hypercholesterolemia</td>
<td>0.09</td>
</tr>
<tr>
<td></td>
<td>9708</td>
<td>Poisoning by other specified central nervous system stimulants</td>
<td></td>
<td></td>
</tr>
<tr>
<td>96502 Poisoning; Methadone</td>
<td>4275</td>
<td>Cardiac Arrest</td>
<td>9654 Poisoning By Aromatic Analgesics</td>
<td>0.013</td>
</tr>
<tr>
<td>96509 Poisoning; Other opiates and related narcotics</td>
<td>7270</td>
<td>Pure Hypercholesterolemia</td>
<td>7242 Lumbago</td>
<td>0.013</td>
</tr>
</tbody>
</table>
Table 3 presented the top co-existed diagnoses for poisoning by different substance. Poisoning by heroin should be mainly caused by drug abuse as it generally co-existed with opioid abuse and resulted in a life-threatening cardiac arrest. Poisoning by opium more likely co-existed with poisoning by barbiturates or other sedative-hypnotic drugs. On the other hand, poisoning by methadone mainly co-existed with accidental poisoning by methadone. Such findings may require further research for the potential driving factors.

**Spatial Analysis**

In this section, we performed zip code level spatial analysis to detect local spatial clusters, and studied potential correlations between opioid poisoning and demographic and social-economic factors.

Figure 4 demonstrated the results of spatial clusters and outliers. The cluster/outlier type field distinguishes between a statistically significant cluster of high values (High-High cluster), cluster of low values (Low-Low cluster), outlier in which a high value is surrounded by low values (High-Low outlier), and outlier in which a low value is surrounded by high values (Low-High outlier). Statistical significance is set at the 95 percent confidence level. We applied the False Discovery Rate (FDR) correction to reduce this p-value threshold from 0.05 to a value that better reflects the 95 percent confidence level given multiple testing. The FDR procedure will potentially reduce the critical p-value to account for multiple testing and spatial dependency. 13

![Clusters and Outliers](image)

**Figure 4.** Spatial clusters and outliers of opioid poisoning incidence rate by ZIP code, New York State, 2010-2014.

We found that most areas of New York City had lower opioid poisoning hospital visits (Low-Low clusters in blue color) except several scattered high risk areas. Such findings were consistent with the decreasing trend of opioid overdose19 in New York City and the steep rise in opioid hospital visits outside New York City20. Targeted public health interventions might be effective in lowering opioid overdose mortality rates19. However, a small part of Staten Island and most part of southeastern Long Island were identified as high risk areas (High-High clusters in red color).
For the scattered outliers across New York City (High-Low outlier or Low-High outlier), further investigation and intervention efforts may be required in the future.

The spatial clustering analysis tried to answer the question of “where” by identifying potential high or low risk areas. The following question should be finding out why there were higher or lower opioid poisoning hospital visits at the discovered spatial clusters or outliers. We then used spatial regression analysis by linking demographic and socio-economic factors to opioid poisoning.

We did exploratory regression analyses between the opioid poisoning and exemplary demographic and socio-economic factors as shown in Table 4. Among the 13 spatial impact factors, the OLS regression results showed that household income, Asian population and the youth population age 20 to 24 had most significant negative correlation relationship with the opioid poisoning.

We then used GWR to model the local trends of correlation relationship between household income and opioid poisoning rates. The local coefficient map in Figure 5 showed that urban residents who made more money had lower rates of opioid poisoning, while country residents with higher income tended to had higher rates of opioid poisoning.

**Table 4.** Statistics of demographic and socio-economic factors by ZIP Code in New York State, 2010-2014.

<table>
<thead>
<tr>
<th>Spatial Impact Factors</th>
<th>Mean (Std. Error)</th>
<th>Moran’s I Index</th>
</tr>
</thead>
<tbody>
<tr>
<td>Opioid Poisoning Incidence Rate per 1,000</td>
<td>1.68 (2.02)</td>
<td>0.09</td>
</tr>
<tr>
<td>% Male</td>
<td>49.61 (5.41)</td>
<td>0.07</td>
</tr>
<tr>
<td>% Age 15-19 years</td>
<td>6.73 (4.32)</td>
<td>0.07</td>
</tr>
<tr>
<td>% Age 20-24 years</td>
<td>6.47 (4.06)</td>
<td>0.14</td>
</tr>
<tr>
<td>% Age 25-34 years</td>
<td>11.71 (5.39)</td>
<td>0.44</td>
</tr>
<tr>
<td>% Age 35-44 years</td>
<td>12.27 (3.74)</td>
<td>0.13</td>
</tr>
<tr>
<td>% White</td>
<td>78.25 (25.22)</td>
<td>0.73</td>
</tr>
<tr>
<td>% Black</td>
<td>6.89 (14.14)</td>
<td>0.63</td>
</tr>
<tr>
<td>% Asian</td>
<td>3.69 (7.01)</td>
<td>0.72</td>
</tr>
<tr>
<td>% Hispanic</td>
<td>8.90 (12.63)</td>
<td>0.69</td>
</tr>
<tr>
<td>Socio-economic Factors</td>
<td>Household Income ($10,000 US)</td>
<td>76,934 (35,610)</td>
</tr>
<tr>
<td>% Poverty</td>
<td>12.5 (10.01)</td>
<td>0.43</td>
</tr>
<tr>
<td>% Uninsured</td>
<td>9.15 (5.59)</td>
<td>0.30</td>
</tr>
</tbody>
</table>

**Discussion**

There has long been a demand for large-scale data driven approach to discover both spatial and temporal patterns and identity potential causes of diseases. This study provided our preliminary results of a large-scale patient level study on opioid poisoning hospital visits in the New York State. We examined the demographic disparities for the patients with opioid poisoning. We compared the historical trends of opioid poisoning for hospital emergency departments visits and inpatient stays. We used frequent item mining to find co-occurrences of diagnoses for possible causes of poisoning or effects from poisoning. We performed zip code level spatial analysis to detect local spatial clusters, and studied potential correlations between opioid poisoning and demographic and social-economic factors.

This work focuses on a large-scale study of opioid poisoning related hospital visits in New York State. We found that men were more likely to have opioid poisoning hospital visits than women. Whites made 72.2% of patients that was higher than the whites proportion in general population. Asians, however, had a very low percentage of patients. We also found a peak of opioid poisoning hospital visits for young adults (21 to 25, peak at 22). A second peak is for ages
starting from middle age (46-55, peak at 51). We found several diagnoses that were well-established comorbidities of opioid dependence. The spatial clustering analysis showed that most areas of New York City had lower opioid poisoning hospital visit rates compared to that in areas outside New York City. The spatial regression results showed that urban residents who made more money had lower rates of opioid poisoning, while country residents with higher income tended to have higher rates of opioid poisoning.

However, our results were based on the patients with their primary diagnosis as opioid poisoning. In our ongoing work, we will also include the patients with their secondary diagnosis as opioid poisoning. We will examine more types of disease diagnoses and treatments and provide multi-dimensional analysis by grouping patients per their demographic attributes. We will take advantage of the street level location information and the full history of each patient for more fine grained spatial and temporal analysis. For example, after geocoding the patient addresses into latitude and longitude coordinates, we will perform point based spatial clustering analysis. By mapping patients’ residential locations into census block group boundaries, we will link the health records with census data and study potential risk factors with fine-grained spatial resolutions.

**Figure 5.** The choropleth map that visualizes Geographically Weighted Regression (GWR) local coefficient of household Income for opioid poisoning incidence rate by ZIP code in New York State, 2010-2014.

**Conclusion**

Increased accessibility of health data made available by the government provides unique opportunity for data driven discovery of disease patterns and identify potential causes of opioid related deaths. Large-scale patient level analysis could provide new insights and create new forms of value to support governments on decision making. In this paper, we present our analytical results on opioid poisoning related hospital visits using New York State SPARCS data. Our results not only provide quantitative measurements on demographic-based distributions, and spatial and temporal trends, but also present top common co-occurrences of diagnoses with opioid poisoning, which provides a groundwork to study possible causes of poisoning and effects from poisoning. Our results provide essential knowledge and guidelines to support stakeholders for improving prevention, intervention and recovery of opioid poisoning.

**Acknowledgments**

This work is supported in part by NSF ACI 1443054 and by NSF IIS 1350885.
References

18. EDITION F. Diagnostic and statistical manual of mental disorders.
Determining Burden of Commuting for Treatment Using Online Mapping Services – A Study of Breast Cancer Patients

Alex C. Cheng, MS MEM, Mia A. Levy, MD PHD
1Vanderbilt University, Nashville, TN

Abstract

For patients with breast cancer who must frequent medical centers for care, commuting is a significant burden. This burden could affect their decisions during treatment. We developed a method to use census tracts and zip codes to determine commuting burden for patients with breast cancer with online mapping services, while protecting patient addresses from third parties. We found that patients who lived farther from Vanderbilt had fewer unique appointment days and more appointments scheduled per day. Total burden decreased over time after diagnosis, but advanced stage patients had sustained high levels of commute time until ten months after diagnosis. Additionally, we found that patients who lived far from Vanderbilt were less likely to receive radiotherapy from Vanderbilt. With the amount of work patients put into traveling for care, understanding commuting burden could help healthcare organizations form strategies to improve access to care and compliance with care plans.

Introduction

Traveling to and from a medical center for treatment is a significant burden to many patients with chronic conditions. In 2014, The Center for Disease Control and Prevention showed that 67% of adults in the United States had at least one encounter with the healthcare system within 6 months of the survey. The percentage of patients who saw a healthcare provider increases significantly for patients with chronic conditions. In 2013, 99% of patients with hypertension had an office-based physician visit and 47% had four or more visits. Similarly, 55% of patients with diabetes visited a physician four or more times in a year. Elderly patients, who often have difficulty traveling for care, had to travel more frequently than the average patient. In addition to the sheer number of times patients must travel to medical centers for care, patients also perceive commuting as a burden. In a survey of 1053 patients regarding factors that contributed to their treatment burden, 41% expressed that they had difficulty adapting to new routines for care that involved planning and organizing travel. Additionally, 30% of patients surveyed indicated they had difficulty with access to health care centers citing distance and parking as barriers to receiving care. Interest in treatment burden goes beyond just providing convenient care for patients. Patients who receive care within their means and are not overburdened tend to be more compliant with their treatment plans which could lead to better outcomes.

Breast cancer patients experience a high level of treatment burden. In our prior work, we demonstrated that stage I-III breast cancer patients receiving care at Vanderbilt underwent an average of 59 appointments over the course of 18 months after their diagnosis. During this time, these patients had to travel to the medical center an average of 39 times and spent approximately 49 hours in clinic. Stage III patients experienced the most time in clinic, followed by stage II and stage I patients. One reason for the intensity of treatment burden in breast cancer patients is the complexity of their treatment. Encounters included radiology diagnostics, laboratory tests, surgery, radiation therapy, and chemotherapy. Furthermore, many patients experienced additional treatment burden due to complications to their care that led to hospitalizations or the need for physical therapy.

Traveling for care is a challenge for cancer patients, particularly those living far from metropolitan areas. Travel also contributes to the burden of treatment through transportation costs. An Australian study showed that the median cancer patient spent 956 Australian dollars (about 727 US dollars) in travel costs over 16 months after diagnosis, which accounted for 71% of all out-of-pocket costs. Distance traveled could even affect patient treatment choices. One study determined that driving distance from a radiotherapy facility resulted in more patients with breast cancer choosing mastectomy instead of breast conserving surgery. While our previous studies looked at burden of treatment due to time spent in inpatient and outpatient encounters, we did not factor in the work patients put into traveling to the medical center for those encounters. The goal of this study was to use web services to calculate commuting burden over the course of treatment for patients with breast cancer.
In addition to the duration of the commute to the medical center, the mode of transportation could also be a factor that healthcare providers should consider. Researchers demonstrated United States counties and English districts where more households had access to a car had a higher rate of screening for cervical cancer. Conversely, breast cancer screening was lower in English districts with higher public transportation usage. While public transportation may be less convenient for patients receiving care for cancer, it may be some patients’ only option. While we assume that most patients who receive care at Vanderbilt arrive by car, we will explore the possibility of commuting by public transportation for our population of breast cancer patients.

Healthcare researchers have used mapping web services to improve the delivery of care. One group from the Netherlands used Google Maps to calculate the difference between driving time and helicopter flight time to help paramedics decide the most effective way to transport patients to the hospital. Services such as Google Maps are excellent at keeping up with changing traffic patterns and new roads that may affect commute times both for driving and public transportation. However, one challenge with using online services in healthcare delivery and research is that sending patient addresses to companies without a Business Associate Agreement (BAA) is a violation of the Health Information Portability and Accountability Act (HIPAA) privacy rules. Additionally, the American Journal of Public Health released an editorial stating that sending patient addresses to a third party is inappropriate and that some method of geographic imputation should be used to protect patient privacy. In our calculations of work due to travel burden, we propose a method of geographic imputation using zip codes, census blocks, and bus stops as landmarks to protect personally identifiable information (PII).

Methods

Validating the Landmarks Method

Our landmarks method is similar to aggregation techniques in public health to anonymize locations. First, we obtained all zip codes and census tract centroid coordinates publicly available on the US Census Bureau website. Nashville Metropolitan Transit Authority (MTA) granted us access to their application program interface (API) where we could pull the list of bus stops and their coordinates. We queried the Google Maps API (Gmaps) for driving times from every landmark to the Vanderbilt University Medical Center (VUMC) and back. Next, we used the Data Science Toolkit (DSTK) geocoder to determine the latitude-longitude coordinates for each of our patient addresses. The DSTK also returns a confidence level for how sure it is that the address is geocoded correctly. We installed the DSTK on virtual machine that ran on a local Vanderbilt computer, thus eliminating the need to send patient addresses to a third party.

To test this method, and to compare the accuracy of the various landmark sets (zip code, census tract, and bus stops), we applied the procedure to a set of homes for sale in the Nashville area. We queried 500 random and publicly available addresses from the Redfin.com real estate listing website on January 30, 2017. Then we compared the coordinates given by the DSTK geocoder with those given by the Gmaps geocoder. Based on the agreement between the DSTK and Gmaps geocoders, we determined a DSTK confidence threshold below which we would not trust. We queried Gmaps for driving times for each of the 500 real estate addresses and compared those to the driving times of each addresses’ respective closest landmark.

Applying the Landmarks Method to Breast Cancer Patient Commutes

After testing the method on the 500 real estate addresses, we applied the best landmark set on a cohort of breast cancer patients obtained from the Vanderbilt Tumor Registry. We included patients with stage I-III breast cancer diagnosed from January 1, 1998 to June 1, 2014. To capture only patients who received most of their first course of treatment at Vanderbilt, we only included patients who had at least three appointments each with a medical oncologist and oncology surgeon. We compared commute times for patients who lived within 100 miles of the main VUMC campus. Patient commute time for any given appointment day was the time it took to drive a round trip from the landmark closest to their home address to the Vanderbilt facility where their appointment was held. To get a characterization of the total burden of traveling, we calculated the total amount of time patients would have to spend traveling to the medical center by car over 18 months after their date of diagnosis. We also compared the behavior of commuting in patients that were farther than the median distance from VUMC with those which were closer than the median distance. We compared the frequency that patients received radiation therapy at a Vanderbilt facility between patients who were closer and those who were farther. With the coordinates of bus stops in Nashville, we analyzed the number of patients that could have feasibly traveled to their appointments via public transportation. Finally, using average commute times, we estimated the cost of commuting per patient. Assuming an average speed
of driving in Nashville of 32.4 miles per hour obtained from Google traffic data\textsuperscript{19}, and a cost of operating a vehicle of 54 cents per mile in 2016, we extrapolated the average total cost of commuting by vehicle per patient. A more accurate method would have been to use direct driving distance based on the Gmaps recommended route. However, due to Gmaps query constraints, we inferred the driving distance using average driving speed. The cost per mile comes from the Internal Revenue Service, which sets mileage rates for the cost of operating a passenger vehicle for charitable and medical purposes\textsuperscript{20}.

Figure 1. Patient addresses from the Vanderbilt Tumor Registry within 100 miles of VUMC were included in driving time calculations.

Results

Among the 500 random real estate addresses obtained from Redfin.com, the DSTK geocoder found 495 latitude-longitude coordinates compared to the Gmaps geocoder which found 483. There was generally good agreement between the coordinates found by DSTK and those found by Gmaps. Among the 425 addresses that were found by both the DSTK and Gmaps, 418 had less than a quarter mile straight-line difference between the DSTK and Gmaps coordinates. To exclude addresses that had major disagreement between DSTK and Gmaps, we only verified landmark driving times for addresses where the DSTK geocoder had at least 80% confidence. DSTK geocoded 427 addresses with at least 80% confidence.

For the addresses found by the DSTK with greater than 80% confidence, we compared driving time to VUMC using the true address and the nearest landmark. Figure 2 shows the difference in round trip driving time as calculated by Gmaps using the true address versus using the zip code centroid, census tract centroid, and nearest bus stop coordinates. Using zip codes in place of true addresses tended to overestimate driving time when the true address was close to VUMC and tended to underestimate driving time when the address was farther away. The difference in estimation time was greater than 20 minutes in some circumstances. With census tracts and bus stops, the difference in times compared to the true addresses was generally less than 10 minutes. Differences between bus stop and true address driving times increased as the distance from VUMC increased.
Figure 2. Differences in round trip driving time to VUMC between real estate addresses and nearest zip codes, census tracts, and bus stops.

For our study of commuting burden in breast cancer patients, we used the nearest census track centroid to calculate driving times. The census tract method was more generalizable and robust for addresses far from VUMC than the bus stop method. We used zip codes for patient addresses where the geocoder had less than 80% confidence. There were 768 patient addresses within 100 miles of the main VUMC campus. Among those, we used the nearest census tract to calculate commute time for 644 patients, and used zip code for 124 patients.

Figure 3. Cohort selection

The distribution of patients within 100 miles by stage was similar to the overall distribution for all stage I-III patients. There were 374 stage I patients, 273 stage II patients, and 121 stage III patients within 100 miles of the main VUMC campus. Among these patients, there was not much differentiation between stages in the distribution of a single round trip driving time from VUMC. The median driving time to and from VUMC across all stages was 76 minutes and the median straight-line distance from VUMC was 20 miles.
Figures 4 and 5. Number of patients per stage included in analysis and distribution of round trip driving times from patient addresses to main VUMC campus.

Figure 6 shows that overall burden, consisting of the sum of time in appointments, waiting time between appointments, and driving time, decreased for patients of all stages over the course of 18 months after diagnosis. Stage I and II patients saw peaks in overall burden in the first and fourth months. Overall time spent on encounters reached about 14 hours per month for stage I patients and 15 hours for stage II patients. Stage III patients had more sustained burden through the first eight months of treatment with a peak of seven hours of commute time in month eight.

Figure 6. Hours spent in appointments, waiting, and commuting over 18 months after diagnosis by breast cancer stage.

In Figure 7, there is clear differentiation in the number of unique appointment days over 18 months between patients who lived closer and farther than the median distance from VUMC. Patients who lived farther made fewer trips to a Vanderbilt facility compared to their closer counterparts in all three stages. Figure 8 shows that patients farther away also had more appointments per trip to a Vanderbilt facility across all stages.
Figure 7. Distribution of unique appointment days over 18 months for patients closer (within 20 miles) and farther (greater than 20 but less than 100 miles) from VUMC by stage.

Figure 8. Distribution of mean number appointment per unique appointment day per patient by stage and distance from VUMC.

The percentage of patients who received radiation therapy at Vanderbilt could be indicative of how commute time affected where patients decided to receive care. Table 1 shows the percentage of patients who received radiotherapy at a Vanderbilt facility compared to all patients in that group. Assuming patients within each stage required radiation therapy at approximately the same rate, stage I and stage II patients who lived closer to VUMC received radiation therapy at Vanderbilt at a rate about three times higher than those who lived farther. Stage III patients close to VUMC received radiation therapy at Vanderbilt at a rate five times higher.

Table 1. Percentage of patients who received radiotherapy at a Vanderbilt facility by stage and distance from VUMC.

<table>
<thead>
<tr>
<th></th>
<th>Close (within 20 miles)</th>
<th>Far (&gt;20 and &lt;100 miles)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stage I</td>
<td>46%</td>
<td>17%</td>
</tr>
<tr>
<td>Stage II</td>
<td>44%</td>
<td>14%</td>
</tr>
<tr>
<td>Stage III</td>
<td>78%</td>
<td>15%</td>
</tr>
</tbody>
</table>

For our cohort of patients, 97.5% of appointments take place at facilities accessible by public transportation. However, Table 2 shows how many patients could access public transportation at various walking tolerances, and the percentage of accessible appointments accounted for by those patients.
Table 2. Number of patients within walking distance of a bus stop at various walking tolerances.

<table>
<thead>
<tr>
<th>Patient address distance from bus stop</th>
<th>Patients</th>
<th>% of all appointments accessible</th>
</tr>
</thead>
<tbody>
<tr>
<td>100 yards</td>
<td>25</td>
<td>4.52%</td>
</tr>
<tr>
<td>200 yards</td>
<td>49</td>
<td>8.26%</td>
</tr>
<tr>
<td>.25 miles</td>
<td>124</td>
<td>20.2%</td>
</tr>
<tr>
<td>.5 miles</td>
<td>171</td>
<td>27.4%</td>
</tr>
<tr>
<td>1 mile</td>
<td>223</td>
<td>36.2%</td>
</tr>
<tr>
<td>2 miles</td>
<td>263</td>
<td>42.2%</td>
</tr>
</tbody>
</table>

Finally, we performed a cost analysis based on our calculated driving times. As expected, patients closer and with lower stage had less estimated cost of commuting by motor vehicle.

Table 3. Estimated cost of vehicle expenses per patient over 18 months after diagnosis by stage and distance from VUMC. Mean (range).

<table>
<thead>
<tr>
<th></th>
<th>Close (within 20 miles)</th>
<th>Far (&gt;20 and &lt;100 miles)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stage I</td>
<td>$609 ($51.06 - $954.76)</td>
<td>$1047 ($238.02 - $3187.27)</td>
</tr>
<tr>
<td>Stage II</td>
<td>$824 ($77.25 - $3125.06)</td>
<td>$1455 ($198.29 - $4958.17)</td>
</tr>
<tr>
<td>Stage III</td>
<td>$1050 ($198.29 - $4958.17)</td>
<td>$1625 ($294.48 - $4779.91)</td>
</tr>
</tbody>
</table>

Discussion

Through our attempt to calculate the burden of treatment related to commuting for patients with breast cancer, we succeeded in developing a method for calculating driving times using online mapping services that did not compromise patient PII. When deciding what type of landmark to use in our method, we decided to use a mixture of census block and zip code centroids instead of bus stop coordinates. While Figure 2 shows that bus stops are somewhat more accurate than census blocks for the real estate addresses, the accuracy gets worse the farther the address is from the city center. This effect may be because bus stops fall along major roads which become farther apart in suburban and rural areas. Additionally, since not all cities have a public transportation system, using census tracts and zip codes makes our method more generalizable.

Geomasking methods such as random perturbation or donut masking attempt to hide patient addresses by randomly moving the patient address in a radius around the true address\(^{16}\). These methods of protecting PII have the potential to be more accurate than our landmarks method since the size of the radius is dependent on the population density around the patient’s location. We did not consider using one of these methods since they each require the researcher to define a level of k-anonymity, which is the minimum number of people from which any research subject could be re-identified from\(^{21}\). It is difficult to define such a k-anonymity level that would be necessary to protect patients from an internet service. However unlikely, Google could easily target cancer treatment relevant advertisements to hundreds of people for the 1/k chance that the cancer patient in the area would receive the advertisement. Additionally, geomasking may be more useful in public health studies where spatial precision is necessary to identify sources of outbreaks\(^{22}\), but is less essential for calculating estimated driving times. As demonstrated from the comparison of commute times between true real estate addresses and census blocks, there is only a small effect on the overall commute time.

One major constraint in using online mapping services such as Google Maps for calculating driving distances is that there is a limit on the number of free requests per user. In early 2017 when we performed this study, Google Maps allowed users to make 2,500 free requests per day, with requests over that quota costing $0.50 per 1000 queries\(^{23}\). In order to maintain full de-identification for PII, we had to request driving times from every census tract and zip code centroid to every Vanderbilt location. The constraint of request quotas limited the scope of this project in several ways. Including only patients within 100 miles of Vanderbilt is reasonable since patients who live farther away may not commute daily from home. However, anecdotally, we have seen that patients who live as far as 200 miles away are driving from their homes to Vanderbilt for care. Our method included the 829 census tracts within 100 miles of...
VUMC, which could be queried for one Vanderbilt location in one day. This number jumps to 3639 census tracts within 200 miles which would require three days of free queries per Vanderbilt location (driving times to and from a location count as two requests). As we build a database of driving times, future work will include patient addresses that are farther away.

Another interesting question that we could answer with more Google Maps queries is the effect of traffic on patient commute times. One of the reasons we chose not to use open source projects such as the Open Source Routing Machine (OSRM) is because they do not have the means to collect live traffic information. Geographic Information System (GIS) software such as ArcGIS has a live traffic feed available, but only through a paid subscription. One powerful feature of modern web mapping applications is that they track typical traffic patterns to provide driving time predictions that factor in road congestion. However, ten of the census tracts within 100 miles of VUMC had fewer than 20 people living in them according to the 2010 census. The presence of low population census tracts means we would have to query every census tract for every appointment time in order to achieve full anonymity, which would become expensive to do with Google Maps. Future work could gradually save hourly driving times with traffic data to get an idea for how much traffic affects the work patients put into their care. Alternatively, we may establish a BAA with Google or another company that provides live driving time predictions.

Despite the limitations, we made several observations about the effect of commute time on cancer treatment. Aside from confirming that stage III patients experienced a higher treatment burden than stage II and stage I patients, we observed in Figure 6 that the pattern in commute time over months after diagnosis was different for stage III patients. Stage III patients experienced increased commute times in months six through eight after diagnosis despite a decrease in appointment time during that period. This increased commute time, coupled with the decrease in appointment and waiting time, may be associated with the observation that many stage III patients underwent radiation therapy after surgery. Radiation therapy procedures are typically 15 minute appointments that occur daily in rapid succession. The fact that these encounters are short but still require patients to travel to the medical center could explain the increase in commute time relative to appointment time.

Table 1 showed that patients farther from Vanderbilt received radiation therapy less often at Vanderbilt than their counterparts that lived closer. The rate that patients received radiation therapy may be high overall due to our cohort already being biased toward patients who received a majority of their care at Vanderbilt with the constraint that all patients have at least three appointments with a medical oncologist and oncology surgeon. If we assume that patients of a given stage of breast cancer require radiation therapy at approximately the same rate, then we can conclude that more patients who live farther from Vanderbilt are getting radiation therapy at other institutions. This finding supports the conclusion of Goyal et. al. that driving distance to a radiation therapy center influences breast cancer patients’ treatment path decisions. This type of information would be useful to healthcare organizations that are considering opening new radiation therapy clinics. If a new clinic knew that patients are three to five times more likely to choose to receive radiation therapy at Vanderbilt with a more convenient location, the clinic could plan capacity to meet that demand.

In discussing patient experience for commuting, one important consideration is determining patient capacity to handle a long trip to the doctor. In cities such as Nashville where a typical commute to work was more than 30 minutes in 2014, medical centers may be able to expect patients travel farther for care. However, in a city where traffic is less onerous, patients may be more sensitive to commuting long distances to a medical center regularly. Nevertheless, our calculated round trip to VUMC for the median patient was 76 minutes. This result means that even without traffic, the median patient within 100 miles of VUMC would have to drive longer than the average work commuter during rush hour.

Figures 7 and 8 have implications for care coordination in patients with cancer. The fact that patients within each stage who were farther from VUMC had fewer unique appointment days and more days per appointment suggest that some effort is being made to coordinate appointments to occur on the same day for patients who live farther away. While it may be hard to determine whether the patient or medical center staff is putting in the coordination effort, being able to track outpatient appointment coordination allows organizations to identify areas for improvement. It might be prudent for patient care coordinators or navigators to examine upward outliers in Figure 8 to see what strategies are working for patients who average more than three appointments per visit.
There are several assumptions we made in our study. First, we assumed that patient addresses in the tumor registry were accurate at the time of their diagnosis, and that patients did not move during the first 18 months of treatment. We also assumed that patients traveled from their home address each unique appointment day. It is possible that patients stayed in hotels or with relatives during the more intense parts of their treatment, which would cut down on burden related to commuting. It is also possible that the patient traveled to VUMC from their work address. There were some locations listed in the appointment record that were not primary Vanderbilt locations and thus, we did not have driving time data for them. These appointments were excluded from our analysis. Only one patient had more than 7 appointments at a non-Vanderbilt listed location. That patient still received 89% of their appointments at a Vanderbilt facility and so the influence of this outlier should negligible.

With regards to public transportation in patient commuting, the main takeaway from Table 2 is that only a small proportion of Vanderbilt’s breast cancer population would be able to take advantage of public transportation. Even if patients were willing to walk two miles to their nearest bus stop, only 263 patients would have access to public transportation. Future work might consider what is the maximum reasonable distance to expect patients with different conditions such as cancer to walk before and after their appointments. Additionally, it would be interesting to see whether there is improved access to healthcare facilities via public transportation in more densely populated cities.

Healthcare organizations could also use this method to predict patient commute times on the day of patients’ appointment. These predictions can be used to warn patients who may need to leave their homes earlier in order to avoid traffic, or to anticipate which patients may be late due to abnormal traffic conditions. Informaticians can also use calculations of commuting burden to develop tools that benefit patients. Providing patients with a mobile application to automatically calculate travel time to appointments would require consent to track their locations. However, such an application could help to alert patients of when they should leave their homes to arrive at their appointments on time. With real-time traffic conditions integrated with the appointment record, an online navigation service could recommend a driving route that avoids traffic and minimizes commuting burden.

Finally, being able to track work related to driving could also allow organizations to identify patients who may be overburdened. For example, patients who are high outliers for overall burden from appointments and procedures may benefit from a home visit from a nurse in lieu of outpatient appointment. In addition to the time requirements of cancer care, financial costs for cancer patients may lead to extreme financial distress and worse outcomes, a phenomenon known as financial toxicity. Foundations such as Susan G. Komen provide support to breast cancer patients who have difficulty affording their care. One of the programs provides financial relief to qualified breast cancer patients by giving them gas card vouchers. A healthcare organization could use information from Table 3 to request a grant from the Komen Foundation for patients under their care based on stage of cancer and travel distance.

**Conclusion**

We developed a generalizable method to calculate approximate driving times from patient addresses to VUMC locations using a third party online mapping service without sending PII to that third party. We used this method to determine the burden of treatment related to commuting for patients with breast cancer receiving care at VUMC. We found that radiation therapy made a significant impact on commuting burden due to the frequency of treatment. Also, patient’s living farther from VUMC tended to receive radiation therapy more at other medical facilities compared to those living closer to VUMC. We discovered that patients farther from VUMC had more appointments per unique appointment day, showing that their care was better coordinated. Future applications for travel time computation could equip organizations to better address the needs of their patients and help patients reduce the disruption of treatment on their lives.

**Acknowledgments**

Alex Cheng was supported by the T15LM007450-13 training grant from the National Library of Medicine.

**References**

Predicting Inpatient Acute Kidney Injury over Different Time Horizons: How Early and Accurate?

Peng Cheng, PhD1,3, Lemuel R. Waitman, PhD1, Yong Hu, PhD2, Mei Liu, PhD1
1University of Kansas Medical Center, Department of Internal Medicine, Division of Medical Informatics, Kansas City, KS, USA; 2Jinan University, Big Data Decision Institute, Guangzhou, China; 3Southwest University, School of Computer & Information Science, Chongqing, China

ABSTRACT
Incidence of Acute Kidney Injury (AKI) has increased dramatically over the past two decades due to rising prevalence of comorbidities and broadening repertoire of nephrotoxic medications. Hospitalized patients with AKI are at higher risk for complications and mortality, thus early recognition of AKI is crucial. Building AKI prediction models based on electronic medical records (EMRs) can enable early recognition of high-risk patients, facilitate prevention of iatrogenically induced AKI events, and improve patient outcomes. This study builds machine learning models to predict hospital-acquired AKI over different time horizons using EMR data. The study objectives are to assess (1) whether early AKI prediction is possible; (2) whether information prior to admission improves prediction; (3) what type of risk factors affect AKI prediction the most. Evaluation results showed a good cross-validated AUC of 0.765 for predicting AKI events 1-day prior and adding data prior to admission did not improve model performance.

INTRODUCTION
Acute Kidney Injury (AKI) is a common clinical event among hospitalized patients, affecting 10% to 15% of all hospitalized patients and >50% of patients in intensive care units (ICUs)1-3. AKI is most easily detected on the basis of the acute sustained rise of serum creatinine (SCr). Using a large consecutive sample of 19,982 adults, Chertow et al. found that an increase in SCr of ≥ 0.5 mg/dl is associated with a 6.5-fold increase in the odds of death, a 3.5-day increase in in length of stay, and nearly $7,500 in excess hospital cost4. Over the past two decades, the incidence of AKI has increased significantly in North American and Europe, particularly within the United States, because of the rising prevalence of acute and chronic conditions, such as sepsis, heart failure, and diabetes5-9. Moreover, the role of medications in the changing epidemiology of AKI has increased from 7% to 16% over a 17-year span because of the wide availability of potential nephrotoxic drugs2, 10.

Unfortunately, the current care of patients with AKI is suboptimal characterized by numerous deficiencies and systematic failings that may be avoidable11. In a 2009 review of the care of patients who died in hospital with a primary diagnosis of AKI, an unacceptable delay in the recognition occurred in 43% of the patients and 20% of the cases could have been prevented with early detection12. It is however difficult for clinicians to recognize at-risk patients prior to their AKI episodes. Once an AKI episode occurs, there is no treatment to mitigate or cure AKI13-16. According to the International Society of Nephrology17, recognizing patients at risk of developing AKI and managing these patients according to their susceptibilities and exposures is likely to result in better outcomes than merely treating the established AKI. Therefore, the ability to predict AKI in hospitalized patients and monitor them at an early stage is crucial to AKI prevention.

Safety tools based on electronic medical records (EMRs) for in-hospital AKI surveillance covering kidney injury triggers have been developed; examples include the Global Trigger Tool18, a tool that uses the Acute Kidney Injury Network (AKIN) definition to monitor AKI19, and dosing tools for improving compliance with renal-dosing of medications20, 21. EMR-based AKI monitoring can expedite interventions and lead to a high percentage of patients retaining their baseline kidney function22, 23. However, as the major limitation of these tools, physicians can only react after observing signs of damage. By contrast, risk prediction would recognize high-risk patients for tailored early management. Most existing AKI risk prediction models focus on predicting adverse outcomes following AKI24 or predicting AKI after specific surgeries and interventions1, 3, 25, 26. There exists some predictive modeling work performed on the critical care populations27-29, but much less work on the general inpatient population30. Matheny et al31 proposed the first study to predict general in-hospital AKI using logistic regression models. Recently, Kate et al32 built machine learning models to examine the difference in prediction vs. detection of AKI in hospitalized older adults.
This study aims to build machine learning models to predict the development of AKI among general patient hospitalizations at daily intervals prior to the event. The primary objective is to assess how early and accurately the development of inpatient AKI can be predicted. The secondary objective is to assess whether clinical data prior to admission enhance the predictive models.

METHODS

Data Collection

A retrospective cohort of 60,534 patients was collected by including adult admissions (age at visit between 18 and 64) to a tertiary care, academic hospital (the University of Kansas Medical Center – KUMC) from November 2007 to March 2016 with a length of stay of at least 2 days. Given that a patient may have multiple admissions (encounters) of at least 2 days and develop AKI during one but not another, this study is conducted at the encounter level with the initial cohort of total 109,319 encounters. For each encounter, we queried the KUMC de-identified clinical data repository HERON (Health Enterprise Repository for Ontological Narration) that integrated electronic health records, billing, clinical registries, and national data sources to obtain structured data on admission and discharge dates, patient demographics, medications, laboratory values, vitals, comorbidities and admission diagnosis.

Cohort Inclusion/Exclusion Criteria

From the initial cohort of 109,319 encounters, we selected an analysis cohort of encounters by excluding those (a) missing necessary data for outcome determination – less than two serum creatinine measurements and (b) had evidence of moderate or severe kidney dysfunction – estimated Glomerular Filtration Rate (eGFR) less than 60 mL/min/1.73 m² or abnormal serum creatinine (SCr) level of >1.3 mg/dL within 24 hours of hospital admission. The final analysis cohort consisted of 48,955 encounters (33,703 patients).

AKI Definition

AKI was defined using the Kidney Disease Improving Global Outcomes (KDIGO)-based modifications of the AKIN and Risk, Injury, Failure, Loss, and End-Stage (RIFLE) Kidney classification criteria. According to KDIGO, adults who demonstrate any of the following are undergoing an AKI episode:

- Increase in SCr by ≥ 0.3 mg/dL (≥ 26.4 micromol/L) within 48 hours
- Increase in SCr by ≥ 1.5 times the baseline within the previous 7 days

The baseline creatinine level was set as either the last measurement within 2-day time window prior to admission or the first available measurement during the stay. All creatinine measurements between admission and discharge were evaluated to determine the occurrence of in-hospital AKI. Based on the above AKI definition, this study classifies each encounter as ‘with AKI’ (positive) or ‘without AKI’ (negative). Out of total 48,955 encounters in the final analysis cohort, patients acquired AKI during 4,405 (8.99%) encounters.

AKI Risk Factors

A list of clinical variables used in building the AKI prediction models is described in Table 1. We referred to Matheny et al. to select laboratory tests that may represent potential presence of a comorbidity that is correlated with in-hospital AKI. For example, an elevated white blood cell count (WBC) is associated with bacterial infection that may cause AKI. Serum creatinine was not included as a variable as it was used to determine the positive and negative samples. For laboratory tests and vitals, only the last recorded value before a prediction point was used and their values were categorized. Laboratory values were categorized as either “present and normal”, “present and abnormal”, or “unknown” according to standard reference ranges. Vitals were categorized into groups as shown in Table 2. Missing values in vitals and lab tests were captured as “unknowns” because information may be contained in the choice to not perform the measurement.

Medication variable included inpatient (i.e., dispensed during stay) and outpatient medications (i.e., historical meds). All medication names were normalized by mapping to RxNorm ingredient. Comorbidity and admission diagnosis, i.e., all patient refined diagnosis related group (APR-DRG) variables were collected from the University Healthsystem Consortium (UHC) data source in HERON. Comorbidity, medication, and admission diagnosis variables took either “yes” or “no” values.
Table 1. Clinical variables considered in building predictive models for hospital-acquired AKI

<table>
<thead>
<tr>
<th>Feature Category</th>
<th># of Variables</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Demographics</td>
<td>3</td>
<td>Age, gender, race</td>
</tr>
<tr>
<td>Vitals</td>
<td>5</td>
<td>BMI, diastolic BP, systolic BP, pulse, temperature</td>
</tr>
<tr>
<td>Lab tests</td>
<td>14</td>
<td>Albumin, ALT, AST, Ammonia, Blood Bilirubin, BUN, Ca, CK-MB, CK, Glucose, Lipase, Platelets, Troponin, WBC</td>
</tr>
<tr>
<td>Comorbidities</td>
<td>29</td>
<td>UHC comorbidity</td>
</tr>
<tr>
<td>Admission diagnosis</td>
<td>315</td>
<td>UHC APR-DRG</td>
</tr>
<tr>
<td>Medications</td>
<td>1682</td>
<td>All medications are mapped to RxNorm ingredient</td>
</tr>
</tbody>
</table>

In the final dataset, vitals, lab test, and medication variables were time-stamped (with resolution to the hour and minute) relative to the admission date, referred here as time-dependent variables. Comorbidities, admission diagnosis, and demographics were presumed to be available at admission and not time-dependent.

Table 2. Categories for vital variable categories

<table>
<thead>
<tr>
<th>Vitals Categories</th>
<th>Categories</th>
</tr>
</thead>
<tbody>
<tr>
<td>BMI</td>
<td>&lt;18.5, [18.5–24.9], [25.0–29.9], &gt;30.0, Unknown</td>
</tr>
<tr>
<td>Diastolic BP</td>
<td>&lt;80, [80–89], [90–99], &gt;100, Unknown</td>
</tr>
<tr>
<td>Systolic BP</td>
<td>&lt;120, [120–139], [140–159], &gt;160, Unknown</td>
</tr>
<tr>
<td>Pulse</td>
<td>&lt;50, [50–65], [66–80], [81–100], &gt;100, Unknown</td>
</tr>
<tr>
<td>Temperature</td>
<td>&lt;95.0, [95.0–97.6], [97.7–99.5], [99.5–104.0], &gt;104.0, Unknown</td>
</tr>
</tbody>
</table>

Evaluation Design

Model evaluation was designed to answer three specific questions: (1) Will data prior to admission improve predictive models’ performance; (2) How early and accurately can AKI be forecasted; (3) How strong each type of risk factors affects the model performance. In this study, we introduced a data collection window, denoted as \([\text{lower\_bound}, \text{upper\_bound}]\). For the first question, we assessed model performance by varying the window’s lower bound with a fixed max upper bound set at 1-day prior to AKI event, i.e., making AKI prediction 1-day prior using different amount of clinical data (Figure 1). The initial data collection window’s lower bound is set at admission and we increased its width with resolution to the day.

![Figure 1. An illustration of adjusting the lower bound of data collection window before the admission date](image)

For the second question, we assessed model performance by lengthening the interval between the prediction time and AKI event (Figure 2). For AKI encounters, AKI event date was used as an anchor for moving the prediction point. For non-AKI encounters, discharge date was used as the anchor point.

To illustrate with an example, patient X was admitted to the hospital on 2015-06-05 (\(\text{Admission\_date}\)) and developed AKI on 2015-06-07 (\(\text{AKI\_date}\)), then the initial data collection window for prediction 1-day prior to AKI occurrence would be \([\text{Admission\_date}, \text{AKI\_date} - 1]\) which is [2015-06-05, 2015-06-06]. For a non-AKI example, patient Y
stayed at the hospital between 2014-02-15 (Admission_date) and 2014-02-18 (Discharge_date), then the data collection window for prediction 1-day prior would be [2014-02-15, 2014-02-17]. To assess whether adding data 1-day prior to AKI would improve predictive performance, the data collection window becomes [Admission_date – 1, AKI_date – 1]; thus for patient X, it becomes [2015-06-04, 2015-06-06]. To assess whether an accurate prediction can be made 2-days prior to AKI, the window upper bound will be AKI_date – 2. In conclusion, there are two parameters to adjust for evaluation: Admission_date – m is the mth day before the hospital admission date and AKI_date – n is the nth day before AKI occurrence date.

For the third question, we adopted two approaches to evaluate the strength of each group of risk factors (i.e., demographics, labs, vitals, medications, and comorbidities) used in the model. In the first approach, we removed one group of attributes and trained the predictive models on the remaining four groups of attributes, resulting in five transformed datasets. In the second approach, we trained the predictive models with only one group at a time. Thus, another five transformed datasets were obtained, each of which only contains one group of attributes.

**Experimental Methodology**

Three different machine learning methods – Logistic Regression37, Random Forest38, and AdaboostM139 were used for building predictive models using the Weka software library40. The number of decision trees in Random Forest model was 100. For AdaBoostM1, the iteration number was 500, using DecisionStump as the base learner. All models were evaluated using the standard 10-fold cross-validation.

The area under the receiver operating characteristic curve (AUC), precision and recall are used to report and compare performance of the models. AUC provides a single measurement of the performance of an ROC curve, which is a graphical plot of the sensitivity or true positive rate against the false positive rate (1 - specificity). Sensitivity is the proportion of actual positives that are correctly identified as such (i.e. SN = TP/(TP+FN)) and specificity measures the proportion of actual negatives that are correctly predicted as such (i.e. SP = TN/(TN+FP)). Precision is the proportion of true positives against all predicted positive results (i.e. P = TP/(TP+FP)). Recall is the same as the true positive rate or sensitivity.

Our datasets are highly imbalanced with an approximate 1:10 positive (AKI) to negative (non-AKI) ratio. With such an imbalanced dataset, most classifiers will favor the majority class (non-AKI) because they are designed to maximize the overall number of correct predictions, thus resulting in poor accuracy in the minority class (AKI) prediction. Current state-of-art correction techniques to account for class imbalance are generally data-based and algorithm-based approaches41. The data-based approach uses sampling technique by either under-sampling the majority class or over-sampling the minority class. The algorithm-based approach modifies the classification algorithm such as through adjusting decision threshold. In this study, we under-sampled the majority class by randomly selecting a subset of non-AKI samples such that AKI to non-AKI sample ratio is 1:1. Hence a model performing better than random classifier must achieve an AUC larger than 0.5.

**RESULTS**

**Data Characteristics**

Distribution of patient demographic variables in AKI and non-AKI encounters is listed in Table 3. In our cohort, the odds ratio between AKI and non-AKI is not significant for all age groups; however, it does increase with age which is also observed in Matheny et al31. Odds ratios for gender and race are also within similar range as reported31.
Table 3. Distribution of patient demographic variables in AKI and non-AKI encounters

<table>
<thead>
<tr>
<th>Demographics</th>
<th>AKI (n = 4405)</th>
<th>Non-AKI (n = 44550)</th>
<th>Odds Ratio (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age, n(%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18-25</td>
<td>291 (6.6)</td>
<td>3889 (8.7)</td>
<td>0.74 (0.65 – 0.84)</td>
</tr>
<tr>
<td>26-35</td>
<td>519 (11.8)</td>
<td>6612 (14.8)</td>
<td>0.77 (0.70 – 0.84)</td>
</tr>
<tr>
<td>36-45</td>
<td>734 (16.7)</td>
<td>7736 (17.4)</td>
<td>0.95 (0.88 – 1.03)</td>
</tr>
<tr>
<td>46-55</td>
<td>1296 (29.4)</td>
<td>12863 (28.9)</td>
<td>1.03 (0.96 – 1.10)</td>
</tr>
<tr>
<td>56-64</td>
<td>1565 (35.5)</td>
<td>13450 (30.2)</td>
<td>1.27 (1.19 – 1.36)</td>
</tr>
<tr>
<td>Gender, n (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>1779 (40.4)</td>
<td>20533 (46.1)</td>
<td>0.79 (0.74 – 0.84)</td>
</tr>
<tr>
<td>Male</td>
<td>2626 (59.6)</td>
<td>24017 (53.9)</td>
<td>1.26 (1.18 – 1.34)</td>
</tr>
<tr>
<td>Race, n (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>3133 (71.1)</td>
<td>32680 (73.4)</td>
<td>0.89 (0.84 – 0.96)</td>
</tr>
<tr>
<td>African</td>
<td>739 (16.8)</td>
<td>6771 (15.2)</td>
<td>1.12 (1.04 – 1.22)</td>
</tr>
<tr>
<td>Asian</td>
<td>31 (0.7)</td>
<td>396 (0.9)</td>
<td>0.79 (0.55 – 1.14)</td>
</tr>
<tr>
<td>Other</td>
<td>502 (11.4)</td>
<td>4703 (10.5)</td>
<td>1.09 (0.99 – 1.20)</td>
</tr>
</tbody>
</table>

Number of encounters in which AKI occurred in different number of days from the time of admission is shown in Table 4. The largest proportion (23.8%) of AKI encounters occurred on the 1st day after hospitalization.

Table 4. Number of encounters in which AKI occurred within different intervals from time of admission

<table>
<thead>
<tr>
<th>Days after Admission</th>
<th>Number of AKI events (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1047 (23.8)</td>
</tr>
<tr>
<td>2</td>
<td>959 (21.8)</td>
</tr>
<tr>
<td>3</td>
<td>554 (12.6)</td>
</tr>
<tr>
<td>4</td>
<td>405 (9.2)</td>
</tr>
<tr>
<td>5</td>
<td>296 (6.7)</td>
</tr>
<tr>
<td>&gt;6</td>
<td>1144 (25.9)</td>
</tr>
</tbody>
</table>

Results of AKI Prediction

Table 5 and Table 6 show model performance in terms of AUC values and precision and recall respectively over data collected both after and before hospital admissions. Table 7 and Table 8 show the results for evaluating early prediction.

Table 5. AUC values of prediction models on data collected before and after hospital admission

<table>
<thead>
<tr>
<th>Data Collection Window</th>
<th>Classification Models</th>
</tr>
</thead>
<tbody>
<tr>
<td>[Admission_date, AKI_date-1]</td>
<td>Random Forest (0.765)</td>
</tr>
<tr>
<td>[Admission_date-1, AKI_date-1]</td>
<td>AdaBoostM1 (0.751)</td>
</tr>
<tr>
<td>[Admission_date-7, AKI_date-1]</td>
<td>Logistic (0.763)</td>
</tr>
<tr>
<td>[Admission_date-15, AKI_date-1]</td>
<td>Random Forest (0.747)</td>
</tr>
<tr>
<td>[Admission_date-30, AKI_date-1]</td>
<td>AdaBoostM1 (0.732)</td>
</tr>
</tbody>
</table>

Table 6. Precision and recall of prediction models on data collected before and after hospital admission

<table>
<thead>
<tr>
<th>Data Collection Window</th>
<th>Classification Models</th>
</tr>
</thead>
<tbody>
<tr>
<td>[Admission_date, AKI_date-1]</td>
<td>Logistic (0.704 / 0.711)</td>
</tr>
<tr>
<td>[Admission_date-1, AKI_date-1]</td>
<td>Logistic (0.690 / 0.686)</td>
</tr>
<tr>
<td>[Admission_date-7, AKI_date-1]</td>
<td>Logistic (0.689 / 0.691)</td>
</tr>
<tr>
<td>[Admission_date-15, AKI_date-1]</td>
<td>Logistic (0.690 / 0.692)</td>
</tr>
<tr>
<td>[Admission_date-30, AKI_date-1]</td>
<td>Logistic (0.690 / 0.692)</td>
</tr>
</tbody>
</table>
The time window \([\text{Admission}_\text{date} - m, \ \text{AKI}_\text{date} - n]\) means that data are collected based on the following rule: the time dependent variables have at least one value available after \(\text{Admission}_\text{date} - m\) and before \(\text{AKI}_\text{date} - n\). \(\text{Admission}_\text{date} - m\) is the \(m\)th day before the hospital admission and \(\text{AKI}_\text{date} - n\) is the \(n\)th day before AKI occurrence date.

Table 7. AUC of prediction models when adjusting the prediction points/upper bound of data collection window

<table>
<thead>
<tr>
<th>Data Collection Window</th>
<th>Classification Models</th>
</tr>
</thead>
<tbody>
<tr>
<td>([\text{Admission}<em>\text{date}, \ \text{AKI}</em>\text{date}-1])</td>
<td>Random Forest 0.765</td>
</tr>
<tr>
<td></td>
<td>AdaBoostM1 0.751</td>
</tr>
<tr>
<td></td>
<td>Logistic 0.763</td>
</tr>
<tr>
<td>([\text{Admission}<em>\text{date}, \ \text{AKI}</em>\text{date}-2])</td>
<td>Random Forest 0.733</td>
</tr>
<tr>
<td></td>
<td>AdaBoostM1 0.727</td>
</tr>
<tr>
<td></td>
<td>Logistic 0.731</td>
</tr>
<tr>
<td>([\text{Admission}<em>\text{date}, \ \text{AKI}</em>\text{date}-3])</td>
<td>Random Forest 0.709</td>
</tr>
<tr>
<td></td>
<td>AdaBoostM1 0.705</td>
</tr>
<tr>
<td></td>
<td>Logistic 0.691</td>
</tr>
<tr>
<td>([\text{Admission}<em>\text{date}, \ \text{AKI}</em>\text{date}-4])</td>
<td>Random Forest 0.688</td>
</tr>
<tr>
<td></td>
<td>AdaBoostM1 0.690</td>
</tr>
<tr>
<td></td>
<td>Logistic 0.651</td>
</tr>
<tr>
<td>([\text{Admission}<em>\text{date}, \ \text{AKI}</em>\text{date}-5])</td>
<td>Random Forest 0.670</td>
</tr>
<tr>
<td></td>
<td>AdaBoostM1 0.678</td>
</tr>
<tr>
<td></td>
<td>Logistic 0.633</td>
</tr>
</tbody>
</table>

Table 8. Precision and recall when adjusting the prediction points/upper bounds of data collection window

<table>
<thead>
<tr>
<th>Data Collection Window</th>
<th>Classification Models</th>
</tr>
</thead>
<tbody>
<tr>
<td>([\text{Admission}<em>\text{date}, \ \text{AKI}</em>\text{date}-1])</td>
<td>Random Forest (Precision/Recall) 0.692 / 0.711</td>
</tr>
<tr>
<td></td>
<td>AdaBoostM1 (Precision/Recall) 0.662 / 0.736</td>
</tr>
<tr>
<td></td>
<td>Logistic (Precision/Recall) 0.704 / 0.711</td>
</tr>
<tr>
<td>([\text{Admission}<em>\text{date}, \ \text{AKI}</em>\text{date}-2])</td>
<td>Random Forest (Precision/Recall) 0.675 / 0.661</td>
</tr>
<tr>
<td></td>
<td>AdaBoostM1 (Precision/Recall) 0.643 / 0.714</td>
</tr>
<tr>
<td></td>
<td>Logistic (Precision/Recall) 0.678 / 0.675</td>
</tr>
<tr>
<td>([\text{Admission}<em>\text{date}, \ \text{AKI}</em>\text{date}-3])</td>
<td>Random Forest (Precision/Recall) 0.650 / 0.650</td>
</tr>
<tr>
<td></td>
<td>AdaBoostM1 (Precision/Recall) 0.625 / 0.687</td>
</tr>
<tr>
<td></td>
<td>Logistic (Precision/Recall) 0.651 / 0.646</td>
</tr>
<tr>
<td>([\text{Admission}<em>\text{date}, \ \text{AKI}</em>\text{date}-4])</td>
<td>Random Forest (Precision/Recall) 0.634 / 0.637</td>
</tr>
<tr>
<td></td>
<td>AdaBoostM1 (Precision/Recall) 0.628 / 0.656</td>
</tr>
<tr>
<td></td>
<td>Logistic (Precision/Recall) 0.620 / 0.623</td>
</tr>
<tr>
<td>([\text{Admission}<em>\text{date}, \ \text{AKI}</em>\text{date}-5)]</td>
<td>Random Forest (Precision/Recall) 0.623 / 0.610</td>
</tr>
<tr>
<td></td>
<td>AdaBoostM1 (Precision/Recall) 0.616 / 0.674</td>
</tr>
<tr>
<td></td>
<td>Logistic (Precision/Recall) 0.608 / 0.605</td>
</tr>
</tbody>
</table>

Table 9. The performance of Random Forest model on cohort data by removing one attribute group (Data Collection Window: \([\text{Admission}_\text{date}, \ \text{AKI}_\text{date}-1]\))

<table>
<thead>
<tr>
<th>Removed attributes group</th>
<th>Precision</th>
<th>Recall</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Demographics</td>
<td>0.692</td>
<td>0.701</td>
<td>0.762</td>
</tr>
<tr>
<td>Vitals</td>
<td>0.685</td>
<td>0.699</td>
<td>0.756</td>
</tr>
<tr>
<td>Labs</td>
<td>0.678</td>
<td>0.701</td>
<td>0.755</td>
</tr>
<tr>
<td>Admission DRGs and Comorbidities</td>
<td>0.667</td>
<td>0.681</td>
<td>0.723</td>
</tr>
<tr>
<td>Medications</td>
<td>0.629</td>
<td>0.622</td>
<td>0.679</td>
</tr>
</tbody>
</table>

Table 10. The performance of Random Forest model on cohort data containing only one attribute group (Data Collection Window: \([\text{Admission}_\text{date}, \ \text{AKI}_\text{date}-1]\))

<table>
<thead>
<tr>
<th>Reserved attributes group</th>
<th>Precision</th>
<th>Recall</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Demographics</td>
<td>0.538</td>
<td>0.618</td>
<td>0.551</td>
</tr>
<tr>
<td>Vitals</td>
<td>0.550</td>
<td>0.542</td>
<td>0.558</td>
</tr>
<tr>
<td>Labs</td>
<td>0.587</td>
<td>0.211</td>
<td>0.546</td>
</tr>
<tr>
<td>Admission DRGs and Comorbidities</td>
<td>0.612</td>
<td>0.619</td>
<td>0.657</td>
</tr>
<tr>
<td>Medications</td>
<td>0.659</td>
<td>0.659</td>
<td>0.712</td>
</tr>
</tbody>
</table>

Table 9 and Table 10 show the experimental results of the Random Forest model on processed data by removing one group of attributes and by reserving one group of attributes respectively. These two series of experiments were performed to assess the range of the affect size of each risk factor group on prediction.

**DISCUSSION**

In this study, we built machine learning based AKI prediction models using structured EMR data for patients admitted to a hospital. Experimental results showed a good cross-validated discrimination performance, best AUC of 0.76 achieved by Random Forest for predicting AKI event 1-day prior, which was similar to other AKI risk stratification models that have been created for general inpatients (0.75)\(^{31}\), elderly (0.66)\(^{32}\), more specific clinical scenarios (0.74 to 0.77), and patients who have undergone coronary artery bypass grafting (0.72 to 0.81)\(^{1,3,25,26}\). To ensure that the classifier output is not affected by the different splits generated by the 10-fold cross-validation, we
examined variance of the AUC, precision, and recall measures by applying the Random Forest model on dataset with a collection window of \([\text{Admission\_date}, \text{AKI\_date} – 1]\). The standard deviations of AUC, precision and recall are 0.01, 0.01, and 0.02 respectively, which shows that there was no great variance in the 10-fold cross-validation.

For the primary objective, we assessed how early and accurately general inpatient AKI can be predicted. Results in Tables 7 and 8 showed that model performance indeed degrades as the time window between the prediction time and AKI event time lengthened, from Random Forest’s AUC of 0.76 at 1-day prior to 0.67 at 5-days prior. While comparing different machine learning algorithms, results in Table 7 showed that Random Forest achieved the best cross-validated AUC of 0.76 for predicting AKI 1-day prior but when the time to event horizon is lengthened to 4 or 5-days prior, AdaBoostM1 had a slightly better AUC than Random Forest. However, in terms of precision and recall in Table 8, Logistic Regression actually had better precision with the same or better recall compared to Random Forest for 1 and 2-day prior, respectively.

In order for a predictive model to be clinically useful, the 15th Acute Dialysis Quality Initiative (ADQI) consensus conference\(^{30}\) in 2016 recommended forecasting AKI events with a horizon of 48 to 72 hours. Although it would be advantageous to predict AKI events as early as possible, lengthening the prediction time to event horizon will reduce accuracy and the ADQI consensus group believes that 2 to 3 days would give physicians adequate time to modify practice, optimize hemodynamics, and mitigate potential injury without sacrificing too much in predictive power\(^{30}\). Our study showed that the best performing model with Random Forest can forecast AKI 2-days and 3-days prior with AUC of 0.73 and 0.70, respectively.

For the secondary study objective, we assessed whether adding data prior to admission would improve model performance where the data collection window was extended to 1, 7, 15, and 30 days before admission. One would intuitively think more data is better, but the contrary was observed in results (Table 5 and 6). As the study cohort contains encounters with various length of stay and the number of days AKI occurs relative to admission in the positive samples varies greatly from 1 day to 359 days, we suspect additional data prior to admission may impact model performance differently for patients who develop AKI on the day after admission vs. five days after because amount of data available during stay is dramatically different. Therefore, we conducted an analysis on encounters in which AKI occurred 1 day after admission, comparing Random Forest’s performance with (i.e., 1 and 7 days) and without data prior to admission. Interestingly, analysis results on this sub-cohort exhibited the same trend as observed in the complete cohort, where AUC for only using data after admission is 0.84 vs. AUC for adding data from 1-day before admission is 0.81 and 7-days prior is 0.80. This implies adding data prior to admission does not improve AKI prediction performance in the general inpatient AKI population. This prompts us to further analyze the impact of using data from the entire encounter vs. only previous day on prediction performance.

We further screened data with the time window \([\text{AKI\_date}-1, \text{AKI\_date}-1]\), i.e. using data on the day of prediction and compared the performance with using data collected within \([\text{Admission\_date}, \text{AKI\_date}-1]\). Based on results in Table 11, it seems to imply that prediction using the most recent one-day data can improve AUC and precision at the expense of degrading recall ratio. F-score decreased from 0.701 to 0.687 in contrast to the AUC increase from 0.765 to 0.783. However, this may also suggest the models were overfitting the temporal clinical variables relative to the static demographics, admission diagnoses, and comorbidities. Evaluating the contribution of each clinical variable and refining their representations are future directions for this research.

**Table 11.** Comparing performance (AUC/Precision/Recall) on entire hospitalization data vs. recent one-day data

<table>
<thead>
<tr>
<th>Data Collection Window</th>
<th>Random Forest (AUC/Precision/Recall)</th>
<th>AdaBoostM1 (AUC/Precision/Recall)</th>
<th>Logistic (AUC/Precision/Recall)</th>
</tr>
</thead>
<tbody>
<tr>
<td>([\text{Admission_date}, \text{AKI_date}-1])</td>
<td>0.765 / 0.692 / 0.711</td>
<td>0.751 / 0.662 / 0.736</td>
<td>0.763 / 0.704 / 0.711</td>
</tr>
<tr>
<td>([\text{AKI_date}-1, \text{AKI_date}-1])</td>
<td>0.783 / 0.721 / 0.655</td>
<td>0.768 / 0.674 / 0.720</td>
<td>0.768 / 0.709 / 0.709</td>
</tr>
</tbody>
</table>

For the third objective, we assessed the effect size of each risk factor type in AKI prediction by applying the Random Forest model on datasets with a collection window of \([\text{Admission\_date}, \text{AKI\_date} – 1]\). Based on results in Table 9 and Table 10, medications play the biggest role in the 1-day prior AKI prediction performance followed by the combination of admission DRG and comorbidity. This is promising as medications are modifiable and clinicians may consider alternative therapies. The demographics variables had the least effect on prediction performance, which may be due to the fact that we limited our cohort to a younger cohort of 18 to 64 year olds and utilized encounter-level rather than patient-level data for prediction. A patient may have multiple hospital encounters that satisfy the cohort inclusion criteria and he/she may experience AKI in one encounter and not in another. Thus, the
same patient can belong to both the AKI and non-AKI class, making it difficult for algorithms to distinguish the two classes based on the patient-level demographic information.

**Limitations**

There are several limitations in the interpretation of results in this study. First of all, the predictive models were based off a younger cohort (18 to 64 years old at admission), which may not be generalizable to an older cohort. Elderly is known to be at increased risk for AKI due to longer exposure to chronic diseases and nephrotoxins, thus our future studies will conduct independent subpopulation analysis for the elderly. Second, we limited the analysis to patients who were admitted to the hospital with a minimum eGFR of 60 mL/min/1.73m² and must have normal serum creatinine on the day of admission. Although patients with reduced eGFR are at increased risk for AKI, it is difficult to determine which of these patients had hospital-acquired vs. community-acquired AKI. Third, comorbidity data utilized in the predictive models were obtained from UHC, which is widely known to be well adjudicated, but not immediately available. This study treated UHC comorbidities as non-time dependent as if the clinical team would know all comorbidities at admission, which may misrepresent comorbidities that developed during the admission. Future studies will evaluate the performance of diagnosis codes from the EMR problem list relative to comorbidity information derived from billing systems. This may change the results we observed in this study that data prior to admission does not improve performance and also more accurately represent the information available to the clinical team to support adverse event surveillance. Last but not least, the study did not use urine output to define AKI nor include it as a risk variable. Although urine output is one of the diagnostic criteria of AKI, many members of the Acute Kidney Injury Network (AKIN) concerned that urine output is not specific enough for the designation of AKI because it can be influenced by factors other than renal health.

**CONCLUSIONS**

Predicting AKI early and accurately allows clinicians to take timely preventative or therapeutic measures. This study investigates the impact of data completeness and prediction time points on the performance of forecasting in-hospital AKI in a general inpatient population. Three machine learning algorithms, Random Forest, AdaBoostM1, and Logistic Regression, were built on clinical datasets screened with different data collection windows. The Random Forest classifier outperformed Logistic Regression and AdaBoostM1 with AUC values for 1, 2, and 3-days prior being 0.765, 0.733, and 0.709, respectively. In this study, the data prior to hospital admission did not improve prediction performance and medication played the biggest role in prediction performance.

**ACKNOWLEDGEMENT**

The HERON data repository described in the paper is supported by institutional funding from the University of Kansas Medical Center and CTSA grant UL1TR000001 from NCRR/NIH. The work is partially supported by PCORI CDRN-15-26643.

**REFERENCES**

Specifications of Clinical Quality Measures and Value Set Vocabularies Shift Over Time: A Study of Change through Implementation Differences

Raja A. Cholan, BS, Nicole G. Weiskopf, PhD, Douglas L. Rhoton, Nicholas V. Colin, MA, Rachel L. Ross, BA, Melanie N. Marzullo, BS, Bhavaya Sachdeva, MPH, David A. Dorr, MD, MS; Oregon Health & Science University, Portland, Oregon

Abstract

Clinical quality measures (CQMs) aim to identify gaps in care and to promote evidence-based guidelines. Official CQM definitions consist of a measure’s logic and grouped, standardized codes to define key concepts. In this study, we used the official CQM update process to understand how CQMs’ meanings change over time. First, we identified differences between the narrative description, logic, and the vocabulary specifications of four standardized CQMs’ definitions in subsequent versions (2015, 2016, and 2017). Next, we implemented the various versions in a quality measure calculation registry to understand how the differences affected calculated prevalence of risk and measure performance. Global performance rates changed up to 5.32%, and an increase of up to 28% new patients was observed for key conditions between versions. Updates to definitions that change a measure’s logic and choices to include/exclude codes in value set vocabularies changes measurement of quality and likely introduces variation by implementation.

Introduction

From a public health perspective, the central goal of the health care system is to increase population health; one way of measuring that effect is by determining the quality of care delivered. The Institute of Medicine defines “health care quality” as “the degree to which health care services for individuals and populations increase the likelihood of desired health outcomes and are consistent with current professional knowledge.”1 Electronic clinical quality measures (CQMs) intend to measure quality by tracking various evidence-based elements of structure, process, and outcomes using clinical data recorded in Electronic Health Record systems (EHRs).2, 3 Although this measurement quality goes back more than a decade, CQMs have become fundamental parts of payment and health care; for instance, the Quality Payment Program, following the Medicare Access and CHIP Reauthorization Act of 2015, proposes to adjust clinicians’ Medicare payments in 2019 between -9 and +9% depending on CQM data submitted by individual clinicians in 2017 and 2018.4

Quality-based payment programs like the Quality Payment Program, typically require measures that are endorsed by the Centers for Medicare and Medicaid Services (CMS), and/or the National Quality Forum (NQF), in order to improve the reliability and validity of CQMs.5-7 These standardized electronic measures have specific definitions and logic encoded in the “Health Quality Measure Format” (HQMF), a Health Level Seven International standard for representing CQMs. HQMFs include machine-readable specifications encoded in Extensible Markup Language, as well as human-readable descriptions of the measure logic and components.8 The HQMF for a CQM assessing the use of aspirin or another antithrombotic for patients with ischemic vascular disease (CMS164; NQF0068), for example, provides computable and human-readable definitions of the measure’s denominator criteria (e.g., patients aged 18-85 years of age with ischemic vascular disease), those who meet the measure’s numerator criteria (e.g., patients taking aspirin medication), and those who are excluded from the measure’s calculation (e.g., patients taking anticoagulant medications). Official CQMs are also defined using value sets, which encode pertinent clinical concepts into controlled terminologies. Value sets are grouped codes from standardized vocabularies (e.g., SNOMED CT, RXNORM, ICD-10-CM, LOINC) in the Quality Data Model and consist of clinical concepts (i.e., value set names), and concept definitions (i.e., code groups) to define a measure’s specifications.

Each year, CMS updates the electronic specifications of the CQMs it approves for submission to its various reporting programs and publishes updated HQMFs in their “eCQM Library”.9 The National Library of Medicine’s Value Set Authority Center (VSAC), subsequently, in collaboration with the Office of the National Coordinator for Health Information Technology, updates its electronic value set definitions at regular intervals, based both on these annual updates and on changes in the terminologies themselves.10, 11 For the aforementioned Aspirin measure, between 2015 (i.e., CMS164v4) and 2017 (i.e., CMS164v5), the HQMF was updated once with new exclusion criteria (patients taking anticoagulant medications) – evidence shows that these patients have an increased risk of bleeding.12 The value sets that define the concepts “ischemic vascular disease” and “acute myocardial infarction” also changed twice between versions with the addition of several ICD-10-CM, ICD-9-CM, and SNOMED CT codes.

A CQM is implemented by mapping the right clinical data from the right sources in the EHR to generate a CQM performance score. Many clinics may not have the skills to interpret the encoded HQMFs/value sets, or the technical capacity to implement custom queries for generating CQM performance data. Therefore, they might increasingly rely on the use of Office of the National Coordinator Certified EHR Technology, or a recognized quality
measure calculation registry, such as the American Board of Family Medicine’s PRIME Registry, for CQM reporting. Regardless of the data source used to generate CQM data – whether it is a certified EHR, registry system, or custom query – CQM implementations should reflect the most up-to-date standards and specific criteria encoded in official HQMFs and value sets. However, for clinics that rely on external measure implementers, it can be difficult to verify whether or not CQM specifications align with the standardized, most-updated CQM definitions and versions. As technical assistance providers for a large primary care practice transformation initiative, we helped 93 clinics in Oregon leverage technology to produce and improve CQMs. Of these clinics, 77% have generated and reported aggregate, practice-level data for the Aspirin measure. Of these (as of Quarter 1, 2017) roughly 1/4 have updated to the 2016/2017 version definition and 3/4 still use the 2015 version. Additionally, sampling the 23% that have custom CQM implementations demonstrate partial value sets and approximate measure logic due to current report-building system limitations. EHR and registry vendors, meanwhile, typically include official CQM identifiers (e.g., NQF ID, CMS ID) but do not always explicitly provide a measure’s version, and rarely provide details about query logic, or specific code sets used. As CQM reporting continues to be operationalized, clinicians could unknowingly have outdated implementations of CQMs, and innocuously report these data with the intent of reporting current CQM versions; and, CMS could unknowingly compare data from different versions of the same CQMs to benchmark clinics and adjust Medicare payments. National quality reform programs and practice transformation efforts need standard representations of CQMs to enable consistent reporting and comparison between clinics across the country. However, the impact of these implementation variations is unknown and dependent upon how substantial the differences between versions are.

Many of these version changes are related to terminology design and evolution, while others may be related to actual evidence changes. Vocabularies like ICD-10-CM, RxNorm, and SNOMED CT, for instance, have a strict hierarchical structure – with differing levels of granularity, concept orientation, and consideration of redundancy – to define various clinical concepts. Value set authors and measure developers, therefore, make choices about the specific codes to include/exclude when defining broad clinical concepts such as stroke or diabetes, for example. In previous work, we identified that key differences in the value set creation process for two similar definitions of the same CQM led to variations in the calculated prevalence of patients at high-risk from key conditions. More challenges arise when value set vocabularies are periodically updated, as some terms and codes may change meaning over time. Cimino noted this previously: “if a concept is changed in such a way as to alter its meaning, what happens to the ability of the aggregated patient data that are coded before and after the change?” Yu and Cimino studied code changes to annual ICD-9-CM codes and found that the meaning of terms may drift, but the original codes in the clinical data remained the same: “researchers who make use of multi-year ICD-9-CM data are actually unaware that the meaning of their data are changing over time.” Other changes may reflect evidence changes, such as recommendations to treat patients with coronary artery disease with antiplatelet agents when they are also on antithrombotics.

With the advent of the Quality Payment Program, regular CQM reporting will continue to increase in importance. Standardized and up-to-date CQM implementations are crucial for quality-based reform programs to compare impact on population health and to facilitate performance improvement, but the reality in clinics is that CQM implementations may not always adhere to the latest versions. Studying the impact of CQM definition updates, and examining code drift, and taxonomy shifts, in value set vocabularies might help in understanding one aspect of the vast variation that exists in current CQM implementations. In this study, our objective is to use the official CQM update process to better understand how updates to official CQM definitions and value sets may change the interpretation of CQMs’ meanings over time. Our aims are to:

1. Identify the common differences for select CQM definitions that may occur as part of regular updates;
2. Understand how variations in old and current HQMFs and value sets for the same CQMs change a measure’s meaning through implementation differences in a CQM calculation registry;
3. Discuss suggestions for disseminating updates to CQM definitions and value sets to implementers.

Our hypothesis was that the variation caused in the CQM definition and value set annual update process would lead to differences in measure populations and performance estimates.

**Methods**

**Overview:** We identified differences between the narrative description and logic of four CQM definitions from the 2015, 2016, and 2017 versions of the measures. We then identified differences in the vocabulary specifications – unique identifiers, concepts, code groups, and coding systems – between the various value sets used to define the same global concepts in each annual version. Next, we implemented the various versions in a quality measure calculation registry to understand how the differences in the annual value sets and measure logic affected calculated prevalence of risk and measure performance.
Table 1. Descriptions of the four standardized CQMs included

<table>
<thead>
<tr>
<th>Measure</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Appropriate Aspirin Use” – CMS164v5; NQF9064</td>
<td>Percentage of patients 18 years of age and older who were diagnosed with acute myocardial infarction (AMI), coronary artery bypass graft (CABG) or percutaneous coronary interventions (PCI) in the 12 months prior to the measurement period, or who had an active diagnosis of ischemic vascular disease (IVD) during the measurement period, and who had documentation of use of aspirin or another antiplatelet during the measurement period.</td>
</tr>
<tr>
<td>Ischemic Vascular Disease (IVD); Use of Aspirin or Another Antithrombotic</td>
<td></td>
</tr>
<tr>
<td>“Cholesterol Management by Statin Therapy” – CMS347v0; Quality ID 438</td>
<td>Percentage of the following patients-all considered at high risk of cardiovascular events-who were prescribed or were on statin therapy during the measurement period: Adults aged &gt;= 21 years who were previously diagnosed with or currently have an active diagnosis of clinical atherosclerotic cardiovascular disease (ASCVD); OR Adults aged &gt;21 years who have ever had a fasting or direct low-density lipoprotein cholesterol (LDL-C) level &gt;= 190 mg/dL or were previously diagnosed with or currently have an active diagnosis of familial or pure hypercholesterolemia; OR Adults aged 40-75 years with a diagnosis of diabetes with a fasting or direct LDL-C level of 70-189 mg/dL</td>
</tr>
<tr>
<td>Statin Therapy for the Prevention and Treatment of Cardiovascular Disease</td>
<td></td>
</tr>
<tr>
<td>“Diabetes A1C Control” – CMS122v5; NQF9059</td>
<td>Percentage of patients 18-75 years of age with diabetes who had hemoglobin A1c &gt; 9.0% during the measurement period.</td>
</tr>
<tr>
<td>Diabetes: Hemoglobin A1c (HbA1c) Poor Control (&gt;9%)</td>
<td></td>
</tr>
<tr>
<td>“Heart Failure, Beta Blocker Use” – CMS144v3; NQF0083</td>
<td>Percentage of patients aged 18 years and older with a diagnosis of heart failure (HF) with a current or prior left ventricular ejection fraction (LVEF) &lt; 40% who were prescribed beta-blocker therapy either within a 12 month period when seen in the outpatient setting OR at each hospital discharge.</td>
</tr>
<tr>
<td>Heart Failure (HF); Beta Blocker Therapy for Left Ventricular Systolic Dysfunction (LVSD)</td>
<td></td>
</tr>
</tbody>
</table>

Measure Inclusion: We first selected four standardized CQMs listed in Table 1 to include in this study. Each of these are included in the Quality Payment Program’s Merit-based Incentive Payment System as official 2017 CQMs. We chose these specific measures because they include value sets defining a variety of key conditions: ischemic vascular disease and overlapping cardiovascular disease, diabetes, and heart failure. In this study, we assessed implementation changes between older and newer versions of these four measures, and explored specific code variations of the value sets used to define the clinical concepts, and related them to taxonomy structure.

HQMF and Value Set Analysis: We compared the human-readable definitions contained in HQMFs of the four measures listed above, and noted any changes to the narrative description or encoded logic in the measures’ criteria between three subsequent versions – 2015, 2016, and 2017. We then searched for the four CQMs in the VSAC and downloaded each annual value set included in each version’s specifications. Next, we counted the unique object identifiers (OIDs) used to specify distinct concepts, and the codes from standard vocabularies that define each category.

Measure Implementation and Comparison: Next, we implemented the various versions of the four measures in the Integrated Care Coordination Information System (ICCIS), a quality measure calculation registry, to understand how the differences in the measure specifications affected calculated prevalence of risk and measure performance. ICCIS contains data mapped from a variety of sources from over 500,000 patients into a star database format that facilitates value set queries. Implementation of CQMs in ICCIS currently requires processing of human-readable CQM definitions and writing queries in Structured Query Language (SQL) against clinical sources to calculate CQM performance. After interpreting the HQMFs for the four selected CQMs, we implemented queries in ICCIS following the specific criteria for each annual version, and queried against the specific OIDs used in each version. Then, the “Aspirin Use” and “Cholesterol Management” CQMs – two of the more complex measures – were divided into sub-measures for each aggregate clinical concept to see which concepts, taxonomies, and individual codes were most responsible for the performance rate.

Data Inclusion Criteria: The measures were implemented against data from two primary care clinics serving approximately 25,000 patients in Portland, Oregon. The clinicians in each clinic used a fully functional EHR in their ambulatory work. We chose to implement the measures using data from these two clinics because they had high quality, suitable, structured data in the ICCIS database to generate reliable quality measure performance rates. The
clinics excluded from our analysis did not have an adequate average number of medications per patient stored as structured data in the ICCIS database; therefore we could not have reliably queried whether the patients were taking aspirin, statin, anticoagulant, or beta-blocker medications.

**Analysis.** To compare the changes in the measure definitions over time, we counted the individual codes in each value set, the taxonomies included, and measured the percent change of the value set codes (newly included and excluded codes) from 2015 to 2017. We then calculated performance rates at the same point in time with a 1 year measurement, or look-back period, and compared the included populations, the exceptions/exclusions, and overall performance.

**Results**

Of the four CQMs, the Aspirin measure’s HQMF was the only one with a change in its logic (added exclusion criteria) between 2015 and 2017 versions. The other three measures’ HQMFs remained constant. There were, however, changes in all four measures’ value set vocabulary definitions. An overview of the clinical concepts with code group changes between the 2015, 2016, and 2017 versions of the measures are found in Table 2. Overall, there were 64 concepts that remained constant between versions, and 11 concepts with code changes, of which eight (73%) resulted in an increase in count of codes, and three (27%) decreased in codes between 2015 and 2017 versions. Of those that changed, five (46%) concepts represented denominator criteria, two (18%) were for exceptions, one (9%) was for exclusions, and three (27%) represented numerator criteria. The seven value sets which represented denominator and exception criteria defined key diagnoses and were made up of ICD-10-CM, ICD-10-PCS, ICD-9-CM, and SNOMED CT codes. The concept “Percutaneous Coronary Intervention” (PCI) had the biggest increase in code count with the addition of 96 inpatient procedure (ICD10PCS) codes used to describe “dilation of coronary arteries”. The diagnosis “diabetes” increased in code count by 74%, with 167 new codes, of which 157 were ICD-10-CM codes describing specific diabetes diagnoses involving retinopathy, and the remainder were generic ICD-9-CM codes. The concept “acute myocardial infarction” (AMI) had a 58% increase in code count; all of the newly added codes were ICD-10CM codes.

<table>
<thead>
<tr>
<th>Measure(s)</th>
<th>Value Set Name/Concept</th>
<th>Criteria</th>
<th>Count of Codes 2015 version</th>
<th>Count of Codes 2016 version</th>
<th>Count of Codes 2017 version</th>
<th>% Change of Codes (2015 to 2017)</th>
<th>Taxonomies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aspirin</td>
<td>Anticoagulant Medications</td>
<td>Exclusions</td>
<td>0</td>
<td>92</td>
<td>92</td>
<td>-</td>
<td>RXNORM</td>
</tr>
<tr>
<td></td>
<td>Aspirin and Other Antiplatelets</td>
<td>Numerator</td>
<td>31</td>
<td>24</td>
<td>24</td>
<td>-23%</td>
<td>RXNORM</td>
</tr>
<tr>
<td>Aspirin / Cholesterol</td>
<td>Acute Myocardial Infarction</td>
<td>Denominator</td>
<td>62</td>
<td>98</td>
<td>98</td>
<td>58%</td>
<td>SNOMEDCT, ICD10CM, ICD9CM</td>
</tr>
<tr>
<td></td>
<td>Ischemic Vascular Disease</td>
<td>Denominator</td>
<td>561</td>
<td>744</td>
<td>786</td>
<td>40%</td>
<td>SNOMEDCT, ICD10CM, ICD9CM</td>
</tr>
<tr>
<td>Cholesterol</td>
<td>Ischemic Stroke</td>
<td>Denominator</td>
<td>128</td>
<td>108</td>
<td>126</td>
<td>-1.6%</td>
<td>SNOMEDCT, ICD10CM, ICD9CM</td>
</tr>
<tr>
<td></td>
<td>PCI</td>
<td>Denominator</td>
<td>80</td>
<td>80</td>
<td>176</td>
<td>120%</td>
<td>SNOMEDCT, ICD10CM, ICD9CM</td>
</tr>
<tr>
<td></td>
<td>Pregnancy Dx</td>
<td>Exceptions</td>
<td>1981</td>
<td>1987</td>
<td>2036</td>
<td>2.8%</td>
<td>SNOMEDCT, ICD10CM, ICD9CM</td>
</tr>
<tr>
<td>Cholesterol / Diabetes</td>
<td>Diabetes</td>
<td>Denominator</td>
<td>225</td>
<td>236</td>
<td>392</td>
<td>74%</td>
<td>SNOMEDCT, ICD10CM, ICD9CM</td>
</tr>
<tr>
<td>Heart Failure</td>
<td>Asthma</td>
<td>Exclusions</td>
<td>81</td>
<td>83</td>
<td>83</td>
<td>2.5%</td>
<td>SNOMEDCT, ICD10CM, ICD9CM</td>
</tr>
<tr>
<td></td>
<td>Beta Blocker Therapy</td>
<td>Numerator</td>
<td>80</td>
<td>92</td>
<td>92</td>
<td>15%</td>
<td>RXNORM</td>
</tr>
<tr>
<td></td>
<td>Beta Blocker Therapy Ingredient</td>
<td></td>
<td>50</td>
<td>49</td>
<td>49</td>
<td>-2.0%</td>
<td>RXNORM</td>
</tr>
</tbody>
</table>

Table 2. Differences in CQM vocabularies by concept and criteria for 2015, 2016, and 2017 versions
and ICD-9-CM codes describing “subsequent ST elevation (STEMI)”, “septal defects”, and “other current complications from AMI”. The codes used to define “ischemic vascular disease” (IVD) also increased by 40% with several new ICD-9-CM and ICD-10-CM codes to describe “atherosclerosis of bypass graft(s)”, “cerebral infarction”, and “unspecified atherosclerosis”. While the Aspirin and Cholesterol measures had a net increase in the number of codes included in their respective aggregate value sets, both also had codes deleted between the 2015 and 2017 versions. The Aspirin measure had nine codes removed: two were ICD-9-CM/ICD-10-CM codes to define “angina pectoris” for the concept IVD, and seven were RXNORM codes (Semantic Clinical Drugs or SCDs) to define various dosages of oral aspirin tablets/capsules. In the Cholesterol value sets, 14 ICD-9-CM/ICD-10-CM codes were removed for “pregnancy diagnoses”, “IVD” (2 codes for “angina”), “ischemic stroke” (1 code describing general cerebrovascular disease), and “hemorrhagic stroke” (3 codes describing nontraumatic subarachnoid hemorrhage); three SNOMED CT codes were also deleted for “CABG surgeries” (“aortocoronary artery bypass graft”), “ischemic stroke” (‘other cerebrovascular disease’), and “medical reason” (‘history of drug allergy’). The Heart Failure measure had 15 deleted codes: seven were SNOMEDCT codes to define “CABG surgeries” (‘aortocoronary artery bypass graft’), “ischemic” stroke (‘other cerebrovascular disease’), and “medical reason” (‘history of drug allergy’). The Heart Failure measure did not have any codes deleted between versions.

Table 3 shows the CQM performance results between the 2015 and 2017 versions of various implementations of the composite measures and aggregate concepts; it also shows the number of newly included and excluded patients between versions. The results in Table 3 only reflect performance data for the CQM implementations which exhibited a change in performance rates between versions. The 2017 version of the Aspirin measure, with the added exclusion criteria, had a 5.32% better performance rate than the 2015 version, and the denominator returned 119 (9%) fewer patients. The stand-alone implementation of the Aspirin measure for the AMI concept returned 24 (17%) more patients in version 2017 with an increase in performance rate of 2.62%. The composite Cholesterol measure had similar results between versions, with virtually identical adherence rates and denominators. The AMI-only version of the Cholesterol measure returned 37 more patients (27%) in the 2017 version, while the performance rate decreased by 2.31%. The stand-alone implementations of the Cholesterol measure for IVD and Diabetes had similar results with performance rates increasing by less than 1% between versions. We also implemented stand-alone versions of the Cholesterol measure for the aggregate concepts “ischemic stroke” and “PCI”; however, these versions did not change between 2015 and 2017 versions and the results are not included in Table 3. The Diabetes CQM had very similar performance rates between versions, with the 2017 version increasing in performance rate by 0.14%. The Heart Failure measure is not included in Table 3; it had identical performance results between versions with adherence rates of 83%.

<table>
<thead>
<tr>
<th>Measure Description</th>
<th>2015 version</th>
<th>2017 version</th>
<th>Change from 2015 to 2017</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Appropriate Aspirin Use” Composite (NQF0068)</td>
<td>978/1390 (75.81%)</td>
<td>N/A</td>
<td>950/1171 (81.13%)</td>
</tr>
<tr>
<td>“NQF0068 - AMI only”</td>
<td>115/140 (82.14%)</td>
<td>N/A</td>
<td>139/164 (84.76%)</td>
</tr>
<tr>
<td>“NQF0068 - IVD only”</td>
<td>942/1245 (75.66%)</td>
<td>N/A</td>
<td>907/1113 (81.15%)</td>
</tr>
<tr>
<td>“Cholesterol Management by Statin Therapy” (CMS347)</td>
<td>1518/2062 (73.62%)</td>
<td>113</td>
<td>1524/2066 (73.77%)</td>
</tr>
<tr>
<td>“CMS347 - AMI only”</td>
<td>118/136 (86.13%)</td>
<td>4</td>
<td>145/173 (83.82%)</td>
</tr>
<tr>
<td>“CMS347 - IVD only”</td>
<td>968/1215 (79.67%)</td>
<td>30</td>
<td>966/1204 (80.23%)</td>
</tr>
<tr>
<td>“CMS347 - Diabetes only”</td>
<td>501/741 (67.96%)</td>
<td>93</td>
<td>506/746 (67.89%)</td>
</tr>
<tr>
<td>“Diabetes A1C Control” (NQF0059)</td>
<td>1030/1438 (71.63%)</td>
<td>N/A</td>
<td>1040/1449 (71.77%)</td>
</tr>
</tbody>
</table>

Table 3. CQM performance that changed between 2015 and 2017 versions by composite measures & sub-concepts
Figure 1 shows the amount of overlap, the number of newly included patients, and the number of dropped patients between the 2015 and 2017 versions of the Aspirin composite measure. The 2017 version of the Aspirin denominator included 13 (1.1%) new patients, while dropping 132 (10.2%) patients included in the 2015 version. The 2017 version of the Cholesterol measure included 14 (<1%) new patients while dropping 10 (<1%) patients from the 2015 cohort. For the Diabetes measure, 11 (<1%) new patients were included in version 2017; 0 patients were dropped from the denominator.

Figure 1. Overlap of patients included in denominators between 2015 and 2017 versions of composite Aspirin CQM

The differences in the distinct patients between the 2015 and 2017 versions’ denominators are attributed to the codes in one version’s value set that are not included in that of the other (in addition to the Aspirin measure’s new exclusion criteria). Tables 4a and 4b show some of the distinct ICD-9-CM/ICD-10-CM codes between versions and their prevalence in the composite measure results, and the prevalence in newly added/dropped patients. For the AMI concept, one newly added ICD-10-CM code (I25.2) accounted for 2.58% of all patients included in the 2017 Aspirin composite denominator, and 92% of the newly added patients in version 2017. In the Cholesterol measure, the code details and prevalence in measures.

<table>
<thead>
<tr>
<th>Measure</th>
<th>Value Set Name</th>
<th>Taxonomy</th>
<th>Code</th>
<th>Prevalence in 2015 Composite Measure</th>
<th>Prevalence of newly added patients in 2017 (N)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMI</td>
<td>“Appropriate Aspirin Use” (NQF0068)</td>
<td>ICD-10-CM</td>
<td>125.2</td>
<td>2.58%</td>
<td>92.3% (12)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ICD-9-CM</td>
<td>412</td>
<td>2.54%</td>
<td>92.3% (12)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>410.70</td>
<td>4.07%</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>410.90</td>
<td>3.60%</td>
<td>-</td>
</tr>
<tr>
<td>AMI</td>
<td>“Cholesterol Management by Statin Therapy” (CMS347)</td>
<td>ICD-10-CM</td>
<td>125.2</td>
<td>1.51%</td>
<td>71.4% (10)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ICD-9-CM</td>
<td>412</td>
<td>1.51%</td>
<td>71.4% (10)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>410.70</td>
<td>2.41%</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>410.90</td>
<td>2.11%</td>
<td>-</td>
</tr>
<tr>
<td>Diabetes</td>
<td>ICD-9-CM</td>
<td></td>
<td>250.40</td>
<td>1.36%</td>
<td>14.3% (2)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>257.21</td>
<td>1.00%</td>
<td>7.1 (2)%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Measure</th>
<th>Value Set Name</th>
<th>Taxonomy</th>
<th>Code</th>
<th>Prevalence in 2015 Composite Measure</th>
<th>Prevalence of dropped patients in 2017 (N)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMI</td>
<td>“Appropriate Aspirin Use” (NQF0068)</td>
<td>ICD-10-CM</td>
<td>120.1</td>
<td>0.92%</td>
<td>6.92% (11)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ICD-9-CM</td>
<td>410.1</td>
<td>1.00%</td>
<td>7.28% (11)</td>
</tr>
<tr>
<td>AMI</td>
<td>“Cholesterol Management by Statin Therapy” (CMS347)</td>
<td>ICD-10-CM</td>
<td>160.20</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ICD-9-CM</td>
<td>180.21</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ICD-9-CM</td>
<td>180.22</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Diabetes</td>
<td>ICD-10-CM</td>
<td>ICD-9-CM</td>
<td>187.89</td>
<td>0.15%</td>
<td>-</td>
</tr>
<tr>
<td>Diabetes</td>
<td>ICD-10-CM</td>
<td>ICD-9-CM</td>
<td>120.1</td>
<td>0.58%</td>
<td>100% (10)</td>
</tr>
<tr>
<td>Diabetes</td>
<td>ICD-10-CM</td>
<td>ICD-9-CM</td>
<td>413.1</td>
<td>0.59%</td>
<td>-</td>
</tr>
</tbody>
</table>

Tables 4a & 4b. Distinct ICD-9-CM and ICD-10-CM code details and prevalence in measures
the same ICD-10-CM code for AMI, I25.2, accounted for 1.51% of the overall 2017 denominator, and 71% of the newly added patients in version 2017. For the diabetes measure, two newly added ICD-9-CM codes (250.4, 357.2) made up about 64% of the newly included patients in version 2017, and almost 11% of the composite denominator.

Finally, the three numerator value sets with code changes (“Aspirin and Other Antiplatelets”, “Beta Blocker Therapy”, “Beta Blocker Therapy Ingredient”), and the exclusions value set with code changes (“Anticoagulant Medications”) all represented drugs, and therefore solely consisted of codes from the RXNORM hierarchy. Most of the changes to these RXNORM concepts between 2015 and 2017 versions involved the addition of Semantic Clinical Drugs (SCDs) in the newer value sets. The Heart Failure CQM, however, had one code deleted in the 2017 version at the ingredient level, removing “Hydrochlorothiazide” from the value set. The Aspirin measure had one Generic Clinical Product (GPCK) in its newly added value set for “anticoagulant medications”. Figure 2 shows the RxNorm hierarchy and the counts of codes at various levels for the Aspirin measure and Heart Failure measure for the concepts “Aspirin” and “Beta Blocking Therapy”. In total, all four measures’ 2017 version value sets consisted of 297 distinct SCDs, but only 69% (187) of these were mapped in the clinics’ EHR from which the data in ICCIS was extracted. Of the mapped SCDs, 57% (106) are used in the clinics, and returned patient IDs in the ICCIS database. The top 20 SCDs, of those mapped and used, account for 53% of patients who are associated with one of the mapped codes.

**Discussion**

**Shifts in Knowledge.** In this study, we examined two types of shifts that may occur as part of regular updates to standardized CQM specifications. The first is a shift in knowledge. For example, the Aspirin measure was updated to adhere to the evidence-based guideline to exclude patients who are taking anticoagulants from aspirin therapy. We identified substantial differences between the 2015 and 2017 implementations of the composite Aspirin measure: the performance percent increased by 5.32%, 132 (10.2%) of the patients included in 2015 were dropped, and 13 (1.1%) new patients were included in 2017. As CQM definitions are updated to reflect new evidence-based guidelines, these types of changes in global performance might be expected if the changes to logic are significant. Currently, in order to know whether or not a CQM has been impacted by a knowledge shift, CQM implementers must download the HQMF from CMS’ eCQM Library and open the human-readable or encoded file to identify changes. It would be helpful if CMS or measure developers published an overview highlighting any updated logic in new versions, so that implementations could be easily identified for editing. CMS does publish “Technical Release Notes” at regular intervals announcing updates; however, processing this file requires downloading a PDF, searching for the measure(s) in question, and interpreting all of the listed changes to identify those to measure logic. An emerging Health Level Seven International standard, which might help with electronic processing of these logic updates, is the Clinical Quality Language (CQL), a new specification that focuses on a common model for representing expression logic for CQMs and Clinical Decision Support. According to CMS’ eCQI Resource Center, CQL will be used in all HQMFs in the future, will replace the Quality Data Model, and is intended to reduce the burden on implementers for consuming measure artifacts. System vendors and measure implementers need to evolve to adhere to this standard, thereby reducing variation caused by knowledge shifts in older and newer versions of CQMs. The eCQI proposes to officially publish the transition to CQL in spring of 2018, with CQL measure development launching in the fall of 2017.
Ideally, these types of logic changes could be made automatically to all CQM implementations; however, this is currently difficult due to the various networks, communication channels, set of transformations, and different organizations of clinical data that currently exist. Therefore, moving away from local measure implementations and towards a more centralized measure calculation system might solve these problems. Clinics currently have access to CQM calculation registries and some companies offer services at what seems to be an affordable price for clinics. As new quality-based reimbursements from the QPP/MIPs are rolled out, we will learn whether the net reimbursements received will enable a sustainable business model for practices, CQM registry companies, and for the healthcare system at large. As we wait for new data standards such as CQL and for a potential transition to centralized measure calculations, an interim solution would be to enhance workforce training in Health IT around implementing measure logic and emphasizing a process and timeline for updating to new versions when they are released.

*Shifts in Taxonomies.* The second type of shift we observed in the CQM update process is a shift in taxonomies. We identified changes in the inclusion/exclusion of codes at various levels of the ICD-9-CM/ICD-10-CM, SNOMED CT, and RxNorm hierarchies. When the CQMs were divided into aggregate clinical concepts, we identified differences in the number of patients included for the clinical concept “acute myocardial infarction”. There were 38 (28%) newly included patients in the 2017 AMI-only implementation of the Cholesterol measure, and 35 (25%) newly included patients in the 2017 AMI-only implementation of the Aspirin measure. In fact, one ICD-10-CM and one ICD-9-CM (both describe “old myocardial infarction”) accounted for over 70% of the newly included patients in the 2017 version of the Cholesterol measure, and over 90% of the new patients included in version 2017 of the Aspirin CQM. Interestingly, the 2017 version of the AMI-only Cholesterol measure decreased in performance rate by 2.13% using the updated AMI code group. The 2017 version of the standalone AMI Aspirin measure only increased in performance by 2.62%, while the 2017 version of the Aspirin composite and the 2017 version of the IVD-only Aspirin measures both increased in rate by greater than 5% between versions. Based on these performance changes alone, one might wonder if the new AMI code group (and the code for “old myocardial infarction” in particular) may be too broad and might include patients not appropriate for statin and aspirin therapy. From a clinical standpoint, however, including old AMI seems appropriate in that having a heart attack in the past does relate to increased risk of subsequent events. Therefore, patients with “old AMI” may represent a care gap for the use of Aspirin for these two clinics. The process for making the choice to include/exclude new codes, and the methods for evaluating impact in the decision-making process, are unknown. Across all four measures, no logic was provided about the decisions made to include/exclude codes. If value set developers provided rationale for the inclusion of the new AMI codes, it might help with interpreting these types of performance changes in the AMI-only implementations, and determining whether or not the changes in patient populations are appropriate. In particular, we found it was difficult to traverse the various taxonomies (especially SNOMED CT and RxNorm) to understand the choice of codes at different levels of the hierarchies. This study shows that the consequence of these choices is that they may impact the number of patients included/excluded in a measure’s denominator, and could decrease (or increase) CQM performance rates. Measure developers should examine these potential implementation changes against EHR data before making changes to value set specifications to help understand the impact of taxonomy drift. One way to potentially reduce the variation on CQM data caused by taxonomy drift is to promote the use of application programming interfaces (APIs) to the VSAC which allow for automated downloads of value set expansions. The development of vendor-specific APIs between specific system vendors and the VSAC is needed to make automated terminology updates possible.

*Impact on population health and improving quality.* Compared to the Aspirin measure, there are minor differences in the CQM performance results between the 2015 and 2017 versions of the composite Cholesterol, Diabetes, and Heart Failure measures. The Heart Failure measure was identical between versions, and the Cholesterol and Diabetes measures had very similar results in version 2015 and version 2017. However, the new and old versions of the Cholesterol and Diabetes measures included distinct patients in version 2017 who were not included in version 2015, and the Cholesterol measure included patients in version 2015 who were not included in version 2017. Because CQMs intend to promote evidence-based clinical processes, variations in CQM data caused by having antiquated implementations may impact the ability of clinicians to assess care and improve quality. Jean-Jacques et al showed that health information technology-supported quality improvement (QI) initiatives can decrease disparities for some chronic disease management and preventive measures; and as technical assistance providers, we observe data-driven QI efforts that heavily rely on patient-level data (i.e., lists) generated by CQM reports. If clinicians rely on old implementations of CQMs, then they may have lists with patients intended to be excluded from a measure, and may therefore, target inappropriate patients for therapies, such as recommending aspirin use for someone at high-risk for a fatal bleeding event. Furthermore, their lists will not include the newly added patients who may need certain therapies to improve outcomes. Assuming that the version 2017 specifications of a measure improved to represent “perfect” inclusion, then every newly included patient can be thought of as needing the evidence-based therapy (such as aspirin)
in order to avoid bad outcomes. Under the same assumption, every dropped patient between versions of a measure can be thought of as avoiding potential harm caused by the promoted therapy. For aspirin use, Number-Needed-to-Treat (NNT) statistics show that of patients with known cardiovascular risk who took aspirin, 1.3% were helped by preventing a non-fatal heart attack, and 0.25% were harmed by a major bleeding event. In this study, 121 (92%) of the patients dropped in the 2017 version of the Aspirin measure were also taking an anticoagulant medication, so the Number-Needed-to-Harm (NNH) statistic for this subset of patients is likely much higher, and for these clinics, 1 to 2 people may have been harmed if the old definition persisted, as Hansen et. al showed that patients with combinations of aspirin, warfarin, and clopidogrel are associated with up to a three-fold higher risk of bleeding for patients on dual therapy and triple therapy. With statin therapy, 1 in 21 people have a repeat heart attack, stroke or death avoided, so even 10 missed people have significant risk of events. Similarly, 10% were harmed by muscle damage or pain, or ~1 of the 14 inappropriately included. Even in this small study, failure to include or exclude patients could have led to real harm. With CQM implementation and QI infrastructure increasing, the problem of having, and using, antiquated CQM versions could have significant potential negative impact on population health by not avoiding events, and avoiding harms for patients.

Limitations. Another type of shift that may impact various versions of CQM implementations can be categorized as “usage shifts”, since the use of specific codes by clinicians might change over time. In this study, we did not aim to look at usage shifts. Despite the fact that the use of codes might have changed between 2015 and 2017, we still sought to understand the differences caused by knowledge shifts and taxonomy shifts by implementing the 2015 and 2017 versions of the measures. Another limitation is that this study only includes data for CQMs implemented against data for two primary care clinics in Oregon. Future analysis could include data from several other clinics, potentially from other parts of the country, to further assess the differences in the performance rates, number of patients included, and distinct patients. Additionally, we could have tested whether each difference between CQMs versions was statistically different. But, the primary goal of this paper was to show the differences themselves when CQMs fail to update, and whether these differences were large or small. For the case of the Aspirin measure, a chi-square test does show a significant difference in performance between versions, and this difference may be important clinically.

Conclusion

We discovered changes in the official update process for four standardized CQMs. Between older and newer versions, CQM definitions can shift in logic, and value sets can change in their inclusion/exclusion of certain codes to define clinical concepts. Both of these changes can result in differences in CQM performance results and impact the patients included. Without more integrated updates to the latest CQM versions, comparability of CQMs for payment programs and utility of CQM data for targeted quality improvement may be limited. Therefore, we need CQM implementations with standard representations, maintained by automated updates. To help minimize variations in CQM implementations, we suggest that system vendors, payers, policymakers, informaticians, measure developers/authors, and developers should work on: (1) validating new data interoperability standards for representing CQM logic to facilitate standardized implementations; (2) assisting measure developers with making/explaining choices at the various hierarchies, and validating value sets against extant EHR data sets prior to dissemination; (3) supporting a sustainable model to pragmatically provide centralized CQM calculation; (4) developing vendor-specific API queries to the VSAC and/or enhanced VSAC API documentation to facilitate automated value set updates; and, (5) enhancing the Health IT workforce to help achieve more standard and timely CQM implementations and updates.

References


18. Cholan R, Weiskopf, NG, Rhoton, DR. From concepts and codes to healthcare quality measurement: understanding variations in value set vocabularies for a statin therapy clinical quality measure. eGEMs. Accepted.


23. What is Clinical Quality Language (CQL)? eCQI Resource Center: Centers for Medicare & Medicaid Services (CMS) and the Office of the National Coordinator for Health Information Technology (ONC); U.S. Department of Health and Human Services; [updated 2017]. Available from: https://ecqi.healthit.gov/cql.


Doodle Health: A Crowdsourcing Game for the Co-design and Testing of Pictographs to Reduce Disparities in Healthcare Communication

Carrie Christensen¹, Doug Redd⁶, Erica Lake², Jean P. Shipman², Heather Aiono³, Roger Altizer, Jr⁷, Bruce E. Bray¹,⁵, Qing T. Zeng⁶

¹Department of Biomedical Informatics, University of Utah, Salt Lake City, UT
²Eccles Health Sciences Library, University of Utah, Salt Lake City, UT
³Genetic Science Learning Center, University of Utah, Salt Lake City, UT
⁴Utah Poison Control Center, Skaggs Pharmacy Institute, Salt Lake City, UT
⁵Division of Cardiovascular Medicine, Department of Internal Medicine, School of Medicine, University of Utah, Salt Lake City, UT
⁶Biomedical Informatics Center, George Washington University, Washington, DC
⁷Entertainment Arts and Engineering Program, University of Utah

Abstract

Supplementing patient education content with pictographs can improve the comprehension and recall of information, especially patients with low health literacy. Pictograph design and testing, however, are costly and time consuming. We created a Web-based game, Doodle Health, for crowdsourcing the drawing and validation of pictographs. The objective of this pilot study was to test the usability of the game and its appeal to healthcare consumers. The chief purpose of the game is to involve a diverse population in the co-design and evaluation of pictographs.

We conducted a community-based focus group to inform the game design. Game designers, health sciences librarians, informatics researchers, clinicians, and community members participated in two Design Box meetings. The results of the meetings were used to create the Doodle Health crowdsourcing game. The game was presented and tested at two public fairs.

Initial testing indicates crowdsourcing is a promising approach to pictograph development and testing for relevancy and comprehension. Over 596 drawings were collected and 1,758 guesses were performed to date with 70-90% accuracies, which are satisfactorily high.

Introduction

Patients, especially those with limited health literacy skills, often experience difficulty understanding health information¹,². Some health information can be represented efficiently using graphics that are easier to understand. In consumer health informatics research, pictographs have been used to improve patient comprehension and recall. In their review of the literature regarding the role pictures play in improving health communication, Houts et al. found that “pictures closely linked to written or spoken text can, when compared to text alone, markedly increase attention to and recall of health education information”¹.

We have conducted several studies related to the development and testing of pictographs to be used in patient instructions³⁵. Our findings confirmed the value of enhancing patient education materials with pictographs for health communication. At the same time, generating informative and comprehensible pictographs for a diverse population is a time consuming and costly process.

People from diverse cultural backgrounds have different communication styles and standards. Some health topics or
images that are part of the vernacular for one group of people may be offensive or culturally taboo for others. Individuals can be challenged to interpret pictures from various cultural and educational backgrounds. The symbol for *prescription*, commonly seen on medication labels, for example, was taken to mean *rubbish bin* by a low-literate South African population with limited English proficiency. A library of high quality, culturally appropriate pictures could address these differences.

The Doodle Health crowdsourcing game was designed to engage a diverse population, including under-represented, vulnerable, and disadvantaged populations, to elicit and test health-related pictures in an effort to build this library. By engaging under-represented minorities in the co-design and evaluation of pictographs, thereby getting their active involvement in the production of the health education material intended for their use, we hope to increase the likelihood that the information they receive will be relevant and meaningful.

**Methods**

**Focus Group**
At the beginning of the design process, we conducted a focus group study to assess the health information needs of group members from under-represented communities, and to obtain their thoughts regarding development of the crowdsourcing game. We recruited 11 participants from a partnership representing five diverse demographic communities residing in the Salt Lake Valley and scientific researchers from the University of Utah and the Utah Department of Health.

The focus group demographics included African Americans, refugees from Africa, Hispanic members, Pacific Islanders, and Native Americans. The participants ranged from 30 to 69 years of age. The majority had acquired education beyond the 12th grade. By race and ethnicity, two American Indians participated (18.1%), as well as five Black or African Americans (45.4%), and two Pacific Islanders (18.1%). English was the primary language spoken by the majority (72.7%), while one person reported knowing English and Tongan equally well (9%); there were two native Spanish speakers (18.1%), and one person told the study team that “African” was their first language.

The two-hour focus group discussion was transcribed and qualitatively analyzed. The focus group helped us identify points of cultural sensitivity. For instance, health-related images or topics that are part of the vernacular for some groups may be offensive or culturally taboo for others. In a healthcare setting, topics of sexuality, conception, and reproduction must be approached delicately.

The group expressed their interest in long-term community health. They wanted to prevent common diseases in their populations, namely heart disease and diabetes, and they wanted to promote healthy behavior, diet and lifestyle choices. They noted that socioeconomic (and sociogeographic) factors are crucial to consider when tailoring health education materials, and they desired information for their communities about the health insurance market and other financial resources to help them overcome barriers to receiving quality healthcare.

We consulted with community representatives after the focus group concluded at regular intervals throughout the game development process. Researchers attended monthly meetings with them in order to report on study progress and obtain feedback.

**The Design Box**
Informed by the results of our focus group, we conducted two iterative Design Box meetings. The Design Box is a conceptual tool for brainstorming and refining game design ideas. The four constraints, presented as four sides of a square, were: Technology, Aesthetics, Audience, and Theory. Technology refers to the digital technical systems of the game. Aesthetics are the formal elements of the game content the game user will interact with, and this category focuses on the emotions they may encounter. The Audience wall of the Design Box is concerned with those who will
play the game. In this case, the target audience is the under-represented communities. Theory refers to the research problem: crowdsourcing pictographs for under-represented populations through gaming.

The Design Box setting encouraged interaction among collaborators. In one Design Box meeting, the Entertainment Arts and Engineering team dialogued with informatics researchers, health sciences librarians, and clinicians. In a separate session, individuals from the five communities recruited for the initial focus group session completed the process. The meeting participants worked together to define the purpose of the crowdsourcing game and to focus the constraints of the Design Box.

**Game Design**
Our game design was based on general game design principles as well as the results from the focus group and Design Box meetings. A key consideration for us was to keep the design simple. In our target audience, many are occasional gamers who require easy-to-use gaming mechanisms. The popularity of games such as Angry Birds™ and Draw Something™ also showed us that simple interface design is attractive to consumers. Short, gradually increasing levels of difficulty, interesting sound effects and colorful graphics are also factors in successful games. The final Doodle Health design is modelled after the Draw Something™ app and the board game Pictionary™, both of which engage participants in the drawing and guessing of pictures. In addition, Pictionary™ is well-known to most of our target audience, and many in the younger audience are also familiar with the Draw Something™ app.

**Implementation**
The Doodle Health game was implemented to run on a variety of mobile devices, standard desktop and laptop computers. The front-end user interface was written in HTML5, CSS3 and Javascript. The Bootstrap framework was selected to make the web pages have a standard appearance across different browsers. jQuery was used to add additional AJAX and dynamic interactivity to the web pages. The canvas element of HTML5 was used for the drawing page, making it so that very little needs to be downloaded to a browser in order for a user to create drawings. PHP scripts were created for the backend of this web application. These PHP scripts used the standard MySQLi library to connect to a MySQL database. Seed image files were uploaded to the server through one of these PHP pages as a stream in order to control what could be uploaded. The result of this architecture (Figure 1) is an adaptable application that can be run on any web browser.

Two versions of the Doodle Health game exist online presently. We tested the version that is open to the public at two local fairs, while the other version was tailored to meet the needs of our individual community partners (Figure 2).

**Game Flow**
For the purposes of the pilot testing, the game first asks the user to self-identify with a community. In the game testing, we listed seven racial/ethnic groups which allows us to link drawings with racial/ethnic backgrounds. Following the advice of our minority community partners, there is no user name or password created and as such players have total anonymity. The player continues by selecting “Draw”, or they can choose to “Guess” a health-related term. For either function there are easy, medium, and hard levels as options.

Upon selecting “Draw”, a player is prompted to draw a word or a phrase (Figure 3). Players working on a touchscreen device, such as a tablet, use a stylus or finger for drawing, while those on a non-touchscreen computer use a mouse to draw. There is a palette of colors and line thicknesses to choose from on the left bottom corner of the drawing screen, and a player is given the options to “Skip” a term, “Clear”, or “Save” and submit their drawing. The Doodle Health user can skip terms up to three times if they don’t feel comfortable drawing them, or if they don’t understand them.
Once a drawing is submitted a progress screen is displayed (Figure 4) showing the points they have earned for their community. The player may continue by selecting “Draw” or “Guess” again.

A player guesses what an image represents by clicking on the corresponding term from a multiple-choice list. The person views a screen showing if their guess was correct and the points earned for their community so far (Figure 5). Because some users make wrong guesses because the pictograph was not easy to understand, they are also given the option of contributing their own drawing for the same term.
To jump start the game, we originally uploaded a set of 50 professionally drawn seed images (from our previous pictograph studies) for guessing. Drawings from the crowdsourcing participants themselves were added to the pictograph library as the game progressed. Because obscenity is a concern when anonymous players are allowed to submit content for others to view, all drawings must be reviewed and approved by a moderator.

Preliminary Testing and Feedback

We tested the Doodle Health crowdsourcing game at two public fairs (Be Well Utah health fair and the Utah State Fair) using tablet computers. A convenience sample was recruited to take part in the study. The participants were asked to play the Doodle Health game for as long as they wanted. Following the play, each participant was asked to fill out an anonymous, Likert-scale survey questionnaire. They were also given the opportunity to provide additional comments about the game. Personal identifying information was not collected from participants and compensation was not provided. Additionally, we kept the Doodle Health game online and invited the fair participants to continue playing the game. We also handed out flyers at the fairs to invite those who could not participate on-site to try out the game at home. To make the testing experience more inviting in this preliminary stage, we did not collect individual demographic information.

Figure 5. Guessing Progress Screen

Results

A combined total of 114 volunteers were recruited via the fairs. We could not ascertain the number of users during the three-month open testing phase. In all, 596 pictures were drawn. 176 were not approved for guessing because they were poorly drawn, unfinished, or because they contained inappropriate content. The data for the images drawn by the community during this time period is shown in Table 1. Some of the drawings we received from crowdsourcing participants were artistic and complex, while others were rendered as basic stick figures. Commonly understood words, such as *ambulance*, yielded similar interpretations from game users (Figure 6). Stranger terms, such as *pinkeye*, showed more variation among drawings (Figure 7).

A total of 1,758 guesses were made. 1,488 matched the correct terms, while 270 were incorrect. Many more guesses were made than drawings, because guessing appears to be simpler and easier for people to do than drawing. The images we included in the game as drawn by community members were largely recognizable by other players. One game tester commented, “I am comfortable guessing drawings and less effective at drawing.” Some of the images, however, were easier for participants to guess than others. The pictures guessed correctly most often were concrete terms like *ambulance*, *loose tooth*, *vegetables*, and *wash hands*. Those frequently missed were conceptually more difficult words and phrases to interpret, such as *fatigue*, *nausea*, *depression*, *mealtime*, and *cool down*. This correlates with our prior research in pictograph development. The number of guesses made by community members is shown in Table 2.

Fair participants answered usability questions. Most people were comfortable with the technology used for the game and they were able to play it without difficulty. The details are shown in Figure 8 and 9.
We received 37 additional comments from fair participants. They were sorted into four main groups and qualitatively analyzed (Table 3.). The fair participants also identified a few software bugs. For instance, it was discovered that a vertical screen orientation was necessary to see all of the buttons on a tablet interface without having to scroll down.

Table 1. Images Drawn by Community

<table>
<thead>
<tr>
<th>Community</th>
<th># of images drawn</th>
<th># of images rejected by m</th>
<th>% of rejected images</th>
<th>% of total images</th>
</tr>
</thead>
<tbody>
<tr>
<td>American Indian/ Alaskan</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0.76%</td>
</tr>
<tr>
<td>Asian</td>
<td>119</td>
<td>31</td>
<td>26.05%</td>
<td>22.67%</td>
</tr>
<tr>
<td>Native Hawaiian/ Pacific</td>
<td>7</td>
<td>5</td>
<td>71.43%</td>
<td>1.33%</td>
</tr>
<tr>
<td>Black or African American</td>
<td>49</td>
<td>14</td>
<td>28.57%</td>
<td>9.33%</td>
</tr>
<tr>
<td>Latino</td>
<td>84</td>
<td>23</td>
<td>27.38%</td>
<td>16.00%</td>
</tr>
<tr>
<td>White</td>
<td>227</td>
<td>90</td>
<td>39.64%</td>
<td>43.24%</td>
</tr>
<tr>
<td>Other</td>
<td>35</td>
<td>11</td>
<td>31.44%</td>
<td>6.67%</td>
</tr>
<tr>
<td>Total</td>
<td>525</td>
<td>174</td>
<td>33.14%</td>
<td>100%</td>
</tr>
</tbody>
</table>

Figure 6. Community Drawings for the Term *Ambulance*
Table 2. Doodle Health Guesses by Community

<table>
<thead>
<tr>
<th>Community</th>
<th>Correct Guesses</th>
<th>Incorrect guesses</th>
<th>Total # of guesses made</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>American Indian/Alaska Native</td>
<td>23</td>
<td>9</td>
<td>32 (1.82%)</td>
<td>71.88%</td>
</tr>
<tr>
<td>Asian</td>
<td>212</td>
<td>46</td>
<td>258 (14.68%)</td>
<td>82.17%</td>
</tr>
<tr>
<td>Native Hawaiian/Pacific Islander</td>
<td>43</td>
<td>8</td>
<td>51 (2.90%)</td>
<td>84.31%</td>
</tr>
<tr>
<td>Black or African American</td>
<td>46</td>
<td>5</td>
<td>51 (2.90%)</td>
<td>90.20%</td>
</tr>
<tr>
<td>Latino</td>
<td>221</td>
<td>41</td>
<td>262 (14.90%)</td>
<td>84.35%</td>
</tr>
<tr>
<td>White</td>
<td>878</td>
<td>149</td>
<td>1027 (58.42%)</td>
<td>85.49%</td>
</tr>
<tr>
<td>Other</td>
<td>65</td>
<td>12</td>
<td>77 (4.38%)</td>
<td>84.42%</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>1488</strong></td>
<td><strong>270</strong></td>
<td><strong>1758 (100%)</strong></td>
<td><strong>84.64%</strong></td>
</tr>
</tbody>
</table>
Figure 8. Game Testing Results from the Be Well Utah Health Fair (n = 79)

Figure 9. Game Testing Results from the Utah State Fair (n = 35)
Table 3. Types of Comments Received from Game Testers

<table>
<thead>
<tr>
<th>Category</th>
<th>Total # of comments</th>
<th>% of Total (n=37)</th>
<th>Sample comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Praise for the game</td>
<td>23</td>
<td>62.2%</td>
<td><em>It was very interesting and fun.</em></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><em>I think this is a great tool to use, both for non-native English speakers and for kids!</em></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><em>[…] the interface was engaging and easy to use! TECHNOLOGY!!</em></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><em>The pictures were very creative.</em></td>
</tr>
<tr>
<td>2. Technical complaints</td>
<td>8</td>
<td>21.6%</td>
<td><em>Not being familiar with a scrolling touch screen made me a little uneasy.</em></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><em>Would like the [drop down] options not to block the picture/image.</em></td>
</tr>
<tr>
<td>3. Suggested health topics</td>
<td>3</td>
<td>8.1%</td>
<td><em>I think it is a good way to invite people to do more activities in our daily life.</em></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><em>Topics to help us eat healthy</em></td>
</tr>
<tr>
<td>4. Disinterest</td>
<td>3</td>
<td>8.1%</td>
<td><em>Not interested.</em></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><em>I don’t like games – no time for it.</em></td>
</tr>
</tbody>
</table>

Discussion

We developed the Doodle Health game to crowdsource the pictograph development and testing tasks while providing entertainment value to players. Without any compensation, and only being marketed at two local public events, the game was able to collect 596 drawings over a three-month testing period. The drawings were evaluated using the same game 1,758 times. Given the 70 to 90% of the guess accuracies, the quality of the drawing and guesses are satisfactorily high.

Based on our prior experience, traditional pictograph creation and testing is a very time-consuming process. Even experienced graphic designers and medical illustrators are limited by their own cultural biases. Using a crowdsourcing game to engage the larger community allows us to rapidly gather and test pictures from diverse backgrounds. Through crowdsourcing, under-represented healthcare consumers are able to contribute to the development of culturally-sensitive and understandable pictographs for health communication.

The results we have gathered show that the Doodle Health game is a promising approach to crowdsourcing for pictograph development and validation. The feedback we collected from those who tested the game was quite positive. Most participants found the game engaging and easy-to-use. Most felt they would like to play it again. These ratings are consistent with the enthusiastic comments we received from the game testing participants.

We hope this game may indeed result in a more meaningful resource library of easily comprehended pictographs for use in health communication materials. The online library of pictographs developed through the crowdsourcing game will be freely available to clinicians and other health care workers to supplement text in patient discharge instructions, for example, and other health-related education materials. The ultimate goal of this project is to have a positive impact on clinical practices by facilitating and enhancing communication between health care providers and consumers.

One limitation of the Doodle Health game is that it is not suitable for medical concepts or content that is totally unfamiliar to an audience. For instance, the ability to draw and/or guess the word *defibrillator* requires specialized knowledge. Moreover, it is unlikely that we will be able to incorporate publicly-drawn images directly into the pictograph library, as many people with high representation may not be expert illustrators. We can identify relevant concepts and cultural sensitivities with the game, and we can develop pictographs through iterative cycles of game play and data analysis.
We asked the crowdsourcing participants in this study to self-identify their community affiliation, because African Americans, American Indians/Alaska Natives, Asian Americans, Hispanics, and Native Hawaiian/Pacific Islander populations carry a disproportionate burden of disease, premature death, and disability. They also receive poorer health care overall, as compared to other ethnic or racial groups \(^{10}\). On the other hand, the needs of each community vary greatly. A Spanish version of the game is needed, for instance, for non-English speaking Hispanics.

To minimize the burden on our game testers and maximize the number of participants, we did not collect demographic information such as age and gender. At the advice of our community partners, we only asked players to identify one community they would like to play for (to earn points for the community), because some people are turned off by the question of race/ethnicity. At the fair, our observation is that the participants were reasonably diverse in terms of race, ethnicity, gender, and age. In future formal trials, we will offer voluntary demographic surveys for crowd sourcing participants.

We are trying out novel applications for the Doodle Health game and experimenting with various promotional ideas and incentives. We have installed a gallery, to present drawings and award crowdsourcing participants, and we are exploring social networking as a tool for marketing the game within particular communities.

**Acknowledgements**
The authors of this paper would like to thank the National Institutes of Health (NIH) for grant support of this project (G08 LM011546-03 Graphics to Enhance Health Education Materials for Underrepresented Populations).

**Conclusions**
*We created a Web-based game, Doodle Health, for crowdsourcing the drawing and validation of pictographs. Initial testing collected a large number of drawings and guesses, with positive user feedbacks, demonstrating the approach to be feasible and promising.*

**References**
Evaluation of a systematic methodology to detect in near real-time performance changes during electronic health record system implementations: a longitudinal study

Tiago K Colicchio¹, MBA, MS, Guilherme Del Fioli¹, MD, PhD,
Greg J Stoddard², MS, Scott P Narus³, PhD
¹Biomedical Informatics, University of Utah, Salt Lake City, UT; ²Department of Internal Medicine, University of Utah, Salt Lake City, UT; ³Medical Informatics, Intermountain Healthcare, Salt Lake City, UT

Abstract

Introduction. Although Electronic Health Record (EHR) adoption has increased in the U.S., our understanding of how it affects health care organizations is still limited. Current literature has produced mixed-results due to the use of simple, non-standardized measurements and poor research designs.

Methods. We propose the use of a systematic methodology that combines measures of quality, productivity and safety processes, tracked over time using an interrupted time-series design with multiple control sites.

Results. Our methodology successfully detected performance changes during an EHR implementation on 17 (77%) outcomes, including a significant increase in Emergency Department length of stay immediately after go live by 0.19 hours [95%CI (0.12, 0.27), p<0.001], and an improvement in time to complete radiology tests, which significantly decreased per month by 0.19 minutes [95%CI (-0.26, -0.12), p<0.001].

Conclusion. The proposed methodology was able to detect several changes immediately after an EHR implementation and over time. The method is a promising and robust approach to assessing the impact of EHR implementations on a wide range of health care quality, productivity, and safety care processes.

Introduction

Health Information Technology (health IT) Adoption in the U.S. Previous studies have demonstrated advantages associated with Electronic Health Record (EHR) systems adoption, such as improving quality and productivity outcomes, and decreasing errors and health care cost1-4. Such improvements have attracted attention from government agencies and policy makers5. In 2004, the U.S. federal government issued an executive order to provide financial incentives to increase health IT adoption in the U.S.6, and in 2009, the Health Information Technology for Economic and Clinical Health (HITECH) act was signed into law establishing the Meaningful Use program7. The program contributed to increasing EHR adoption among U.S. care delivery systems, and, as a result, the literature in this area is also rapidly increasing8. Several systematic reviews have analyzed studies on the impact of health IT. The studies report on different outcomes, such as quality, productivity and safety; review different health IT tools, including multifunctional EHRs; cover both ambulatory and non-ambulatory care settings; and include U.S. and non-U.S. health care organizations8-11. Such reviews found that health IT adoption studies more frequently report positive outcomes associated with EHR adoption and use8, 11, and that the Meaningful Use program has contributed to the increased positive outcomes reported11. However, despite the increasing number of health IT evaluations and positive findings, the same reviews also found several studies that produced mixed or negative results, leaving unanswered questions on the full impact of IT interventions in health care8, 11.

The problem. The lack of a more complete understanding of health IT impact has several contributing factors. Studies available frequently report simple, non-standardized measurements, creating obstacles to the comparison of health IT outcomes across different studies12. In addition, quantitative hypothesis-testing studies more frequently assess the impact of their interventions using pre-post comparisons8. However, similar to large-scale changes observed in other industries13, an EHR implementation is a complex, ongoing process that introduces sociotechnical changes that iteratively evolve over time14, exposing end-users to a learning curve of up to two years15. Previous research suggests that when an intervention has a longitudinal effect – which is the case for EHR implementations – interrupted time-series design is the most suitable design to avoid biases caused by time-sensitive variations not detected.
by simple pre-post statistical comparisons. However, an effective interrupted time-series design requires specific components. As demonstrated by Zhang et al., to achieve 80% power with an effect size = 0.5, interrupted time-series studies must have at least 24 time periods for equal pre-post time periods, and at least 30 time periods if the post-intervention period is smaller than pre-intervention. Other studies suggest the use of a control site whenever possible, to overcome the confounding effect of other events not detected by assessing isolated intervention sites. Despite the recommendations, a recent systematic review by Jones et al. reported that only 27 (11%) of 236 health IT adoption studies used a time-series design; out of those, only 8 include 24 data instances or more, and only 2 have control sites. Therefore, a more robust methodology is necessary to increase our understanding of the full impact of health IT interventions. The method should use a robust longitudinal analysis to more effectively monitor the impact of health IT interventions and to identify unexpected effects introduced both during the transition and after the new system has been stabilized, seasonal effects, and time to recover to baseline performance. In addition to improved design, studies should include a compressive set of outcome measurements, covering quality, productivity and safety.

**Potential Solution: Systematic Methodology for Near Real-Time Monitoring of EHR Implementations.** We have previously developed an inventory of measures with potential to detect the impact of EHR implementations in near real-time using data that are routinely captured in electronic format. These measures provide a comprehensive coverage of quality, productivity and safety processes likely impacted by EHR implementations. Measures were identified through semi-structured interviews with health care leaders, and input from informatics experts. By combining these measures with other measures commonly reported in the literature (as detected by our previous study), our final inventory included 102 outcome measures, and a taxonomy to classify these outcomes. In the present study, we use a subset of measures from the larger inventory to detect time-sensitive changes in a group of health care settings implementing a commercial EHR system using an interrupted time-series design with multiple control sites. Our aim is to verify if our proposed methodology detects and quantifies what changes are introduced to the performance of the care delivery organization while it is transitioning from long-used and stable homegrown systems to a commercial EHR.

**Methods**

**Setting and Study Design.** We performed this study at a medium-size (140 beds) non-academic hospital and five ambulatory care clinics implementing a commercial EHR system, and compared them with a medium-size (243 beds) non-academic hospital and five ambulatory care clinics that had not implemented the new EHR at the time of this study. All settings are part of Intermountain Healthcare, a not-for-profit integrated care delivery system of 22 hospitals and over 185 ambulatory care clinics covering the entire state of Utah and southern Idaho. We used an interrupted time-series design, analyzing monthly data for the period of February 1st, 2013 to January 31st, 2017. The baseline period is represented by data instances prior to the new EHR go live, and was compared against the intervention period, which includes data after go live. Intermountain Healthcare’s institutional review board approved the study under protocol 1040351.

**Description of intervention.** Intermountain Healthcare has extensive experience developing and operating EHR systems, including the Health Evaluation through Logical Processing (HELP) system, first introduced at the LDS Hospital in Salt Lake City in 1975. An enterprise, longitudinal EHR was later developed and first deployed in 1994. Intermountain is replacing its legacy systems with the commercial, multifunctional Millennium EHR developed by Cerner Corporation, Kansas City, MO, U.S. Given the size of the Intermountain care delivery system, the implementation of the new EHR follows a staggered schedule with multiple phases, each phase comprising a group of hospitals and clinics from the same geographical area. While the enterprise implementation will be phased, the introduction of the new EHR in each region will use a “big bang” strategy, replacing all legacy systems at once within that region. Such a staggered approach
allowed for the inclusion of one hospital and five ambulatory care clinics from the first region of the implementation, which started using the new EHR in February 2015, and one hospital and five ambulatory care clinics from another region that has not started its implementation as a comparison control.

**Outcome Measures.** We collected data for 22 measures (Tables 1 and 2) of quality, productivity and safety outcomes. These measures were extracted from a rigorous and comprehensive inventory of outcome measures likely impacted by EHR implementations, combining measures from the literature\(^1\)\(^2\) and subject-matter experts’ suggestions\(^3\)\(^4\). Data were collected from existing Intermountain analytics reports and from Intermountain’s Enterprise Data Warehouse (EDW). To test the ability of our methodology to detect various performance changes during the new EHR implementation, we collected data for all measures that were available before and after the Cerner EHR go live for both the implementation and control sites. We were not able to include more measures due to incomplete mapping between Millennium’s database and Intermountain’s EDW. The number of data instances before and after go live varied by measure, but all measures had at least 20 instances of data in each phase. All patients eligible for one or more measures were included; no individual patient health information was accessed for this study.

### Table 1. Ambulatory care outcome measures

<table>
<thead>
<tr>
<th>Measure</th>
<th>Description</th>
<th>Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Quality</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
| Hemoglobin A1c control | Rate of diabetes patients with hemoglobin A1c under control | N: diabetes patients with Hemoglobin A1c below 8%  
D: diabetes patients with Hemoglobin A1c measured |
| Medication for asthma  | Rate of asthma patients using appropriate medication   | N: asthma patients who received controller reliever medication  
D: eligible asthma patients |
| **Productivity**       |                                                       |                                                                          |
| Employee movement      | Rate of employees transferred to a different department of facility | N: ambulatory employees transferred to a different work location  
D: total ambulatory employees |
| Employee turnover      | Rate of employee contracts terminated                 | N: ambulatory employees with voluntary contract termination  
D: total ambulatory employees |
| New patients visits    | Rate of new patient visits to ambulatory settings      | N: new patient visits  
D: all patient visits |
| Patient visits         | Number of patient visits                              | Number of patient visits to ambulatory care clinics                      |
| Radiology tests        | Number of orders of imaging tests                     | Number of imaging tests ordered in ambulatory care clinics               |

Abbreviations: N: Numerator; D: Denominator.

### Table 2. Hospital outcome measures

<table>
<thead>
<tr>
<th>Measure</th>
<th>Description</th>
<th>Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Quality</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
| Patient satisfaction  | Rate of patients who gave their hospital a rating of 9 or 10 on a scale from 0 (lowest) to 10 (highest) | N: patients who rated their hospital as 9 or 10  
D: patients who answered satisfaction survey |
| Pressure ulcer rate   | Rate of patients who developed pressure ulcer during hospitalization | N: hospital-acquired pressure ulcer cases  
D: 100 total inpatient discharges |
| Readmission rate      | Rate of heart failure patients readmitted within 30 days | N: unplanned heart failure readmissions  
D: 100 unplanned heart failure patient discharges |
| Sepsis bundle         | Composite measure for sepsis care measured as compliance to all composite items | N: patients with severe sepsis or septic shock in compliance with all items of the sepsis bundle  
D: total sepsis patients |
| Sepsis mortality rate | Rate of patients who died during hospitalization due to severe sepsis or septic shock | N: sepsis patients who died during hospitalization  
D: total sepsis patients |
<p>| <strong>Productivity</strong>      |                                                     |                                                                          |
| ED LOS                | Length of stay of patients in emergency departments | Median length of stay of patients in the emergency department in hours  |
| ED visits             | Number of patient visits to emergency departments   | Number of emergency department visits                                    |
| ED wait time          | Time between patient arrival and seen by provider in emergency departments | Median time between patient check-in and seen by qualified provider in the emergency department |</p>
<table>
<thead>
<tr>
<th>Employee movement</th>
<th>Rate of employees transferred to a different department of facility</th>
<th>N: hospital employees transferred to a different work location</th>
<th>D: total hospital employees</th>
</tr>
</thead>
<tbody>
<tr>
<td>Employee turnover</td>
<td>Rate of employee contracts terminated</td>
<td>N: hospital employees with voluntary contract termination</td>
<td>D: total hospital employees</td>
</tr>
<tr>
<td>Hospitalizations</td>
<td>Number of patients hospitalized</td>
<td>Number of patients hospitalized</td>
<td></td>
</tr>
<tr>
<td>Radiology tests</td>
<td>Number of imaging tests</td>
<td>Number of imaging tests completed</td>
<td></td>
</tr>
<tr>
<td>Time to complete radiology tests</td>
<td>Mean time between radiology test started and completed</td>
<td>Mean time between patient arrival and imaging test completed</td>
<td></td>
</tr>
<tr>
<td>Time to sign radiology tests</td>
<td>Mean time between radiology test completed and report signed</td>
<td>Mean time for issuing imaging test report</td>
<td></td>
</tr>
</tbody>
</table>

Safety

| Safety    | Rate of adverse drug events | N: adverse drug events | D: 1000 inpatient days |

Abbreviations: N: Numerator; D: Denominator.

**Data analysis.** To test the ability of our methodology to detect performance changes during the new EHR implementation, we calculate the effect immediately after the intervention and the pre-post effect over time on the outcome measures. The immediate effect after intervention is the difference between the last predicted value before the introduction of the intervention and the first predicted value after its introduction. Predicted values are automatically generated by the statistical model described below. To calculate the effect over time, we first calculate the monthly change in the period before and in the period after the introduction of the intervention and their difference within each group/setting separately; and then calculate the difference between intervention and control sites on these pre-to-post differences. We combined the five clinics of each region by calculating the arithmetic average of each measure for each month in each group (intervention vs. control). We analyze these monthly data instances using interrupted time series analysis (ITSA) with an ordinary least squares (OLS) model²² with Newey-West standard errors to adjust for autocorrelation, which assumes the error structure to be heteroskedastic and possibly autocorrelated up to some lag²³. After fitting our model, we check if the number of lags included in the model to account for autocorrelation was correctly specified and adjusted accordingly, using the Cumby-Huizenga general test for autocorrelation²⁴. Data analysis was performed using Stata version 14.2 statistical software [College Station, TX: StataCorp LP].

**Results**

From the 22 measures assessed, 17 (77%) measures showed a significant difference in the intervention group when compared to control sites. Nine (41%) measures showed a significant effect immediately after the introduction of the intervention and 14 (61%) measures showed a significant effect over time.

**Ambulatory care measures.** Hemoglobin A1c (Figure 1-A) in the intervention group had a significant immediate effect with a decrease in the proportion of patients under control by 3.03 [95%CI (-4.35, -1.71), p<0.001], and a decrease per month by 0.16 [95%CI (-0.28, -0.05), p=0.006] when compared to the control group. Although the proportion of patients receiving appropriate medication for asthma (Figure 1-B) in the intervention group had a significant immediate effect, decreasing by 25.81 [95%CI (-36.63, -14.99), p<0.001] when compared with control sites, such a difference is attributable to a significant increase in control sites by 28.13 [95%CI (17.65, 38.61), p<0.001].

Both employee movement (Figure 1-C) and employee turnover (Figure 1-D) had a significant effect over time. The proportion of employees transferred to a different department or facility in the intervention group increased per month by 0.11 [95%CI (0.01, 0.21), p<0.05], and the proportion of employees with voluntary turnover increased per month by 0.12 [95%CI (0.34, 0.21), p=0.007] when compared to control sites. Employee movement has an outlier between months 10 and 15 in the pre-intervention period, and although it did not seem to affect the predicted baseline, it may warrant further investigation. Both groups had a
significant immediate effect in the proportion of new patient visits (Figure 1-E) when assessed individually; however, the intervention group decreased by 1.02 more than control sites [95%CI (-1.87, -0.16), p<0.02], and decreased per month by 0.18 more than control sites [95%CI (-0.23, -0.13), p<0.001]. Although the number of patient visits (Figure 1-F) in the intervention group showed a significant immediate effect, increasing by 458.05 [95%CI (364.84, 551.26), p<0.001] when compared to control sites, such a difference is attributable to a significant decrease in the number of visits in control sites by 393.15 [95%CI (-448.22, -338.08), p<0.001]. Both groups had a significant immediate effect in the number of radiology tests completed (Figure 1-G) when assessed individually; however, the intervention group decreased by 61.30 tests more than control sites [95%CI (-99.18, -23.41), p<0.002], and increased per month by 3.45 tests more than control sites [95%CI (0.82, 6.08), p<0.01]. Tables 3 and 4 describe the analysis for immediate effect and pre-post effect over time in ambulatory care settings.

![Figure 1. Ambulatory measures with significant difference between intervention and control groups](image)

**Table 3. Immediate effect of the intervention in ambulatory care clinics**

<table>
<thead>
<tr>
<th>Measure</th>
<th>Intervention Clinics Coefficient (95% CI)</th>
<th>P Value</th>
<th>Control Clinics Coefficient (95% CI)</th>
<th>P Value</th>
<th>Intervention vs. Control Coefficient (95% CI)</th>
<th>P Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hemoglobin A1c control</td>
<td>-3.39 (-4.63, -2.15)</td>
<td>&lt;0.001</td>
<td>-0.36 (-0.81, 0.10)</td>
<td>0.12</td>
<td>-3.03 (-4.35, -1.71)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Medication for asthma</td>
<td>2.32 (-0.39, 5.02)</td>
<td>0.92</td>
<td>28.13 (17.65, 38.61)</td>
<td>&lt;0.001</td>
<td>-25.81 (-36.63, -14.99)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Employee movement</td>
<td>-0.01 (-0.31, 0.28)</td>
<td>0.93</td>
<td>-0.70 (-2.07, 0.67)</td>
<td>0.31</td>
<td>0.69 (-0.72, 2.09)</td>
<td>0.34</td>
</tr>
<tr>
<td>Employee turnover</td>
<td>0.06 (-0.22, 0.34)</td>
<td>0.69</td>
<td>-0.64 (-1.77, 0.50)</td>
<td>0.27</td>
<td>0.70 (-0.48, 1.87)</td>
<td>0.24</td>
</tr>
<tr>
<td>New patient visits</td>
<td>-1.94 (-2.34, -1.54)</td>
<td>&lt;0.001</td>
<td>-0.92 (-1.67, -0.16)</td>
<td>0.01</td>
<td>-1.02 (-1.87, -0.16)</td>
<td>0.02</td>
</tr>
<tr>
<td>Patient visits</td>
<td>64.90 (-10.30, 140.10)</td>
<td>0.09</td>
<td>-393.15 (-448.22, -338.08)</td>
<td>&lt;0.001</td>
<td>458.05 (364.84, 551.26)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Radiology tests</td>
<td>-48.39 (-84.94, -11.83)</td>
<td>0.01</td>
<td>12.91 (2.94, 22.88)</td>
<td>0.01</td>
<td>-61.30 (-99.18, -23.41)</td>
<td>0.002</td>
</tr>
</tbody>
</table>
Table 4. Pre-post effect of the intervention over time in ambulatory care clinics

<table>
<thead>
<tr>
<th>Measure</th>
<th>Intervention Clinics Coefficient (95% CI)</th>
<th>p Value</th>
<th>Control Clinics Coefficient (95% CI)</th>
<th>p Value</th>
<th>Intervention vs. Control Coefficient (95% CI)</th>
<th>p Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hemoglobin A1c control</td>
<td>-0.37 (-0.48, -0.26)</td>
<td>&lt;0.001</td>
<td>-0.21 (-0.24, -0.17)</td>
<td>&lt;0.001</td>
<td>-0.16 (-0.28, -0.05)</td>
<td>0.006</td>
</tr>
<tr>
<td>Medication for asthma</td>
<td>0.55 (0.33, 0.77)</td>
<td>&lt;0.001</td>
<td>0.79 (-0.05, 1.64)</td>
<td>0.07</td>
<td>-0.24 (-1.12, 0.63)</td>
<td>0.58</td>
</tr>
<tr>
<td>Employee movement</td>
<td>0.04 (0.01, 0.07)</td>
<td>0.01</td>
<td>-0.07 (-0.17, 0.03)</td>
<td>0.17</td>
<td>0.11 (0.01, 0.21)</td>
<td>0.05</td>
</tr>
<tr>
<td>Employee turnover</td>
<td>0.09 (0.06, 0.11)</td>
<td>&lt;0.001</td>
<td>-0.03 (-0.12, 0.04)</td>
<td>0.39</td>
<td>0.12 (0.34, 0.21)</td>
<td>0.007</td>
</tr>
<tr>
<td>New patient visits</td>
<td>-0.09 (-0.12, -0.06)</td>
<td>&lt;0.001</td>
<td>0.09 (0.04, 0.13)</td>
<td>&lt;0.001</td>
<td>-0.18 (-0.23, -0.13)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Patient visits</td>
<td>-3.24 (-9.11, 2.64)</td>
<td>0.27</td>
<td>1.74 (-2.05, 5.52)</td>
<td>0.37</td>
<td>-4.97 (-11.96, 2.02)</td>
<td>0.16</td>
</tr>
<tr>
<td>Radiology tests</td>
<td>1.58 (-0.94, 4.10)</td>
<td>0.22</td>
<td>-1.87 (-2.60, -1.14)</td>
<td>&lt;0.001</td>
<td>3.45 (0.82, 6.08)</td>
<td>0.01</td>
</tr>
</tbody>
</table>

Hospital measures. There was no significant difference between the two hospitals in the proportion of
patients satisfied (Figure 2-A) immediately after the go live; however, there was a significant difference
over time with satisfaction rate in the intervention hospital increasing per month by 0.21 [95% CI (0.06,
0.37), p=0.008] when compared to the control site. Compliance to sepsis bundle (Figure 2-B) significantly
increased over time in the intervention group by 4.08 [95% CI (1.97, 6.20), p=0.008] when compared to
the control site; however, the difference is attributable to a significant decrease per month in the control hospital
by 3.44 [95% CI (-4.76, -2.13), p<0.001]. Sepsis mortality rate (Figure 2-C) significantly decreased per
month in the intervention group by 0.80 [95% CI (-1.31, -0.30), p=0.02] when compared to the control site.

ED length of stay (Figure 2-D) significantly increased in the intervention hospital immediately after the go live by 0.19 hours [95% CI (0.12, 0.27), p<0.001] when compared to the control site; however, this
difference was not observed over time because both sites had a similar significant decrease in the post-
intervention period. ED wait time (Figure 2-E) showed no significant immediate effect when the two
hospitals were compared; on the other hand, it showed a significant effect over time with the intervention
group decreasing per month by 0.32 minutes [95% CI (-0.54, -0.11), p=0.004] when compared to the control
site. Both hospitals had a significant effect over time in the proportion of employees transferred to a
different department or facility (Figure 2-F) when assessed individually, and this effect was significantly
higher in the intervention hospital which decreased per month by 0.43 more than the control hospital
[95%CI (-0.06, -0.02), p=0.001]. Hospitalizations (Figure 2-G) showed a significant effect over time in the
intervention group, decreasing per month by 5.82 admissions [95% CI (-8.83, -2.83), p=0.001] when
compared to the control site. The number of radiology tests completed (Figure 2-H) significantly decreased
in the intervention group immediately after the go live by 259.01 tests [95% CI (-464.28, -53.72), p=0.01],
and significantly increased per month by 18.9 tests [95% CI (1.02, 35.15), p=0.04] when compared to the
control site. Time to complete radiology tests (Figure 2-I) improved with a significant decrease in the
intervention group by 2.75 minutes [95% CI (-3.66, -1.86), p<0.001] immediately after the go live, and a
decrease per month by 0.19 minutes [95%CI (-0.26, -0.12), p<0.001] when compared to the control site.

Although the rate of adverse drug events (Figure 2-J) significantly increased in the intervention group by
1.30 [95% CI (0.33, 2.27), p=0.01] immediately after go live, such an effect may be explained by the
somewhat high variability of the outcome, with rate ranging from 0.80% to 8.90%. In addition, it improved
over time with a significant decrease per month by 0.09 [95%CI (-0.16, -0.02), p=0.02]. Tables 5 and 6
describe the analysis for immediate effect and pre-post effect over time in hospital settings.
Figure 2. Hospital measures with significant difference between intervention and control groups

Table 5. Immediate effect of the intervention in the hospital setting

<table>
<thead>
<tr>
<th>Measure</th>
<th>Intervention Hospital Coefficient (95% CI)</th>
<th>p Value</th>
<th>Control Hospital Coefficient (95% CI)</th>
<th>p Value</th>
<th>Intervention vs. Control Coefficient (95% CI)</th>
<th>p Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient satisfaction</td>
<td>0.94 (-0.47, 2.35)</td>
<td>0.26</td>
<td>1.75 (0.36, 3.11)</td>
<td>0.01</td>
<td>-0.81 (-2.77, 1.15)</td>
<td>0.42</td>
</tr>
<tr>
<td>Pressure ulcer rate</td>
<td>-0.20 (-0.27, -0.14)</td>
<td>&lt;0.001</td>
<td>-0.26 (-0.33, -0.18)</td>
<td>&lt;0.001</td>
<td>0.05 (-0.05, 0.16)</td>
<td>0.32</td>
</tr>
<tr>
<td>Readmission rate</td>
<td>-2.84 (-11.69, 6.01)</td>
<td>0.52</td>
<td>-1.50 (-12.57, 9.56)</td>
<td>0.79</td>
<td>-1.33 (-15.50, 12.82)</td>
<td>0.85</td>
</tr>
<tr>
<td>Sepsis bundle</td>
<td>0.38 (-23.29, 24.05)</td>
<td>0.97</td>
<td>-15.14 (-35.96, 5.66)</td>
<td>0.15</td>
<td>15.53 (-15.99, 47.05)</td>
<td>0.33</td>
</tr>
<tr>
<td>Sepsis mortality rate</td>
<td>-3.69 (-8.00, 0.62)</td>
<td>0.09</td>
<td>1.27 (-3.51, 6.07)</td>
<td>0.60</td>
<td>-4.97 (-11.42, 1.47)</td>
<td>0.12</td>
</tr>
<tr>
<td>ED LOS</td>
<td>0.24 (0.17, 0.30)</td>
<td>&lt;0.001</td>
<td>0.04 (0.01, 0.08)</td>
<td>0.02</td>
<td>0.19 (0.12, 0.27)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>ED visits</td>
<td>49.48 (-3.48, 102.44)</td>
<td>0.07</td>
<td>82.88 (37.39, 128.17)</td>
<td>0.001</td>
<td>-33.40 (-103.22, 36.40)</td>
<td>0.34</td>
</tr>
<tr>
<td>ED wait time</td>
<td>4.49 (2.27, 6.71)</td>
<td>&lt;0.001</td>
<td>2.22 (0.56, 3.89)</td>
<td>0.009</td>
<td>2.26 (-0.51, 5.03)</td>
<td>0.11</td>
</tr>
<tr>
<td>Employee movement</td>
<td>-0.06 (-0.31, 0.20)</td>
<td>0.65</td>
<td>-0.28 (-0.48, -0.09)</td>
<td>0.004</td>
<td>0.23 (-0.09, 0.55)</td>
<td>0.15</td>
</tr>
<tr>
<td>Employee turnover</td>
<td>0.09 (-0.07, 0.25)</td>
<td>0.26</td>
<td>0.21 (0.08, 0.34)</td>
<td>0.002</td>
<td>-0.11 (-0.32, 0.09)</td>
<td>0.26</td>
</tr>
<tr>
<td>Measure</td>
<td>Intervention Hospital Coefficient (95% CI)</td>
<td>p Value</td>
<td>Control Hospital Coefficient (95% CI)</td>
<td>p Value</td>
<td>Intervention vs. Control Coefficient (95% CI)</td>
<td>p Value</td>
</tr>
<tr>
<td>------------------------------</td>
<td>-------------------------------------------</td>
<td>---------</td>
<td>--------------------------------------</td>
<td>---------</td>
<td>-----------------------------------------------</td>
<td>---------</td>
</tr>
<tr>
<td>Patient satisfaction</td>
<td>0.14 (0.2, 0.26)</td>
<td>0.03</td>
<td>-0.08 (-0.17, 0.02)</td>
<td>0.12</td>
<td>0.21 (0.06, 0.37)</td>
<td>0.008</td>
</tr>
<tr>
<td>Pressure ulcer rate</td>
<td>0.01 (-0.01, 0.01)</td>
<td>0.69</td>
<td>0.01 (0.01, 0.01)</td>
<td>0.03</td>
<td>-0.01 (-0.01, 0.01)</td>
<td>0.14</td>
</tr>
<tr>
<td>Readmission rate</td>
<td>0.51 (-0.29, 1.32)</td>
<td>0.20</td>
<td>0.43 (-1.00, 1.87)</td>
<td>0.55</td>
<td>0.08 (-1.56, 1.72)</td>
<td>0.92</td>
</tr>
<tr>
<td>Sepsis bundle</td>
<td>0.64 (-1.03, 2.30)</td>
<td>0.45</td>
<td>-3.44 (-4.76, -2.13)</td>
<td>0.001</td>
<td>4.08 (1.97, 6.20)</td>
<td>0.001</td>
</tr>
<tr>
<td>Sepsis mortality rate</td>
<td>-0.43 (-0.73, -0.13)</td>
<td>0.006</td>
<td>0.37 (-0.03, 0.78)</td>
<td>0.07</td>
<td>-0.80 (-1.31, -0.30)</td>
<td>0.02</td>
</tr>
<tr>
<td>ED LOS</td>
<td>-0.02 (-0.03, -0.02)</td>
<td>&lt;0.001</td>
<td>-0.02 (-0.02, -0.02)</td>
<td>&lt;0.001</td>
<td>-0.01 (-0.01, 0.01)</td>
<td>0.18</td>
</tr>
<tr>
<td>ED visits</td>
<td>-4.02 (-8.58, 0.55)</td>
<td>0.08</td>
<td>-5.41 (-9.29, -1.53)</td>
<td>0.007</td>
<td>1.40 (-4.59, 7.39)</td>
<td>0.64</td>
</tr>
<tr>
<td>ED wait time</td>
<td>-0.37 (-0.56, -0.17)</td>
<td>&lt;0.001</td>
<td>-0.04 (-0.14, 0.05)</td>
<td>0.37</td>
<td>-0.32 (-0.54, -0.11)</td>
<td>0.004</td>
</tr>
<tr>
<td>Employee movement</td>
<td>-0.07 (-0.08, -0.05)</td>
<td>&lt;0.001</td>
<td>-0.03 (-0.04, -0.01)</td>
<td>&lt;0.001</td>
<td>-0.04 (-0.06, -0.02)</td>
<td>0.001</td>
</tr>
<tr>
<td>Employee turnover</td>
<td>-0.03 (-0.05, -0.02)</td>
<td>&lt;0.001</td>
<td>-0.02 (-0.03, -0.01)</td>
<td>0.001</td>
<td>-0.01 (-0.03, 0.01)</td>
<td>0.13</td>
</tr>
<tr>
<td>Hospitalizations</td>
<td>-1.84 (-4.14, 0.46)</td>
<td>0.12</td>
<td>3.98 (2.05, 5.91)</td>
<td>&lt;0.001</td>
<td>-5.82 (-8.83, -2.83)</td>
<td>0.001</td>
</tr>
<tr>
<td>Radiology tests</td>
<td>-1.38 (-9.16, 6.41)</td>
<td>0.73</td>
<td>-19.47 (-34.65, -4.29)</td>
<td>0.01</td>
<td>18.9 (1.02, 35.15)</td>
<td>0.04</td>
</tr>
<tr>
<td>Time to complete radiology tests</td>
<td>-0.13 (-0.18, -0.08)</td>
<td>&lt;0.001</td>
<td>0.06 (0.01, 0.10)</td>
<td>0.02</td>
<td>-0.19 (-0.26, -0.12)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Time to sign radiology tests</td>
<td>-0.01 (-0.04, 0.01)</td>
<td>0.32</td>
<td>-0.04 (-0.05, -0.03)</td>
<td>&lt;0.001</td>
<td>0.02 (-0.02, 0.05)</td>
<td>0.07</td>
</tr>
<tr>
<td>ADEs rate</td>
<td>-0.10 (-0.15, -0.05)</td>
<td>&lt;0.001</td>
<td>-0.01 (-0.06, 0.04)</td>
<td>0.72</td>
<td>-0.09 (-0.16, -0.02)</td>
<td>0.02</td>
</tr>
</tbody>
</table>

### Discussion

We successfully tested a methodology consisting of an interrupted time-series design with multiple control sites and a wide range of outcome measures for monitoring the effect of EHR implementations on quality, productivity and safety outcomes. We detected changes to the performance of several measures over time, seasonal effects, and outcome variability. Such time-sensitive effects would not have been detected by pre-post or short-term time series designs, and demonstrate the importance of long-term studies of health IT evaluations. We also detected different effects on different measures, demonstrating the importance of using a broad set of measures to cover processes likely impacted by EHR adoption. The use of our methodology in future studies will benefit the broader informatics community from knowing what and how to monitor future EHR implementations. Furthermore, it will allow for earlier and potentially more precise detection of unexpected effects introduced by the intervention, and implementation of effective response to mitigate negative impacts, leading to a potentially faster return to, or improvement over, baseline performance.

Our methodology detected seasonal effects that were maintained in control sites in the post-intervention period, whereas the intervention sites seem to have been disrupted and affected for several months following the new EHR go live. While hemoglobin A1c control in the intervention group dropped and had a pattern similar to its initial baseline, ED length of stay and ED wait time stayed above baseline levels for seven months in the post-intervention period. Our methodology also detected specific patterns and outcome variability. Examples include medication for asthma and patient visits that varied even in the control settings; such a variability may have influenced detection of immediate effect in some cases and warrant further investigation of potential indirect effects of the implementation (e.g. organizational resources diverted to implementation regions). Such time-sensitive events would not have been detected if intervention sites were assessed without a control, or in a simple pre-post design, nor would the disruption on ED measures have been detected if data were assessed in a more spaced frequency, e.g. with yearly data collection. Therefore, we suggest future studies focusing on time-sensitive effects, as opposed to conducing
simple pre-post comparisons; we also recommend the use of more comprehensive measurements and more frequent data collection, for at least one year after go live. Otherwise, our understanding of the full impact of health IT adoption will be incomplete, and such evaluations may not be cost-effective. Our analysis included monthly data instances; however, several measures from our larger inventory\(^ {19}\) may be readily available in electronic format allowing for an even nearer real-time assessment (daily or weekly), dependent on the underlying count or frequency of the measurement. In addition, due to the broad coverage of our larger inventory\(^ {19}\), our methodology will be useful for monitoring implementations of new EHRs, upgrading of current EHRs, and other important changes both in the EHR and in the sociotechnical environment.

We also detected interesting improvements. Although patients may have been affected by delayed processes, patient satisfaction increased over time. Other improvements such as faster radiology tests may have contributed to improving patient experience. However, a more complete analysis must be conducted to confirm this hypothesis. As opposed to our previous study\(^ {19}\), where employee measurements were rated by a panel of informaticists as least relevant for monitoring EHR implementations, in the present study, such outcomes were affected over time, although it is difficult to infer how the implementation may have caused such changes. An explanation may be elicited by analyzing the subsequent implementation regions.

Studies evaluating IT adoption in other industries\(^ {25}\) demonstrate that IT adoption rarely produces positive results if not accompanied by complementary changes or investments. In addition, some outcomes assessed, such as ADEs rate, may have been affected by other effects that may or may not have been influenced by the new EHR implementation. Therefore, a more comprehensive analysis is necessary to account for possible confounders or complementary changes that may confirm or discard some of the effects we observed. We are currently evaluating the subsequent implementation regions and expect to report our results in more comprehensive future studies, including a qualitative analysis to identify and classify complementary changes introduced by the new EHR implementation and other potential confounders.

**Limitations.** Our study includes a small number of care settings. Our methodology effectively detected and quantified what changes happened, but not how and why they happened. We expect to improve its ability to detect how and why EHR implementations affect care delivery organizations by expanding our analysis in future studies including other implementation regions, measures, potential confounders, and a qualitative analysis of complementary changes introduced by the new EHR implementation. The two groups are located in different geographical areas and have different patient volume; we mitigated this limitation by comparing each group separately before comparing one against the other. Patients voluntarily answer the satisfaction survey, which could lead to selection bias.

**Conclusions**

We successfully tested a methodology that combines the use of 22 measures of quality, productivity and safety outcomes for monitoring performance changes during EHR implementations. We evaluated a commercial EHR implementation in a medium-size hospital and five ambulatory care clinics with multiple control sites. Nine (41\%) measures had an immediate effect after the go live, and 14 (61\%) measures an effect over time. We detected seasonal effects and outcome variability that would not have been detected by studies using simple measurements or pre-post comparisons. We recommend the use of our methodology in future studies to hopefully increase our understanding of the full impact of health IT interventions. A larger evaluation with the present measures and several others in 6 hospitals and over 50 clinics is underway.

**Acknowledgements**

Authors would like to acknowledge Intermountain’s data analysts for their support to data collection. Intermountain Healthcare, Salt Lake City, UT, USA, supported this research.
References

7. HITECH Act § 13400(5).
Evolution of an Implementation-Ready Interprofessional Pain Assessment Reference Model

Sarah A Collins, RN, PhD, Karen Bavuso RN, MSN, Mary Swenson RN, MBA, Christine Suchecki, MSN, RN, Perry Mar, PhD, Roberto A. Rocha, MD, PhD

Clinical Informatics, Partners eCare, Partners Healthcare Systems, Boston, MA; Clinical Content, Partners eCare, Partners Healthcare Systems, Boston, MA; Division of General Internal Medicine and Primary Care, Brigham and Women’s Hospital, Boston, MA; Harvard Medical School, Boston, MA

ABSTRACT

Standards to increase consistency of comprehensive pain assessments are important for safety, quality, and analytics activities, including meeting Joint Commission requirements and learning the best management strategies and interventions for the current prescription Opioid epidemic. In this study we describe the development and validation of a Pain Assessment Reference Model ready for implementation on EHR forms and flowsheets. Our process resulted in 5 successive revisions of the reference model, which more than doubled the number of data elements to 47. The organization of the model evolved during validation sessions with panels totaling 48 subject matter experts (SMEs) to include 9 sets of data elements, with one set recommended as a minimal data set. The reference model also evolved when implemented into EHR forms and flowsheets, indicating specifications such as cascading logic that are important to inform secondary use of data.

Introduction

In light of today’s prescription Opioid epidemic, evidence-based and standardized assessment and management of patients’ pain is a high priority from clinical, safety, quality, and regulatory perspective, as discussed in the 2016 Joint Commission Statement on Pain Management. Joint Commission standards emphasize the importance of comprehensive pain assessments consistent with the organization’s scope of care, treatment, and services, as well as with the patient’s age, condition, and ability to understand. Management includes treatment and reassessment of pain, noting that treatment may include pharmacologic and nonpharmacologic approaches, and considering risks and benefits, including potential risk of dependency, addiction, and abuse. Importantly, these standards do not require pain assessment for all patients, but rather based on organizational policy that considers the population served and services delivered. Joint Commission standards also do not require treatment until the pain score reaches zero, but rather emphasizes an individualized patient-centric approach to determine the most appropriate pain goal.

Standardizing pain assessment can support clinician judgment and decision making for individualized patient-centric approaches to pain management by providing consistent and comparative data capture. However, standardization of pain assessment requires acknowledgement that the criteria for what comprises a comprehensive pain assessment may vary across settings, patient populations, and scope of interprofessional practice. Variations in setting, population, and professional practice are a significant challenge for the definition of consistent sets of data elements to represent a comprehensive pain assessment. Moreover, implementation of a comprehensive pain assessment on Electronic Health Record (EHR) artifacts (e.g., forms, flowsheets) results in further variation due to functional differences across EHR tools used in different clinical settings and by different types of health professionals (e.g., nurses, physicians, physical therapists, pharmacists, pain specialists).
Background

A comprehensive pain assessment includes the use of a validated pain severity scale, as well as collection of data to understand the context of the pain experience, such as the type of pain, how it originated, how often it occurs, factors that alleviate or exacerbate pain, and how the pain impacts the patient’s quality of life. As preliminary work, our team performed a literature review of pain assessment data elements that should be captured as part of a comprehensive pain assessment. This review included published standards from Health Level 7 (HL7), Logical Observation Identifiers Names and Codes (LOINC), and the Intermountain Clinical Element Model, among other resources, resulting in a comprehensive reference model for pain assessment subsequently validated by local SMEs. Further, there is a lack of an implementation ready comprehensive reference model for pain assessment that is applicable across clinical settings, patient populations, interprofessional teams of clinicians, and EHR tools.

Such an implementation ready model requires a balance between data constraints and flexible implementation approaches. To achieve this balance, our team adheres to principles for consistent and high quality data capture, while also leveraging and relying on interprofessional clinical subject matter expertise (SME). Here we report results from this process, highlighting the iterative development and validation of an implementation-ready enterprise-wide comprehensive interprofessional pain assessment reference model.

Methods

Our team previously published a 10-step approach for the development, validation, implementation, and evaluation of clinical reference models: 1) identify clinical topics, 2) create draft reference models for clinical topics, 3) identify downstream data needs for clinical topics, 4) prioritize clinical topics, 5) validate reference models for clinical topics, 6) perform gap analysis of EHR data elements compared against reference model, 7) communicate validated reference models across project members, 8) request revisions to EHR data elements based on gap analysis, 9) evaluate usage of reference models across projects, and 10) monitor for new evidence requiring revisions to reference model. Here we report on the application of these steps for the clinical topic of “Pain Assessment,” and the subsequent evolution and refinement of our Pain Assessment Reference Model. Specifically, we report on the reference model evolution at the following points in time: a) initial draft (step 2; model version 1), b) interprofessional SME Validation Delphi Rounds 1 and 2 (step 5; model version 2), c) EHR form implementation (step 6; model version 3), d) specialty SME Validation (step 5; model version 4), and e) EHR flowsheet implementation (step 6; model version 5). Notably, this is an iterative process, hence SME validation in step 5 was repeated when appropriate, e.g. to expand the scope of the model to include pediatric populations, and step 6 was repeated to perform a gap analysis for implementation in flowsheets. In addition, step 5 validation when performed on an initial draft reference model leverages a large group of interprofessional SMEs from across our enterprise and uses a modified Delphi technique with online voting conducted in two rounds to achieve consensus. As we perform follow-up SME validation for refinements to the model, the groups of SMEs are smaller and more targeted, typically requiring a single validation session. Reference model terms were based on standards and SME voting.

Our team has also reported on methodologies for gap analyses of EHR data elements compared against reference model. These methods use the MUC-5 (Fifth Message Understanding Conference) Evaluation Metrics, originally used to express error rates as part of a scoring system for automated information extraction, to compare two sets of structured data elements to determine how well they are aligned. Given the applicability of MUC-5 Evaluation Metrics to compare and quantify variation between sets of information, we used the MUC-5 Evaluation metrics to compare the 5 successive
versions of our reference model. These metrics identify if data elements are a match, a partial match, conflicting, extra, or missing.

In this study we will report the overall iterative process end-result, which includes collaborative sessions with EHR analysts to review results of gap analyses and determine if EHR functionality constraints will require modifications to the reference model in order to support its implementation using current EHR tools.

**Results**

A total of 32 SMEs participated in the initial interprofessional validation sessions, resulting in version 2 of the reference model (see Table 1 and Figure 2). An additional 16 SMEs participated in the specialty sessions focused on the pediatric population, resulting in version 4 of the reference model. These SMEs represented the following roles: physicians, nurse practitioners, registered nurses, pharmacists, physical therapists, occupational therapists, and analysts; these SMEs worked in the following clinical or operational areas: inpatient nursing, ambulatory nursing, inpatient medicine, ambulatory medicine, emergency medicine, acute and chronic pain management, home health, inpatient physical therapy, ambulatory physical therapy, pharmacy, nurse education, oncology, hand therapy, osteopathic medicine, psychology, palliative care, patient safety and quality, and corporate compliance.

Table 1. Results of SME Validation and EHR Implementation: Data Element Types and Counts per Version

<table>
<thead>
<tr>
<th>Data Element Types</th>
<th>Initial Draft Model (v1)</th>
<th>Interprofessional SME Validation (v2)</th>
<th>Form Implementation (v3)</th>
<th>Pediatrics Specialty SME Validation (v4)</th>
<th>Flowsheet Implementation (v5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total (n)</td>
<td>21</td>
<td>26</td>
<td>29</td>
<td>33</td>
<td>42</td>
</tr>
<tr>
<td>Free Text (n)</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>8</td>
</tr>
<tr>
<td>Date/Time (n)</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Time with units (n)</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>Boolean (n)</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Value Set (n)</td>
<td>15</td>
<td>21</td>
<td>23</td>
<td>27</td>
<td>28</td>
</tr>
<tr>
<td>Validated Pain Severity Scales* (n)</td>
<td>1</td>
<td>5</td>
<td>13</td>
<td>13</td>
<td>16</td>
</tr>
</tbody>
</table>

*Number of pain scales not counted toward total data elements in reference model; v = version

The Pain Assessment reference model evolved from 21 data elements initially to 47 data elements in its final version (v5) (see Table 1). Throughout these iterative refinements, the data type of 1 data element changed, the names of 8 data elements were revised, and value sets adjusted for 16 data elements.

Figure 1 reports our MUC-5 Evaluation metrics comparing the initial model (v1) to the final model (v5). Figure 2 compares all 5 successive versions of our reference model using the MUC-5 Evaluation metrics. Overtime, 30 data elements total were added to the model and 17 data elements total were modified (Figure 1). Twelve new data elements were added to version 2 during interprofessional SME sessions, and 9 new data elements were added to version 4 during specialty SME sessions. Four new data elements were added to version 3 during implementation as an EHR form, and 5 new data elements were added to version 5 during implementation as a flowsheet (Figure 2).

As the model evolved and was implemented as EHR forms and flowsheets, intuitive sets of data elements emerged. The final model included 9 sets of data elements (see Table 2). The sets included an
average of 5.2 data elements with a range of 2 to 11 data elements each. The first set of data elements includes the recommendation for being “always included” on a documentation artifact (i.e. form or flowsheet). The remaining 8 sets of data elements can be optionally included if relevant to the clinical context of the documentation artifact. The “always include” set is comprised of 9 data elements that SMEs determined capture a minimal set of pain assessment data, namely concepts related to duration, location, quality, temporality, severity, and evaluation of pain relief.

Notably, two of the data elements in the “always include” set were added during our EHR flowsheet implementation step: 1) “Pain episode duration time unit”, and 2) “Pain assessment scale selection for cascade”. The form-based implementation required creation of a separate data element for the end-user to assert the unit of time associated with a numerical data entry for “Pain Duration”, such as minutes, hours, days, or weeks. There are 5 of this type of data element in the model: “Pain episode duration time unit”, “Pain onset time unit”, “Time period for best pain score”, “Time period for worst pain score”, and “Time period for average pain score”. Importantly, a clinical reference model may not require creation of separate data elements to handle units of measure, however, the instantiation of a clinical reference model ready for implementation in an EHR requires this level of specification. The data element ‘Pain assessment scale selection for cascade’ also was added during the last phase of refinement, readiness for Flowsheet Implementation, as a data element necessary to guide end-user to navigation using logic-based cascading within a flowsheet. However, we do recognize that these extensions to the reference model might need to be adjusted depending on the EHR system.
<table>
<thead>
<tr>
<th>Data Element</th>
<th>Data Type</th>
<th>Value Set List</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>SET: Pain Assessment [ALWAYS INCLUDE]</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pain Episode Duration</td>
<td>Numeric</td>
<td>NA</td>
</tr>
<tr>
<td>Pain Episode Duration Time unit</td>
<td>Time with unit</td>
<td>Seconds</td>
</tr>
<tr>
<td>Pain Location</td>
<td>Category</td>
<td>abdomen</td>
</tr>
<tr>
<td>Pain Location Qualifier</td>
<td>Category</td>
<td>Right</td>
</tr>
<tr>
<td>Pain Quality</td>
<td>Category</td>
<td>Ache</td>
</tr>
<tr>
<td>Relative Temporal Context</td>
<td>Category</td>
<td>Post-operative/procedure</td>
</tr>
<tr>
<td>Pain Assessment Severity Scale</td>
<td>Category</td>
<td>List of validated scales:</td>
</tr>
<tr>
<td>Selection for Cascade</td>
<td>Category</td>
<td>Use Numeric 0-10 Scale Scores:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>unless other scale from list below is required for patient population:</td>
</tr>
<tr>
<td>Is Pain Relief Acceptable?</td>
<td>Boolean</td>
<td>Yes</td>
</tr>
<tr>
<td><strong>SET: Pain Type and Count</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Pain Type</strong></td>
<td><strong>Category</strong></td>
<td>**Acute pain</td>
</tr>
<tr>
<td>---------------</td>
<td>--------------</td>
<td>---------------</td>
</tr>
<tr>
<td><strong>Pain Type not listed above</strong></td>
<td><strong>String</strong></td>
<td><strong>NA</strong></td>
</tr>
<tr>
<td><strong>Multiple Pain Sites</strong></td>
<td><strong>Boolean</strong></td>
<td>**Yes</td>
</tr>
</tbody>
</table>

**SET: Pain Initiation**

| **Pain Initiating Events** | **String** | **N/A** |
| **Rate of Pain Onset** | **Category** | **Sudden | Gradual** |
| **Pain Onset (how long ago)** | **Numeric** | **Seconds | Minutes | Hours | Days | Weeks | Months | Years** |
| **Pain Onset Time unit** | **Time with unit** | **Seconds | Minutes | Hours | Days | Weeks | Months | Years** |

**SET: Pain Course and Frequency**

| **Pain Course** | **Category** | **Resolved | Rapidly improving | Slowly improving | Relapsing and remitting | Not improving | Worsening** |
| **Pain Frequency** | **Category** | **Constant | Intermittent | Isolated (one-time)** |

**SET: Patient Stated Pain Goals**

| **Patients Tolerable Pain Score at Rest** | **Category** | **0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10** |
| **Patient Stated Functional Goal** | **String** | **NA** |

**SET: Pain Alleviating, Exacerbating and Associated Symptoms**

| **Pain Alleviating Factors** | **Category** | **Exercise | Eating | Medication | Positioning | Rest | Sleep | Therapy (e.g. massage, relaxation, hot/cold, please specify type) | Environment adjustment | Immobilization | Nothing** |
| **Pain Alleviating Factors comment** | **String** | **NA** |
| **Pain Exacerbating Factors** | **Category** | **Ambulating | Anxiety | Awakening | Deep breathing | Eating | Empty/Full stomach | Labor | Positioning | Rest | Sleep | Straining at stool | Childbirth labor** |
| **Associated Signs and Symptoms** | **Category** | **Anxiety | Arthritis | Bleeding Symptoms | Change In Consciousness | Claudication | Decreased Activity | Decreasing Blood Flow | Depression | Diabetes | Disability | Discomfort | Disease of immune system | Dyspnea | Fatigue | Fear | Hopelessness | Increased Need To Rest | Inflammation | Insomnia | Irritable | Mood change | Soreness | Stiffness | Stress | Tightness | Withdrawal From Activities** |
| **Expression of Pain – Pediatrics** | **Category** | **agression | agitated | anorexia | anxious | arched/rigid | body stiffness | change in activity pattern | change in sleep pattern | clenching teeth/lips | contracted limbs | crying | eyes wide open | facial expression | flailing | grimace | guarding | jerking | moaning | muscle tension | restless | rocking | rubbing | squirming | withdrawn | verbalization | not listed above (see comments)** |
| **Expression of Pain – Pediatrics comment** | **String** | **NA** |

**SET: Pain Radiation**

<p>| <strong>Pain Radiation</strong> | <strong>Category</strong> | <strong>abdomen | achilles | ankle | arm | axilla | back | breast | buttocks | calf | chest | chin | coccyx | ear | elbow | eye</strong> |</p>
<table>
<thead>
<tr>
<th>Pain Radiation Qualifier</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Right</td>
</tr>
<tr>
<td>SET: Pain Intervention and Reassessment</td>
<td></td>
</tr>
<tr>
<td>Pain Intervention</td>
<td>Category</td>
</tr>
<tr>
<td></td>
<td>Alternate therapy</td>
</tr>
<tr>
<td>Pain Intervention not listed above</td>
<td>String</td>
</tr>
<tr>
<td></td>
<td>NA</td>
</tr>
<tr>
<td>Neonatal Pain Intervention</td>
<td>Category</td>
</tr>
<tr>
<td></td>
<td>Comfort Measures</td>
</tr>
<tr>
<td>Neonatal Pain Intervention not listed above</td>
<td>String</td>
</tr>
<tr>
<td></td>
<td>NA</td>
</tr>
<tr>
<td>Previous Intervention Response</td>
<td>String</td>
</tr>
<tr>
<td></td>
<td>NA</td>
</tr>
<tr>
<td>Aberrant Behaviors</td>
<td>Category</td>
</tr>
<tr>
<td></td>
<td>None</td>
</tr>
<tr>
<td>Adverse Effects from Therapy</td>
<td>Category</td>
</tr>
<tr>
<td></td>
<td>None</td>
</tr>
<tr>
<td>Adverse Effects from Therapy comment</td>
<td>String</td>
</tr>
<tr>
<td></td>
<td>NA</td>
</tr>
<tr>
<td>SET: Patient Best and Worst Pain Scores</td>
<td></td>
</tr>
<tr>
<td>What number best describes, during the past week, how your pain has interfered with your enjoyment of life?</td>
<td>Category</td>
</tr>
<tr>
<td>What number best describes, during the past week, how your pain has interfered with your general activity?</td>
<td>Category</td>
</tr>
<tr>
<td>What was the patient's best pain score over the past week?</td>
<td>Category</td>
</tr>
<tr>
<td>What was the patient's worst pain score over the past week?</td>
<td>Category</td>
</tr>
</tbody>
</table>
Discussion

Pain assessment documentation is a well-scoped clinical domain, ripe for standardization using a standard reference model. Similar models from HL7, LOINC, and other organizations have informed the development of our initial model. In working closely with SMEs that represent care needs across health professions, including medicine, nursing, and physical therapy, and inpatient and outpatient settings, including chronic pain care clinics, we identified twice as many data elements compared to our initial version. Overall, a larger number of data elements were added to the model for the SME validated versions compared to the EHR implementation versions, indicating that the original reference model was lacking clinical concepts needed for assessing pain in different contexts.

Our experience confirms the need to develop comprehensive clinical topic reference models in stages and with a broad set of interprofessional SMEs that represent the breadth of clinical settings in which that topic is used. We found that the use of online voting, supplemented with targeted and structured group discussion that included guidance and recommendations from informaticians on our team, was successful in driving consensus and resolving disagreements among SMEs. While our team is comprised of informaticians who perform multi-site data science research studies, future work may benefit from inclusion of researchers and data scientists as SMEs, formally representing important consumers of secondary EHR data. We also confirmed that additional sessions with SMEs from relevant specialty areas were beneficial in later stages of model refinement, contributing additional data elements while maintaining the overall structure of the model. We found the overall decreasing rate of change as versions were refined by SMEs to be a positive indicator of reaching a satisfactory level of model completeness. Our team is in the process of replicating these findings using the same methodological steps for other clinical topic reference models. Importantly, continuous reference model optimization is expected and we recommend scheduled updates to incorporate new evidence and models of care, as well as future work to expand model scope to other relevant patient populations and specialty areas.

With the aim of driving adoption of clinical reference models within EHRs, we consider the application of a reference model to at least one EHR instance an important exercise that uncovers the complex decisions required to bridge the inherent gaps between a model and an implementation of that model. In our study we found that both EHR implementations of the model using different types of documentation functionality (e.g., forms versus flowsheets) required a similar number of modifications.
Functionality, user interfaces, and data structures can vary significantly - even within the same EHR system – impacting the ability to standardize implementation-ready clinical reference models.

The types of new data elements added for implementation-ready versions 3 and 5 could be categorized as assisting with end-user workflows and navigation logic, as well as data capture where units of measure are mandatory. We can argue that only SME validation sessions were needed to standardize the capture of important clinical concepts for pain documentation. However, an implementation-ready reference model that reflects decisions made to accommodate end-user workflows and use of cascading logic to support documentation navigation is important to guide subsequent interpretations required for data reporting and analytics. For example, in our work we added a data element “Pain Assessment Severity Scale Selection for Cascade” and a separate data element to capture the Pain Severity Score from whichever pain assessment scale was selected. To simplify our model we did not include each pain severity scale in its entirety, as these are all validated published scales. However, we needed to explicitly include data elements to: 1) support scale selection and use, and 2) capture pain score data in one field for all scores on a scale of 0-10 to support reporting of pain scores across patients, settings, and time.

Our ongoing work also includes mapping our reference model data elements to existing reference terminologies, such as SNOMED CT and LOINC. Importantly, EHR implementation challenges are a barrier to reusing generic data elements, such as anatomical location that is independent of the pain domain. Leveraging generic concepts in our reference terminology mappings is ideal, including post-coordinating when appropriate, however, some EHR data capture and reporting functionality does not support – or is significantly complicated by - duplicate instances of a generic field across multiple contexts. In 2015, Westra and colleagues outlined a number of challenges and lessons learned from a study modeling flowsheet data for quality improvement and research, including the overall need for standards to represent flowsheet data. Our approach identifying an “always Include” set comprised of the minimal data elements available for clinicians to document a pain assessment addresses, at least in part, the challenge of documentation inconsistencies within and across systems. Westra and colleagues identified cascading logic as a critical challenge for secondary use of flowsheet data, leading to ambiguity if data were missing or not applicable for a given patient. EHR flowsheet data comprises a large portion of recorded structured clinical data, particularly for hospitalized patients. As secondary use of these data increases, implementation-ready data element reference models that include relevant implementation specifications, such as cascading logic, will be critically important for reliable and accurate data interpretation. Our work in validating an implementation-ready reference model for pain assessment confirms that “implementation-ready” specifications include variation across EHR forms and flowsheets. Currently, our team is collaborating with the Big Data Science Consortium lead by Westra and colleagues from the University of Minnesota to validate and standardize a flowsheet-based Pain Reference Model across multiple health systems. This final version of our SME validated Pain Assessment Reference Model is being used to inform that multi-site work. This collaboration, in addition to our interprofessional approach and initial model development agnostic to the EHR is intended to minimize silos of EHR data. Future work should include definition and use of metrics to evaluate if the implementation of reference models meets the needs for secondary use of these data for analytic and research activities and incorporation reference models as an EHR standard.

Conclusion

Our work to develop an implementation-ready Pain Assessment Reference Model resulted in a model with 47 data elements organized into 9 data sets. Significant modifications to clinical concepts were made after sessions with interprofessional SMEs, including the addition of data elements and revisions to existing data elements and their value sets. The model further evolved when implemented into EHR forms and flowsheets, adding data elements to support documentation workflows and cascading logic.
Refinements based on SME expertise and EHR functionality are important and complementary in clinical data element reference model development and should be well documented to increase standardization of structured clinical data, while also informing accurate secondary uses of EHR data.

Acknowledgments

Thank you to the Partners eCare Structured Clinical Data Element Workgroup and to numerous clinical subject matter experts that participated in this work from at Partners Healthcare System.

References

Towards Analytics of the Patient and Family Perspective: A Case Study and Recommendations for Data Capture of Safety and Quality Concerns

Brittany Couturea, Maureen Fagan MHA,WHNP-BC,FNP-BCb, Esteban Gershanik MD, MPH, MMSc,e, Cathy Yoon MSa, James Benneyan PhDd, David W. Bates MD, MPH,e, Sarah A Collins, RN, PhDab,ec,

aDivison of General Internal Medicine and Primary Care, Brigham Health, Boston, MA; bCenter for Patients and Families, Brigham Health, Boston, MA; cHarvard Medical School, Boston, MA; dHealthcare Systems Engineering Institute, Northeastern University, Boston, MA; eClinical Informatics, Partners eCare, Partners Healthcare Systems, Boston, MA;

Abstract
Safety reporting systems are improving our current understanding of safety in hospital settings, although mostly from the clinician perspective. Patient Family Relations (PFR) programs provide the opportunity to capture patient/family concerns in the hospital. Descriptive statistics were completed of PFR concern submissions over a 20 month period, as well as a comparison of structured data fields to those of the AHRQ Common Format. We identified statistically significant differences in rates of concern submissions, methods of submission, and role of submitter across patient populations. Overall, the most frequent concerns submitted to PFR were care/treatment and communication concerns. There was very little overlap of the PFR data elements with those of the AHRQ Common Format (overall rate of mismatch approached 80%). These results emphasize both the unique information that PFR data provides, as well as the need for enhancement and continuity of reporting systems for more effective analysis of safety data.

Background/Introduction
Estimates suggest that up to 400,000 preventable deaths may occur each year in the U.S., and consistent vigilance, monitoring, and system improvements are required to improve patient safety. (1–5) Voluntary event reporting systems and patient safety organizations have greatly advanced our understanding of safety from the clinician perspective in the hospital setting. (6–9) However, detailed understandings of the perspective of patients and families who experience a wide range of undocumented safety concerns every day are limited. A recent JAMA study found that patients whose surgeons have large numbers of unsolicited patient concerns and dissatisfaction are more likely to suffer from medical or surgical complications. (10) Patient reported concerns are not limited to individual clinicians, rather they typically capture perceptions of clinical and non-clinical services and processes in a hospital setting that can elucidate systemic weaknesses experienced by patients and families that are not readily clear to staff and administrators.

Root cause analysis (RCA) has recently been criticized for failing to deliver benefits of preventing future harms ‘at-scale’ and oversimplifying based on linear thinking that inappropriately leads to the identification of a single root cause. (11,12) Rather, systems thinking recognizes the multilayered complexity of health services and seeks to solve unrecognized systems issues. (12) In addition, RCA does not typically include the patient and family perspective, and is weakened by hindsight bias and the focus on single incidents which limits opportunity for understanding vulnerability in the context of prevalence of other similar incidents. (11,12) Electronic capture of patient and family reported concerns in data sets that grow overtime could mitigate these weaknesses by enabling real-time analytics aligned with systems thinking frameworks for improvement and evaluation of solutions that consider the patient and family perspective. The National Patient Safety Foundation and Peerally and colleagues recommend the creation
of safety data sets for aggregated analyses at multiple levels (e.g., department, institution, system, state, and national).(11,12)

Hospital-based Patient and Family Relations (PFR) programs provide a mechanism for patients (or their family, friends or staff members on the patients’ behalf) to share any concerns they have while receiving care and serve as a resource to coordinate the resolution of outstanding concerns. These essential programs provide a unique and highly valuable service for patients and families to contact in real-time and at the point of care to resolve clinical and non-clinical concerns. In addition to resolving concerns in real-time, the data captured as part of Patient and Family Relations work is highly valuable from a patient safety perspective to complement existing safety reporting systems. Several data integrity issues exist when capturing safety incidents and near misses/concerns in hospital reporting systems, such as incomplete data in clinician reporting systems.(13) Triangulation of safety data sets from different sources with varied perceptions of care delivered could mitigate the impact of missing data in each data set. Yet, the triangulation of these safety data sets requires standardization of data capture for aggregated and comparative analyses. The AHRQ’s Common Format for capturing safety events provides standards to capture ‘incidents’, ‘near misses’, and unsafe conditions’. (14) An ‘incident’ is defined as a patient safety event that reached the patient, ‘near miss’ as a patient safety event that did not reach the patient, and ‘unsafe condition’ as a circumstance that increases the probability of a patient safety event.

The aims of this study are to: 1) describe reporting trends to Patient and Family Relations across inpatient populations to increase our understanding of patient perceived safety concerns, and 2) compare Patient and Family Relations data set to AHRQ’s Common Format to identify critical gaps that constrain aggregated analyses and learning opportunities of the patient and family perspective and experience in the context of these quality and safety concerns.

Methods
As part of a study focused on hospitalized patient and family perceptions of safety concerns and reporting tools we analyzed a data set of concerns shared with a PFR program at a large academic medical center. These data were de-identified and represented a time period of 20 months (January 2015 - August 2016) for hospital units including the following patient populations: critical care medical and surgical patients, and acute care cardiovascular, medical, surgical, neurology, and oncology patients. The de-identified data set contained the following fields for each submission: clinical unit, date received, resolution date, report submitter, method of submission, description of concern, outcome notes, and up to 4 concern categories and subcategories (selected by PFR staff that entered the reported concern into the system). Method of submission is a structured field, with the following value set: telephone, walk-in, email, letter, and website. Descriptions of concern and outcome notes are free-texts field. The options for report submitter were bucketed into the following values: patient, family/friend, staff, and other.

We performed two sets of analyses: 1) descriptive statistics of role of submitter, method of reporting, and categories of concerns stratified by patient population, and 2) comparison of structured data capture fields from Patient and Family Relations database to AHRQ Common Format. To enable comparative descriptive statistics across patient populations with varying number of patient beds and lengths of stay we utilize counts per 1000 patient days calculated from average nightly census data. This rate was calculated by total number of submissions divided by the total number of patient days multiplied by 1000. We also performed a small sub analysis on the reliability of the structured report submitter field by comparing it to the information recorded in the description of concern field.

We used the MUC-5 (Fifth Message Understanding Conference) Evaluation Metrics to compare PRF data fields to AHRQ Common Format.(15) These metrics are used to express error rates and were originally developed as part of a scoring system for automated information extraction when compared to manual
Results

Descriptive Statistics

A total of 554 concerns were submitted during the 20 months of the study data across all seven inpatient populations: critical care medical and surgical patients, and acute care cardiovascular, medical, surgical, neurology, and oncology patients. Table 1 displays submission rates per 1000 patient days for each patient population along with the range.

<table>
<thead>
<tr>
<th>Patient Population</th>
<th>Cardiovascular</th>
<th>Neurology</th>
<th>Oncology</th>
<th>Medical</th>
<th>Surgical</th>
<th>Surgical</th>
<th>Medicine</th>
</tr>
</thead>
<tbody>
<tr>
<td>Submission rate (per 1000 patient days)</td>
<td>2.709</td>
<td>5.112</td>
<td>2.886</td>
<td>4.235</td>
<td>5.634</td>
<td>3.116</td>
<td>2.247</td>
</tr>
</tbody>
</table>

Tests used: ¹Poisson test for equal rates; ²Fishers Exact test

Table 1: Concern submissions, submission method, and submitter per patient population

<table>
<thead>
<tr>
<th>Submission Method</th>
<th>Total Count</th>
<th>Acute Care</th>
<th>Critical Care</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Telephon e</td>
<td>40</td>
<td>114</td>
<td>48</td>
<td>216</td>
</tr>
<tr>
<td>Walk In</td>
<td>7.5%</td>
<td>9.48%</td>
<td>4.17%</td>
<td>12.04%</td>
</tr>
<tr>
<td>Email</td>
<td>11.43%</td>
<td>4.31%</td>
<td>12.5%</td>
<td>4.63%</td>
</tr>
<tr>
<td>Letter</td>
<td>0%</td>
<td>1.72%</td>
<td>10.42%</td>
<td>2.31%</td>
</tr>
<tr>
<td>Website</td>
<td>2.5%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Submitter</th>
<th>Total Count</th>
<th>Acute Care</th>
<th>Critical Care</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient</td>
<td>40</td>
<td>20.18%</td>
<td>10.42%</td>
<td>35.19%</td>
</tr>
<tr>
<td>Family and Friend</td>
<td>30%</td>
<td>30.7%</td>
<td>22.92%</td>
<td>28.24%</td>
</tr>
<tr>
<td>Staff</td>
<td>30%</td>
<td>49.12%</td>
<td>66.67%</td>
<td>34.72%</td>
</tr>
<tr>
<td>Other</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>1.85%</td>
</tr>
</tbody>
</table>

1Submission rate calculated by total number of submissions divided by total number of patient days multiplied by 1000

Tests used: ¹Poisson test for equal rates; ²Fishers Exact test
We identified statistically significant differences in rates of concern submissions overall across all patient populations types (Fisher’s Exact Test, p-value <0.0001). We also found statistically significant differences in methods of submission used across patient populations, as well as differences in role of submitter across patient populations (Fisher’s Exact Test p-value 0.0258 and 0.0006, respectively). See Table 1 for the rates and statistical results. Acute care surgical patients had the highest rate of submission per 1000 patient days, followed by acute care neurology patients. The critical care medical patient population had the lowest submission rate across the given time period at 2.25 submissions per 1000 patient days. The telephone was the most used submission method for reporting a concern across all populations, with usage rates for telephone between 62.5% and 84.5%. Acute care oncology and neurology frequency of staff submissions (66.7% and 49.1%, respectively) stand out as an anomaly from the other acute care patient populations, which otherwise have staff submission frequencies at 30%, 34.7%, and 35.2%. The role of submitter that was most frequent varied across the patient populations. As expected due to patient acuity, we found that families and staff submitted concerns on behalf of critical care patients at higher rates than critical care patients submitted themselves. Subsequently, acute care surgical and cardiovascular patients submit more frequently than family and staff for those populations.

A sub analysis was performed on the accuracy of the report submitter field in the PFR data. In about half of all the submissions (278 out of 554) the description of concern field directly stated who contacted PFR about the concern. Out of these 278 submissions, we looked at a random 20% (55) and checked the accuracy of the report submitter structured field against the description of concern. 8 out of these 55 had conflicting documentation on who submitted the particular concern.

The PFR data set included a total of 19 potential concern categories that could be attributed to a concern submission. The 5 most frequently coded categories overall were Care/Treatment, Communication, Coordination and Continuity of Care, Discharge, and Attitude. The top 2-4 categories across each patient population were consistently comprised from these 5 categories, with slight variation in specific ranking (see Table 2).

<table>
<thead>
<tr>
<th>Table 2: Top concern categories overall and per patient population</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Care/Treatment</td>
</tr>
<tr>
<td>2. Communication</td>
</tr>
<tr>
<td>3. Coordination and Continuity of Care</td>
</tr>
<tr>
<td>4. Discharge</td>
</tr>
<tr>
<td>5. Attitude</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Acute Care</th>
<th>Cardiovascular</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Communication</td>
<td>22.4% (17)</td>
</tr>
<tr>
<td>2. Attitude</td>
<td>10.5% (8)</td>
</tr>
<tr>
<td>3. Coordination and Continuity of Care</td>
<td>10.5% (8)</td>
</tr>
<tr>
<td>4. Discharge</td>
<td></td>
</tr>
</tbody>
</table>

| Neurology | 1. Communication | 17.7% (31) |
|-----------|-----------------|
| 2. Care/Treatment | 14.9% (26) |
| 3. Discharge | 13.1% (23) |

| Oncology | 1. Care/Treatment | 29.3% (22) |
|----------|-----------------|
| 2. Attitude | 13.3% (10) |
| 3. Communication | 12.0% (9) |

| Medical | 1. Care/Treatment | 17.2% (51) |
|---------|-----------------|
| 2. Communication | 14.8% (44) |
| 3. Coordination and Continuity of Care | 12.5% (37) |

| Surgical | 1. Coordination and Continuity of Care | 16.3% (21) |
|----------|-----------------|
| 2. Communication | 15.5% (20) |
| 3. Care/Treatment | 17.8% (23) |
The breakdown of frequency of concern category used per patient population (normalized across populations by 1000 patient days) can be seen in Figure 1. This shows all the categories and their uses by the various patient populations. It is important to note that up to four categories could be attributed to one submission (i.e., many to one), so the number of categories used does not correlate directly to the number of submissions per patient population.
Figure 2: PFR data elements mapped to AHRQ Common Format data elements

Total primary error represents the number of wrong elements over total elements; undergeneration is the number of elements missing over the number of elements possible; overgeneration is the number of extra elements over number of actual elements. Substitution is defined by incorrect matches and partial matches divided by correct, partial, and incorrect elements.

Mapping to AHRQ Common Formats

The comparison of data elements from the AHRQ Common Formats compared to Patient and Family Relations data elements indicate very little overlap with an overall rate of mismatch of almost 80% (see Figure 2). Interestingly, from our data set of Patient and Family Relations submissions the most frequently selected safety event/concern categories were also the most generic safety event/concern categories from within the categorization schema (see Figure 1). Table 3 below shows the detailed mapping results between PFR data elements compared to the AHRQ Common Formats data elements with shading to note unmatched (missing) data elements.

<table>
<thead>
<tr>
<th>Data Type</th>
<th>AHRQ Common Format Data Elements</th>
<th>Patient Family Relations Data Elements</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient Information</td>
<td>Name</td>
<td>Name</td>
</tr>
<tr>
<td></td>
<td>Date of Birth</td>
<td>MRN</td>
</tr>
<tr>
<td></td>
<td>Gender</td>
<td></td>
</tr>
<tr>
<td>Reporter</td>
<td>Anonymous</td>
<td>Attending MD</td>
</tr>
<tr>
<td></td>
<td>Name</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Telephone Number</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Email Address</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Job or Position</td>
<td>Report Method</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Communicated By</td>
</tr>
<tr>
<td>Event Information</td>
<td>Event Discovery Date</td>
<td>Date Received</td>
</tr>
<tr>
<td></td>
<td>Event Discovery Time</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Narrative Description of event</td>
<td>Description</td>
</tr>
<tr>
<td></td>
<td>Narrative Description of event location*</td>
<td>Location*</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Outcome Notes</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Resolution Date</td>
</tr>
<tr>
<td></td>
<td></td>
<td>File Status</td>
</tr>
<tr>
<td></td>
<td></td>
<td>File Owner</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Entered By</td>
</tr>
<tr>
<td>Safety Event Type</td>
<td>Incident</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Near Miss</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Unsafe Condition</td>
<td></td>
</tr>
<tr>
<td>Safety Event/Concern Categories</td>
<td>Blood or Blood Product</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Device or Medical/Surgical Supply</td>
<td></td>
</tr>
<tr>
<td>Event Type</td>
<td>Cause</td>
<td></td>
</tr>
<tr>
<td>------------------------------------------------</td>
<td>--------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>Fall*</td>
<td>Accident/Injury*</td>
<td></td>
</tr>
<tr>
<td>Healthcare-associated Infection</td>
<td>Medication</td>
<td></td>
</tr>
<tr>
<td>Medication or other substance</td>
<td>Medication</td>
<td></td>
</tr>
<tr>
<td>Perinatal</td>
<td>Other</td>
<td></td>
</tr>
<tr>
<td>Pressure Ulcer</td>
<td>Care/Treatment</td>
<td></td>
</tr>
<tr>
<td>Surgery or Anesthesia (includes invasive procedure)</td>
<td>Coordination and Continuity of Care</td>
<td></td>
</tr>
<tr>
<td>Venous Thromboembolism</td>
<td>Discharge</td>
<td></td>
</tr>
<tr>
<td>Other (Please specify)</td>
<td>Patient Room</td>
<td></td>
</tr>
</tbody>
</table>

Following are two examples from our data that demonstrate the implications of the high rate of mismatch between the Patient and Family Relations and AHRQ Common formats data elements. One narrative description from our Patient and Family Relations was: “The care coordinator left a message stating that the patient’s sister is upset that nursing staff did not understand the healthcare proxy documentation policy.” This description was categorized as ‘Care/Treatment’ with a subcategory of ‘Participation in care plan’. While this categorization is arguably insufficient as it does not capture the detailed nature of the submission or indicate its relevance as a healthcare proxy concern, it does categorize the issue at a high level. However, there are no safety event/concern categories from the AHRQ Common formats that would apply to this issue, requiring that it would be coded as ‘other’. It could be argued to be an ‘unsafe condition’ that increases the probability of patient safety event from the perspective that the sister’s ability to act as a patient advocate is being limited in this situation.

A second example is: “Patient is concerned about how to care for her shunt once she is discharged”. In the Patient and Family Relations data this was categorized as: ‘Care/Treatment’ with a subcategory of ‘Outcome of surgery/procedure’. These categorizations are also arguably insufficient, although not wrong. Using the AHRQ Common Formats this could be an ‘unsafe condition’ increasing the probability of complication or infection and a readmission to the hospital. However, we consider the safety event

---

**Shaded boxes indicate no match (missing). *indicates partial match**
categories ‘Healthcare-associated Infection’ and ‘Surgery or Anesthesia (includes invasive procedure)’
tangential, or only slightly connected, and not appropriate to categorize the issue.

Discussion
This data set of patient and family concerns contains a vast amount of information that can be used to
learn more about the patient and family experience and patient-perceived concerns across inpatient
populations. There was interesting variation in patient and family concern reporting across patient
populations. One trend in particular that stands out is the high level of staff submissions seen on the acute
care oncology and neurology units. Perhaps due to the nature of the conditions for these specialty
populations there is a greater need for patient advocates. Oncology care can be very complex and of
significant gravity, difficult to navigate, and there are likely many cognitive issues on neurology, both
making it complicated for the patient to advocate for his or her self. There are also some variations in
reporting method: telephone use is most common across all patient populations, but there is a higher use
of walk-ins for acute and critical care medical populations than, for example, acute and critical care
surgical populations. Communication, care/treatment, coordination and continuity of care, attitude, and
discharge are the most commonly reported categories of concern, with good consistency across the patient
populations looked at in this study.

The discrepancies found in the sub analysis on accuracy of the report submitter field highlight the need to
review the structured field definitions to confirm if they meet the needs of the data set and the individual
entering the data, either through usability analysis or detailed interviews to better understand the
workflow of Patient and Family relations and barriers to this data entry. Patient concerns are often
complex and not straightforward, and may include multiple parties. The addition of a structured field
indicating who is concerned, or who is followed up with after the initial contact could give a better picture
of the situation, e.g. the nurse calls at the request of the family and then the family talks to PFR to
describe the details of the concern. Since these situations are not always as simple as one person
reporting, we need a way to reflect this in the structured data. In the long run that will allow for better
and more accurate data analysis using big data analytics with less need of retrospective qualitative
checking.

Upon detailed analysis, the current categorizations in the PFR data set are too generic to adequately
capture the context and content of concerns. Therefore, aggregated quantitative analyses of these
categories and subcategories provide limited information and utility. In fact, qualitative analyses are
needed to learn relevant trends from this rich data set that could increase our understanding of the patient
experience and inform system improvement. There exists an opportunity for a data-driven categorization
of concerns based on these data sets.

Further, the data elements used by Patient and Family Relations appears tangential to the AHRQ
Common Formats, likely due to variation in the purpose of data collection for these data sets. The
significant differences in concern categories between the AHRQ Common Format and the PFR data show
there is much need for intensive work on providing better and more informative categorization for the
PFR data. It is also clear that there is a large need for the information the PFR data supplies, since it is
not well represented in the AHRQ Common Format. Demonstrated through the descriptive PFR
examples such as ‘Patient is concerned about how to care for her shunt once she is discharged’, we can
see that the AHRQ common format categories are likely more designed to capture near misses and
incidents, which are typically gathered from clinician-type reporting systems. The type of data that PFR
captures can help inform on the “unsafe conditions” incident put forth by the AHRQ Common Format in
providing the unique view of hospital conditions from the patient and family perspective, and differences
and tendencies among patient populations.
Patient and Family Relations does not explicitly have a method to submit anonymously (a phone call or a letter could always be unidentified or a patient/family member could request to stay anonymous, but that is not the main pretense of the system). The AHRQ Common Format does have the option to submit an anonymous concern. A straightforward way to anonymously submit safety concerns through PFR might provide an important view into the group of patients that might not report concerns otherwise. This can be for a number of reasons, but most prominent is the fear of repercussion; in reporting a concern, some patients and family members might worry that his or her care will be compromised or will be treated differently. Our team is leading work to provide patients with an electronic reporting app called MySafeCare on each clinical unit to complement existing hospital-wide reporting methods. MySafeCare is an electronic safety reporting tool that aims to 1) provide patients and families that might otherwise not speak up with an anonymous way to address concerns they have about their hospital stay, and 2) create a quick and easy to use method of reporting that can facilitate real-time capture of worrisome or concerning events. (16,17) This tool is helping to add to the pool of patient-generated data that is becoming more recognized as important and useful in patient care. The structured Patient and Family Relations data is extremely beneficial to learn from because it already established as an operational system in the hospital. Initiatives like MySafeCare aim to target areas of need, i.e. the patients that wouldn’t otherwise report a concern because they do not want to do it in person or identified. We can leverage analysis of PFR data to learn about areas of need and effective methods to implementing these new types of tools. Current development work is also being done with MySafeCare to enable data exports using AHRQ Common Formats, further aligning data for ease of analysis.

Table 5 below gives a summary of the recommendations for patient reporting systems and their potential benefits based on the analysis completed for this paper. A robust and comprehensive patient and family reporting system will greatly improve our knowledge of patient perspectives in the hospital setting.

<table>
<thead>
<tr>
<th>Table 5: Recommendations for patient reporting systems</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Recommendation</strong></td>
</tr>
<tr>
<td>Optimization of complete and accurate data capture for report submitter(s) structured field</td>
</tr>
<tr>
<td>Addition of structured field for who is concerned</td>
</tr>
<tr>
<td>Anonymous submission option</td>
</tr>
<tr>
<td>Data-driven refinement of categories and subcategories</td>
</tr>
<tr>
<td>Frequent and continuous (bi-yearly, annual, etc) analysis of submission trends</td>
</tr>
<tr>
<td><strong>Potential Benefits</strong></td>
</tr>
<tr>
<td>1. Ability to describe reporting patterns based on role across patient populations with confidence</td>
</tr>
<tr>
<td>2. Inform specific outreach to patients and families by patient advocates based on the patient populations characteristics on how to report concerns</td>
</tr>
<tr>
<td>1. In depth understanding of the nature of the concern (i.e. is the family member calling on behalf of the patient?)</td>
</tr>
<tr>
<td>2. In real time, helpful for the PFR representative that is following up on the case to know who to address</td>
</tr>
<tr>
<td>1. Ability to capture concerns of patients/family members that may otherwise be uncomfortable speaking up (structured field option)</td>
</tr>
<tr>
<td>1. Will allow for big data analytics to better study patient concern patterns</td>
</tr>
<tr>
<td>2. Reduce the need for extensive qualitative analysis</td>
</tr>
<tr>
<td>1. Better understanding of patterns of submissions</td>
</tr>
<tr>
<td>2. Opportunity for continuous improvement both at the unit level and hospital wide</td>
</tr>
<tr>
<td>3. Inform development of tools like MySafeCare for areas of need</td>
</tr>
</tbody>
</table>

**Conclusion**

Patient and family generated safety concern data can be utilized to help systemically improve patient care and satisfaction at both a unit and hospital-wide level. We can learn, among other things, which types of concerns are most prevalent on different clinical units, which patient populations are in the most need for patient advocates, and the best methods for increasing communication of these issues. Development and enhancement of reporting systems that capture concerns from the patient and family perspective are needed to improve on the quality and accuracy of data capture, so we can learn important reporting.
patterns across patient populations in a standardized manner. This information can complement data from clinician-reported events and offer critical insight into safety and the patient experience in the hospital.

Acknowledgements
Thank you to Patient and Family Relations Committee for their collaboration on this study. This work was funded by AHRQ 1P30HS0235335 Making Acute Care More Patient Centered.

References
Calibration Drift Among Regression and Machine Learning Models for Hospital Mortality

Sharon E. Davis, MStat1, Thomas A. Lasko, MD, PhD,1 Guanhua Chen, PhD,1 Michael E. Matheny, MD, MS, MPH1 2
1Vanderbilt University School of Medicine, Nashville, TN; 2 VA Tennessee Valley Healthcare System, Nashville, TN

Abstract

Advanced regression and machine learning models can provide personalized risk predictions to support clinical decision-making. We aimed to understand whether modeling methods impact the tendency of calibration to deteriorate as patient populations shift over time, with the goal of informing model updating practices. We developed models for 30-day hospital mortality using seven common regression and machine learning methods. Models were developed on 2006 admissions to Department of Veterans Affairs hospitals and validated on admissions in 2007-2013. All models maintained discrimination. Calibration was stable for the neural network model and declined for all other models. The L-2 penalized logistic regression and random forest models experienced smaller magnitudes of calibration drift than the other regression models. Calibration drift was linked with a changing case mix rather than shifts in predictor-outcome associations or outcome rate. Model updating protocols will need to be tailored to variations in calibration drift across methods.

Introduction

Risk prediction models are developed across clinical domains1-6 to support patient and provider decision-making, adjust quality metrics for acuity, and augment enrollment strategies for clinical trials1,5. Although model validations have traditionally focused on discrimination, calibration is increasingly recognized as an essential aspect of model accuracy, particularly when models support individual-level patient decision-making1, 7, 8. With the adoption of electronic health records and use of advanced modeling methods, the role of clinical prediction models and our understanding of the challenges presented by the incorporation of predictive analytics into clinical care are rapidly evolving1, 9-11. One such challenge is deterioration of model calibration as characteristics of patient populations shift over time6,12-15. Limited evidence is available regarding how such calibration drift may vary across advanced regression and machine learning models, and how modeling methods may impact the need for routine model updating.

Mortality within 30-days of hospital admission, a key metric of hospital quality and patient safety, is assessed and tracked by the Centers for Medicare and Medicaid Services (CMS) to both inform the public and, since 2013, to adjust reimbursements16. CMS quality metrics rely on prediction models to standardize mortality rates by adjusting for the case mix of each hospital’s patient population16. In addition to quality benchmarking, prediction models for hospital mortality may be used to support decision-making particularly in the critical care setting. Numerous prediction models for hospital mortality have been presented in the literature2, and external validations have documented deteriorating calibration over time for many of these models13,14,17-19. While advanced regression and machine learning methods have been implemented for mortality prediction20-22, studies of calibration drift have primarily focused on logistic regression models and crude measures of average calibration. We sought to extend this existing literature through consideration of additional modeling methods, more stringent assessments of calibration, and systematic quantitative evaluations of the population data shifts driving performance changes.

Methods

We accessed data on eight years (2006-2013) of admissions to Department of Veterans Affairs hospitals nationwide with corresponding pre-admission data from 2002 for admitted patients, and randomly selected 50% of the sites within each Veterans Integrated Service Network for inclusion in this analysis. We modeled 30-day all-cause mortality after hospital admission using a predictor set developed through review of previously published risk models2, 23-25. Predictors were extracted from the national Corporate Data Warehouse26 and included demographics, vital signs during the first 48 hours of admission, laboratory values during the first 48 hours of admission, and diagnoses codes, medications, and care utilization prior to admission. Admissions were excluded if the patient was under 18 years of age, discharged to hospice care, admitted for less than 48 hours, or lacked 30-days of available follow-up (i.e., admitted in the last 30 days of 2013).
Admissions beginning in 2006 served as our training set and admissions beginning in 2007-2013 served as our validation set. Using the common training data and predictors set, we developed seven parallel models for hospital mortality based on logistic regression (LR), L-1 penalized logistic regression (L1; i.e., lasso), L-2 penalized logistic regression (L2; i.e. ridge), L-1/L-2 penalized logistic regression (L1-L2; i.e., elastic net), random forest (RF), neural networks (NN), and naïve Bayes (NB)\textsuperscript{27}. For those models requiring tuning (i.e., L1-L2, NN, RF), hyperparameters were selected with 5-fold cross-validation. Each model was internally validated with the bootstrap (B=200) using only admissions in the training set (2006 admissions). We divided the 7-year validation set into consecutive 3-month periods (n=28) and assessed performance of the models within each. Discrimination was measured with the area under the receiver operating characteristics curve (AUC)\textsuperscript{28}. Calibration was measured with observed to expected outcome ratios (O:E), the Cox logistic recalibration model’s intercept and slope, and flexible calibration curves\textsuperscript{29, 30}. The most stringent measure of calibration in this study, flexible calibration curves assess calibration across the range of probabilities by fitting a logistic model for the observed outcome based on predicted probabilities fit with a restricted cubic spline\textsuperscript{29}. We summarized these calibration curves with the estimated calibration index (ECI; the mean squared difference between predicted probabilities and estimated observed probabilities from the flexible calibration curves)\textsuperscript{29} and by regions of predicted probability where the model was calibrated, overpredicted, or underpredicted (i.e., ranges of probability over which the confidence interval of the flexible calibrations curve captured the 45° line of perfect agreement, was below the line, or was above the line were labeled as regions of calibration, overprediction and underprediction, respectively). In addition, we calculated within-region ECIs and rescaled each region by the volume of observations with predictions in its range. This proportional regional volume assessment emphasized calibration status based on data density and ranges of probability most relevant to the observed data.

Table 1. Patient population at development (2006) and during three validation years

<table>
<thead>
<tr>
<th></th>
<th>2006</th>
<th>2007</th>
<th>2010</th>
<th>2013</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>235,548</td>
<td>235,734</td>
<td>243,631</td>
<td>214,798</td>
</tr>
<tr>
<td>% 30-day mortality</td>
<td>5.0</td>
<td>4.9</td>
<td>4.9</td>
<td>4.7</td>
</tr>
<tr>
<td>Age in years (mean and SD)</td>
<td>62.9 (13.7)</td>
<td>63.0 (13.8)</td>
<td>63.6 (14.0)</td>
<td>63.9 (14.3)</td>
</tr>
<tr>
<td>% Female</td>
<td>4.5</td>
<td>4.7</td>
<td>4.9</td>
<td>5.5</td>
</tr>
<tr>
<td>Race</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>% White</td>
<td>71.7</td>
<td>71.6</td>
<td>72.3</td>
<td>72.1</td>
</tr>
<tr>
<td>% Black</td>
<td>19.8</td>
<td>20.0</td>
<td>19.6</td>
<td>19.8</td>
</tr>
<tr>
<td>% American Indian/Alaskan</td>
<td>1.3</td>
<td>1.4</td>
<td>1.5</td>
<td>1.6</td>
</tr>
<tr>
<td>% Asian/Pacific Islander</td>
<td>1.1</td>
<td>1.2</td>
<td>1.2</td>
<td>1.3</td>
</tr>
<tr>
<td>% Unreported</td>
<td>6.0</td>
<td>5.9</td>
<td>5.5</td>
<td>5.3</td>
</tr>
<tr>
<td>BMI at admission (mean and SD)</td>
<td>28.2 (7.1)</td>
<td>28.3 (7.3)</td>
<td>28.7 (7.2)</td>
<td>28.8 (7.1)</td>
</tr>
<tr>
<td>Health care utilization (prior year)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inpatient visits (mean and SD)</td>
<td>1.3 (2.0)</td>
<td>1.3 (2.0)</td>
<td>1.3 (2.0)</td>
<td>1.3 (2.1)</td>
</tr>
<tr>
<td>Outpatient visits (mean and SD)</td>
<td>36.4 (43.6)</td>
<td>37.1 (43.3)</td>
<td>42.0 (48.2)</td>
<td>43.5 (48.9)</td>
</tr>
<tr>
<td>Select diagnoses (preadmission)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chronic pulmonary disease</td>
<td>28.4</td>
<td>32.4</td>
<td>38.5</td>
<td>41.2</td>
</tr>
<tr>
<td>Congestive heart failure</td>
<td>17.3</td>
<td>19.1</td>
<td>22.0</td>
<td>23.7</td>
</tr>
<tr>
<td>Depression</td>
<td>20.1</td>
<td>24.5</td>
<td>32.6</td>
<td>38.4</td>
</tr>
<tr>
<td>Dyslipidemia</td>
<td>41.8</td>
<td>49.6</td>
<td>61.5</td>
<td>66.4</td>
</tr>
<tr>
<td>Hypertension</td>
<td>61.7</td>
<td>67.4</td>
<td>74.2</td>
<td>76.4</td>
</tr>
<tr>
<td>Renal failure</td>
<td>12.3</td>
<td>15.3</td>
<td>19.5</td>
<td>21.9</td>
</tr>
</tbody>
</table>

We also characterized data shifts in the patient population to link any changes with model performance over time. We documented shifts in the mortality rate and in the distribution of each predictor across the study period. Case mix changes were further assessed by fitting membership models in each 3-month validation period. Membership models are prediction models aimed at discriminating between validation and development observations using the outcome and predictor set from our original model to determine whether case mix shift is present and identify variables contributing to such shift\textsuperscript{31}. We refit the LR, L1, and RF models in each 3-month validation period to explore shifts...
in the strength of associations between predictors and 30-day mortality. Any changes in association were documented as changes in LR odds ratios, L1 variable selection patterns, and RF variable importance ranks.

All analyses were conducted in R 3.2. This study was approved by the Institutional Review Board and the Research and Development committee of the Tennessee Valley Healthcare System VA.

Results

Nationwide, 3,467,142 admissions to VA facilities met all eligibility criteria for our 30-day all-cause mortality models. Restricting to admissions from the randomly selected sites, our analysis set included 1,893,284 admissions (54.6% of all eligible admissions), 235,548 in the 2006 development set and 1,657,736 in the 7-year validation set. The final validation set (i.e., 2013-Q4) was smaller than the other validation set (n=37,442) as it was restricted to admissions beginning on or before December 1, 2013 to allow for sufficient follow-up time for outcome ascertainment. The remaining 27 validation sets consisted of 60,011 admissions on average (range 57,367 – 62,139). A brief summary of the patient population at select points across the study period is presented in Table 1. Admitted patients were primarily male (95.0%), white (72.1%), in their early 60s (mean age: 63.4; standard deviation: 14.0), and diagnosed with at least one chronic medical condition (93.9% diagnosed with one condition, 86.8% diagnosed with multiple conditions). Overall, the 30-day all-cause mortality rate after admission was 4.9%.

Initial performance of each of the seven models is presented in Table 2. Due to large sample sizes, the confidence intervals are narrow for all measures. Discrimination was generally good, with AUCs ranging from 0.768 to 0.847. The NN and NB models had slightly lower AUCs than the regression and RF models. The regression models and the NN model were calibrated based on both O:E ratios and ECIs. The RF model, with an O:E ratio of 0.929 (95% CI: 0.927, 0.931), slightly overpredicted on average. The NB model lacked calibration due to strong overfitting.

Table 2. Initial model performance with 95% confidence intervals

<table>
<thead>
<tr>
<th>Discrimination</th>
<th>LR</th>
<th>L1</th>
<th>L2</th>
<th>L1-L2</th>
<th>RF</th>
<th>NN</th>
<th>NB</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>AUC</strong></td>
<td>0.847</td>
<td>0.844</td>
<td>0.842</td>
<td>0.844</td>
<td>0.834</td>
<td>0.794</td>
<td>0.768</td>
</tr>
<tr>
<td></td>
<td>[0.846, 0.847]</td>
<td>[0.844, 0.844]</td>
<td>[0.841, 0.842]</td>
<td>[0.844, 0.844]</td>
<td>[0.833, 0.834]</td>
<td>[0.794, 0.795]</td>
<td>[0.768, 0.769]</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Calibration</th>
<th>LR</th>
<th>L1</th>
<th>L2</th>
<th>L1-L2</th>
<th>RF</th>
<th>NN</th>
<th>NB</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>O:E ratio</strong></td>
<td>0.998</td>
<td>0.998</td>
<td>0.998</td>
<td>0.998</td>
<td>0.929</td>
<td>0.997</td>
<td>0.997</td>
</tr>
<tr>
<td></td>
<td>[0.996, 1.001]</td>
<td>[0.996, 1.001]</td>
<td>[0.996, 1.001]</td>
<td>[0.996, 1.001]</td>
<td>[0.927, 0.931]</td>
<td>[0.993, 1.00]</td>
<td>[0.338, 0.340]</td>
</tr>
<tr>
<td><strong>Cox intercept</strong></td>
<td>-0.048</td>
<td>0.088</td>
<td>0.215</td>
<td>0.096</td>
<td>0.074</td>
<td>-0.122</td>
<td>-2.548</td>
</tr>
<tr>
<td></td>
<td>[-0.055, -0.042]</td>
<td>[0.080, 0.096]</td>
<td>[0.207, 0.223]</td>
<td>[0.088, 0.105]</td>
<td>[0.065, 0.082]</td>
<td>[-0.133, -0.112]</td>
<td>[-2.551, -2.546]</td>
</tr>
<tr>
<td><strong>Cox slope</strong></td>
<td>0.980</td>
<td>1.039</td>
<td>1.093</td>
<td>1.043</td>
<td>1.072</td>
<td>0.951</td>
<td>0.113</td>
</tr>
<tr>
<td></td>
<td>[0.977, 0.982]</td>
<td>[1.036, 1.042]</td>
<td>[1.090, 1.096]</td>
<td>[1.040, 1.046]</td>
<td>[1.069, 1.076]</td>
<td>[0.947, 0.955]</td>
<td>[0.113, 0.114]</td>
</tr>
<tr>
<td><strong>ECI</strong></td>
<td>0.013</td>
<td>0.010</td>
<td>0.011</td>
<td>0.010</td>
<td>0.034</td>
<td>0.008</td>
<td>7.783</td>
</tr>
<tr>
<td></td>
<td>[0.013, 0.014]</td>
<td>[0.010, 0.010]</td>
<td>[0.010, 0.011]</td>
<td>[0.010, 0.010]</td>
<td>[0.033, 0.034]</td>
<td>[0.007, 0.008]</td>
<td>[7.758, 7.808]</td>
</tr>
</tbody>
</table>

Model Performance Over Time

We observed stable discrimination for all models over the 7-year validation period (Bonferroni adjusted p<0.001; Figure 1). The regression and RF models had comparable AUCs and maintained higher discrimination than the NB and NN models, which exhibited similar levels of discrimination.

We observed calibration drift over time for all models, with the magnitude and pattern of drift varying by modeling method and calibration metric (Figure 2). The NB model substantially underperformed all other models in terms of calibration due to extreme predictions, and thus the calibration of this model is not considered further.

The O:E ratio (ideal value: 1) declined immediately after development and across the study period for all models, indicating increasing average overprediction. At the first 3-month validation period, O:E ratios included the ideal values of 1.0 for only the NN and LR models. These two models achieved calibration according to the O:E ratio at two to four additional time points, however, overpredicted on average for most of the validation period. The trajectory of the O:E ratio was similar for most models, with the exception of the NN model which did not exhibit a significant slope in the O:E ratio over time (Bonferroni adjusted p<0.001). The NN model demonstrated significantly less overprediction than the RF and regression models, particularly in the last three years of the validation period. In
addition to the overall drift, a seasonal pattern was apparent in the O:E ratios. In the first and fourth quarters of most validation years, O:E ratios peaked for all models.

The Cox logistic recalibration intercept (ideal value: 0) and slope (ideal value: 1) also noted different patterns of drift across modeling methods. The Cox intercept declined across the validation period for each of the regression models, indicating increasing overprediction, while remaining stable for the RF and NN models (Bonferroni adjusted p<0.001). The L2 model exhibited a smaller decline in the Cox intercept over time than the other regressions. The RF model did not systematically over or underpredict in most (24 of 28) validation periods, and the NN model systematically overpredicted to a stable degree for the entire study period. Additionally, we observed a seasonal pattern in the Cox intercept similar to that of the O:E ratio, although to a lesser degree. Cox slopes were stable over time (Bonferroni adjusted p<0.001). No significant overfitting was observed for the LR, L1, and L1-L2 regression models. The L2 regression and RF models exhibited some underfitting (i.e., Cox slope>1.0). This underfitting was consistent over time for the L2 regression model and demonstrated a nonsignificant tendency toward increasing for the RF model. The NN model had Cox slopes less than 1, indicating overfitting; however, there was no significant change over time.
Measuring calibration more stringently with flexible calibration curves, we observed drift among the regression and RF models, and stable overall calibration with some seasonal variation for the NN model. ECIs (ideal value: 0) of the regression and RF models increased across the validation period (Bonferroni adjusted p<0.001), indicating declining calibration, and exhibited no changes in the trajectory or rate of ECI drift during the seven validation years. The L2 regression model experienced a smaller magnitude of drift in the ECI compared to the other regression models and a similar magnitude of drift to the RF model. Compared to the L2 regression and RF models, the rate of change in ECI was 50% higher for the LR model (0.006 [95% CI: 0.005, 0.008] vs 0.004 [95% CI: 0.003, 0.005] and 0.004 [95% CI: 0.002, 0.005] for the LR vs L2 and RF, respectively) and 75% higher for the L1 and L1-L2 models (0.007 [95% CI: 0.006, 0.009] for both). For each model, seasonal corrections of the ECI were generally observed in the first and fourth quarters of each year. ECI was stable over time for the NN model (Bonferroni adjusted p<0.001). Although not significantly different from the surrounding time periods, in most validation years, the ECI of the NN model was markedly lower (i.e., closer to the ideal value of 0) during the first and fourth quarters of most years.

The ranges of predicted probabilities and proportion of admissions over which each model was calibrated also changed over time and varied by modeling method (see Figure 3). With the exception of the NB model, which strongly overpredicted for most predicted probabilities, each model moved in and out of regions of calibration, overprediction,
Figure 4. Model calibration over time, measured by regions of calibration scaled by proportion of observations in each region and shaded by the magnitude of the within region ECI.

and underprediction across the range of predicted probabilities. The RF model was the only model with a large range of probabilities over which it strongly underpredicted. During the first half of the validation period, the L2 regression model tended to be calibrated for predictions in the 50% to 90% ranges. The remaining models and the L2 regression model during the second half of the validation period tended to strongly overpredict for predicted probabilities starting at approximately 20%. For each regression model, across the validation period the proportional volume assessment (Figure 4) indicated that nearly half of admissions were in regions of overprediction, with the proportion increasing slowly over time. The majority of these admissions were minimally overpredicted, falling in areas with regional ECI values near the ideal value of 0, as highlighted by the lightest blue shades in Figure 4. For example, in the LR model, a low risk region of overprediction captured at least 40% of admissions in each 3-month validation period and had a mean ECI of 0.005 (range: 0.002 – 0.009). Each of the regression models also experienced growth in an overpredicted region with a larger magnitude of miscalibration among higher predicted probabilities. For the L1 and L1-L2 regression models, the vast majority of admissions were strongly overpredicted in the last year of the study period. The overpredicted region of the RF model captured a growing proportion of admissions over time and increased modestly in the magnitude of overprediction across the validation period. The proportional volume analysis highlighted a seasonal pattern in the calibration of the NN model for admissions with predicted probabilities under
3%. On average, 61.3% of all admissions were in this low risk region for which the NN model was calibrated during the first and fourth quarters of most years and minimally overpredicted during the second and third quarters.

Data Shifts in the Patient Population Over Time

Over the 7-year validation period, there was a statistically significant decline in the 30-day mortality rate (Bonferroni adjusted p<0.0003); however, this change was small, declining from 5.0% in the 2006 development year to 4.8% in the final validation period. Seasonal fluctuations within each validation year were three times larger than the overall change in the mortality rate (mean within year change: 0.6%; overall change: 0.2%). Compared to predictor variables, the outcome was relatively unimportant in RF membership models aimed at distinguishing validation and development observations. The variable importance rank of 30-day mortality declined from 59 to 63 out of 67 over the first validation year and was stable at a mean of 63 (range: 52 – 64) over the next 6 years.

We observed case mix shift across the validation period. Distributions of predicted probabilities generally indicated increasing severity and heterogeneity of risk in the patient population over time (Bonferroni adjusted p<0.002). Membership models also noted the presence of case mix shift, with the logistic membership models increasingly discriminating between admissions from the development and each sequential validation set as the AUC increased from 0.616 (95% CI: 0.613, 0.618) to 0.836 (95% CI: 0.834, 0.839) over the seven validation years. We observed changes in the distributions of 94.5% of predictors during the study period (Bonferroni adjusted p<0.0003). The proportion of admissions to black patients, that were planned, and that were unplanned but not readmissions did not change over time. In addition, the mean of the most recent blood urea nitrogen level was constant over. Changes in vital signs, laboratory values, and body mass index were generally small in magnitude. The forms of these changes were variable, with some having an inflection points at three to four years after model development. The largest changes were observed among the health history variables. With the exception of HIV, which declined by less than 0.5%, the proportion of admissions involving patients with each health condition increased across the validation period. The rates of these increases were generally constant over time. Among health history variables, the largest changes over time occurred for dyslipidemia (41.8% to 66.8%), fluid and electrolyte disorders (19.0% to 38.9%), and depression (20.1% to 39.9%), while the smallest change was observed for lymphoma (1.2% to 1.6%).

Changes in the structure of associations between predictors and 30-day mortality were measured by changes in the confidence intervals of the odds ratio from the original model based on 2006 data. However, a few variables exhibited a non-significant tendency toward strengthening or weakening associations. For some laboratory values, such as chloride and sodium levels during admission, odds ratios were less stable, moving in and out of significance in both directions of association. In L1 regression models refit over time, we observed significant temporal changes in selection patterns for two predictors (Bonferroni adjusted p<0.0004): history of liver disease and mean corpuscular hemoglobin concentration during the admission window. The frequency of selection for inclusion in L1 regression models began to decline starting approximately two years after development for both predictors. The magnitude of decline in selection frequency was larger for liver disease, which was selected in 98.0% of bootstrapped models at development and in 44.5% of models in the final validation period. Refitting RF models in each 3-month validation period revealed temporal changes in variable importance ranking for three predictors (Bonferroni adjusted p<0.0004): history of depression, history of dyslipidemia, and serum creatinine during the admission window. Ranks increased from 51 to 43 for depression and 29 to 17 for creatinine. The variable importance rank of dyslipidemia was steady through the first five years and then declined from 39 to 50 over the last two years of the validation period.

Discussion

In repeated validations over seven years, we observed varying patterns of performance among regression and machine learning models for 30-day all-cause mortality after hospital admission. Among all models, discrimination was stable over time. With the exception of the NN model, calibration drifted across the entire validation period as well as the other overpredicted risk. Seasonal changes in the mortality rate were correlated with cyclical fluctuations in the calibration of all models, including the NN model despite its maintaining stable calibration overall. Case mix shift dominated temporal changes in the patient population. Taken together with the observed calibration drift, the data shift assessments highlight robustness of the NN model, moderate susceptibility of the RF and L2 regression models, and high susceptibility of the other regression models to case mix shift.

Our findings of stable discrimination and increasing overprediction over time are consistent with previous studies of performance drift in logistic hospital mortality models [13, 14, 15-19]. While most prior work focused on logistic regression...
models, Minne et al\textsuperscript{14, 32} provide a comparison of calibration drift between corresponding logistic regression and tree-based rSAPS-II models for mortality among elderly ICU patients. Based on O:E ratios, the logistic rSAPS-II experienced increasing levels of overprediction within four years of development, while the tree-based rSAPS-II demonstrated stable calibration\textsuperscript{14, 32}. Although we did not observe differences in calibration drift between our logistic and RF models based on the O:E ratios, our RF model did experience a smaller magnitude of deterioration in the more stringent ECI metric compared to our logistic regression model.

Performance drift is driven by changes in the prevalence of an outcome, patient case mix, and associations between predictors and outcomes\textsuperscript{3, 5, 6, 13, 15, 31}. In our cohort, we detect limited evidence of predictor-outcome association shifts and observed primarily seasonal changes in the mortality rate. Seasonal event rate shift was correlated with a cyclical pattern in calibration. However, with the exception of the NN model, our models experienced calibration drift across the validation period that could not be explained by the seasonal variation in the mortality rate alone and is likely, therefore, associated with the documented changes in patient case mix. The stability of calibration of the NN model suggests this method is robust to case mix changes, at least to the extent seen in our cohort. The Cox recalibration model and flexible calibration curves indicated less deterioration in calibration over time for the L2 regression and RF models compared to the LR, L1, and L1-L2 regression models. The basic logistic regression and penalized regression methods that include variable selection, therefore, appear to be the most susceptible methods to calibration drift in the presence of case mix shift, while the L2 regression and RF models appear to be moderately susceptible, falling between the other regression approaches and the NN model. Additionally, we note that case mix shifts leading to changes in the variability of risk in the patient population have been previously noted to be of particular concern for discrimination drift, as patients become more difficult to distinguish or more easily separable as they become more homogenous or more heterogeneous, respectively\textsuperscript{31}. Although we observed increased heterogeneity of risk among admissions over time, we did not observe drifting discrimination. This may indicate that the degree of change in the variability of the risk in this cohort was not sufficient to trigger discrimination drift.

An important limitation in interpretation of modeling performance over time in real-world data is that different patterns of data shifts could result in differential model performance. We previously executed a separate study in which we explored calibration drift and data shifts in the clinical domain of hospital-acquired acute kidney injury\textsuperscript{33}. Diverging patterns of calibration drift between regression and machine learning methods were observed in both studies. Our current study extends and complements our previous findings by revealing disparate patterns of calibration drift across modeling methods in the presence of distinctive combinations of data shifts. For both clinical domains, the calibration of all models was susceptible to shifting event rates. While we did not observe predictor-outcome association shifts in our 30-day mortality cohort, association shifts in the acute kidney injury population temporally coincided with diverging patterns of calibration drift between regression and machine learning models. While case mix shift was observed in both study populations, the dominance of this form of data shift in our 30-day mortality population allows us to link case mix shift with differences in calibration drift between across modeling methods. Synthesizing findings across both studies, we found RF and NN models are generally less susceptible to case mix and association shifts than regression models, with the exception of the L2 regression model in the presence of case mix shift. Our findings in this work inform which methods may be most useful for modeling hospital mortality in changing environments and highlight both the variable updating needs of different modeling approaches and the influence of particular combinations of data shifts on our understanding of model susceptibility to deteriorating performance. As model performance patterns emerge among different data, understanding similarities in the assessed domains and a future domain of interest can assist in calibration maintenance.

Our findings have important implications for the integration of prediction models into clinical decision support tools and the design of model updating protocols. With the exception of the NN model, calibration drift impacted all models and began shortly after model development. Miscalibration of risk predictions can be harmful, potentially leading to inappropriate changes in treatment selection, anxiety, over-confidence, or inefficient allocation of limited clinical resources\textsuperscript{3, 7, 8, 13}. Lack of calibration may also mislead benchmarking assessments of facility quality\textsuperscript{14}. Calibration drift must, therefore, be addressed with routine model updating through recalibration, full model revision (i.e., refitting), or even model extension with the incorporation of new predictors\textsuperscript{3, 6, 34}. The form of population data shifts and performance drift should drive the timing and approach to model updating. We recommend implementation of active surveillance tools to track model performance and characterize changes in patient populations over time in order to trigger model updating and inform updating approaches. As our findings highlight, models based on different methods will have different updating requirements. We thus further recommend that the updating protocols built into such surveillance systems be tailored to modeling methodologies and support flexibility in timing of updates rather than being restricted to regularly scheduled intervals.
There were some limitations in this work. In this analysis, we consider model performance under a particular combination of data shifts; however, other combinations of data shifts in patient populations, including different forms and extents of shift, may expose other performance drift patterns. This was highlighted by the different combinations of data shifts observed in this study and prior work with AKI models\textsuperscript{33}. Studies in additional populations with diverse data shift scenarios or in simulated populations with defined forms and extents of data shift would provide a more nuanced understanding of the susceptibilities of various modeling methods. Additionally, we focused on a limited number of common modeling methods, and assessment of additional modeling techniques may reveal more patterns of performance drift. Finally, while we observed significant deterioration in the calibration of most models, we note that statistically significant miscalibration may not translate to clinically relevant changes in performance. Explorations of drift across methods using clinical utility metrics or extensions of calibration metrics to incorporate clinically acceptable margins of error would be particularly informative.

Conclusion

Predictive analytics providing tailored predictions at the individual-patient level are becoming increasingly feasible and commonplace. As we continue to identify and pursue opportunities to incorporate predictive analytics into clinical decision-making, we will require well-calibrated models that consistently deliver highly accurate predictions. In this study, we explore a set of methodologies that can be implemented as an ensemble for active surveillance of clinical prediction models in order to understand potential impacts in the choice of which modeling method to use and how trends in changes in clinical data impact calibration over time in order to make decisions about periodic modeling updating. For our models of 30-day all-cause mortality after hospital admission, case mix shift had little impact on the neural network model, moderate impact on the random forest and L-2 penalized logistic regression models, and a significant impact on most variations of logistic regression. In order to maintain the utility of and user-confidence in model predictions, routine maintenance of implemented clinical prediction models will be essential moving forward. We emphasize the importance of considering the long-term performance implications of modeling methods when developing, implementing, and updating clinical prediction tools. As best practice guidelines for model updating are developed, modeling methods must be a central consideration in order to promote efficient and effective strategies.

Acknowledgements

Financial support for this study was provided entirely by grants from the National Library of Medicine (ST15LM007450-15 and 1R21LM011664-01), the Veterans Health Administration (VA HSR&D IIR 11-292, VA HSR&D IIR 13-052).

References


Does Level of Numeracy and Graph Literacy Impact Comprehension of Quality Targets? Findings from a Survey of Home Care Nurses

Dawn W. Dowding, PhD, RN 1,2, David Russell, PhD 2, Karyn Jonas 3, Nicole Onorato 2, Yolanda Barrón, MS 2, RN Jaqueline A. Merrill, PhD 1, Robert J. Rosati, PhD 3
1Columbia University School of Nursing, New York, NY; 2Center for Home Care Policy & Research, Visiting Nurse Service of New York, New York, NY; 3New York University College of Nursing, New York, NY; 4Visiting Nurse Association Health Group, Holmdel, NJ

Abstract

Clinical dashboards that display targets compared to performance metrics are increasingly used by healthcare organizations in their quality improvement efforts. However, few studies have evaluated the extent to which healthcare professionals can readily understand and interpret these data. This study explored associations between measures of graph literacy and numeracy in home care nurses from two agencies (N=195) with comprehension of quality targets presented in a graphical dashboard format. Data were collected using an online survey. Results from linear regression models indicated that nurses’ levels of graph literacy and numeracy were positively associated with comprehension of quality targets. Nurses with low levels of both graph literacy and numeracy tended to have the lowest target comprehension scores compared to those who had high levels of both graph literacy and numeracy. Nurses with low graph literacy and high numeracy also had significantly lower scores for comprehension of quality targets compared to those with high graph literacy and numeracy. These findings suggest that developers of clinical dashboards that incorporate quality target information need to evaluate users’ ability to understand the information displayed in graphs and tables before they release the product for general use in healthcare settings.

Introduction

Clinical dashboards are a form of Health Information Technology (HIT) that use data visualization techniques to provide feedback to healthcare professionals on their performance compared to quality metrics (1). Dashboards are increasingly used by healthcare providers to enhance clinical productivity and track performance on quality improvement efforts. Most clinical dashboards aim to provide timely information to care providers in a succinct visual format. One feature that is often incorporated is the presentation of performance relative to quality targets or benchmarks. In the context of home healthcare services, quality targets might present clinical performance against outcome metrics like hospital readmission and emergent wound infections, which clinicians can use to benchmark their current performance against organization goals (2).

Despite the widespread use of dashboards in clinical settings, researchers have yet to explore the extent to which healthcare professionals can understand and interpret quality targets presented in a dashboard format. Two factors that may influence comprehension of quality targets include numeracy and graph literacy. Numeracy is defined as the ability to understand basic probability and math concepts (2). Graph literacy represents the ability to comprehend information displayed in graph formats (3). Few studies have explored numeracy and graph literacy among health care professionals in general or nursing specifically (3, 4). Those studies that have explored the impact of graph literacy with either the general population or patients have highlighted that individuals with high graph literacy have greater comprehension of information presented in a graphical format, and individuals with low graph literacy have greater understanding if the same information is presented in a number format (5-7). Studies exploring graph literacy and numeracy in physicians and nurses suggest that the type of graph used to display data interacts with graph literacy to impact data comprehension (3, 4). The present study builds upon earlier research by examining how numeracy and graph literacy skills are related to comprehension of quality targets presented in a graphical format among nurses sampled from two large regional home care agencies. We hypothesized that home care nurses with lower levels of graph literacy and numeracy would be less likely to comprehend graphical data presented to them about clinical quality targets.

Method

The study employed data collected via an on-line survey. Eligibility criteria included: 1) Registered Nurses (RN) who were currently employed at a Certified Home Health Agency (CHHA), and 2) visiting patients once or more per week. Nurses were recruited from two large, not-for-profit home health agencies located in the North East Region of the United States. All potential participants were sent an email invitation describing the purpose of the study and
containing a link to the survey website. Reminders to complete the survey were sent to all non-responders two, three and four weeks after the initial email. Participants first completed screening questions to ensure their eligibility to participate in the study, and if eligible they were directed to a demographic questionnaire followed by both the numeracy scale and graph literacy scale. The survey then randomized participants into one of four groups; each group received information in different graph formats (the results of this experimental study have been reported elsewhere). Finally all participants completed the questions related to comprehension of quality target information (the results reported in this paper). On completion of the survey nurses could provide their contact details (separate from survey responses) to receive a $20 gift voucher in recognition of their time. All study procedures were approved by the institutional review boards at Columbia University Medical Center and the Visiting Nurse Service of New York.

**Study Measures**

**Numeracy**

Numeracy was measured using the expanded numeracy scale developed by Lipkus et al (8). This is an 11-item scale that evaluates numerical skills such as familiarity with probability, the ability to convert percentages into proportions and risk magnitude. It is widely used as a method for assessing objective numerical competence, with high reliability (Cronbach α ranging from 0.7 – 0.75). The median numeracy score was used to categorize respondents with lower (<8) and higher numeracy levels (≥8).

**Graph Literacy**

Graph literacy was measured using a scale developed by Galesic and Garcia-Retamero (9) specifically for the health domain, to measure both basic and advanced graph reading skills and comprehension across different types of graphs (9). The scale consists of 13 items and measures 3 levels of graph comprehension: i) the ability to read the data (to find relationships in the data); and iii) the ability to read beyond the data (e.g., to be able to predict a future trend from a line chart). The scale was originally tested in two samples from Germany and the US, with high internal consistency (α = 0.74 and 0.79 respectively), and has been used previously to evaluate the graph literacy of nurses (9). The median graph literacy score was used to categorize respondents with lower (<9) and higher graph literacy levels (≥9).

**Comprehension of Quality Targets**

Comprehension of quality targets was assessed with four questions that asked nurses to interpret data presented in a table and graph relative to a quality target. The four questions assessing comprehension of quality targets were chosen based on their importance for home health nursing practice. Questions were developed to assess comprehension of quality targets related to the completion of a home health plan of care, visiting high risk home health patients within 24 hours of admission, assessment of hospitalization risk, and tracking the signs and symptoms of patients with chronic illness. An example of one quality target graph and corresponding comprehension question is shown in Figure 1. Binary variables were created to code correct responses and then summed to create a single measure of quality target comprehension that ranged from 0 (no correct responses) to 4 (all correct responses).

**Figure 1.** Example of Graph Displaying Quality Target and Corresponding Question Measuring Comprehension.
Demographic Characteristics

Nurse responders were asked about their demographic characteristics at the start of the survey. These characteristics included sex, age, race/ethnicity, education and nurse training level, number of years in a nursing role, number of years working at the agency from which they were sampled, and full-time or part-time status.

Data Analysis

Descriptive statistics were used to examine the demographic characteristics of the sample population, and chi-square or t-test were used to compare these characteristics between the two agencies. Linear regression models were used to estimate bivariate relationships of quality target comprehension with 1) graph literacy, 2) numeracy, and 3) a four-category interactive variable specifying combinations of graph literacy and numeracy. All data were analyzed using R (version 3.1.3).

Results

The survey was conducted between February 16th and May 26th 2016. Email invitations were sent to 1,052 nurses across the two agencies. 322 individuals accessed the survey (30.6 % response rate); of these 125 failed to complete all of the survey questions. Of the 125 respondents who did not complete the entire survey, 23 participants were disqualified because they did not meet all study inclusion criteria and 102 participants gave partial responses before dropping out of the survey. Of those with partial responses, 44 completed screening and/or demographic questions, 8 dropped out before completing the graph literacy scale, 38 dropped out before completing the numeracy scale, and 12 dropped out before completing the data comprehension questions. Overall we received a total of 195 completed responses (129 from Agency 1 and 66 from Agency 2). Reliability of the numeracy scale for our sample was α= 0.68 and for the graph literacy scale α=0.76. Similar to previous studies (9), the numeracy and graph literacy scales were strongly correlated (r=0.56; p<0.001), but with enough variance to indicate that they were measuring different concepts.

The characteristics of nurses who participated in the study are displayed in Table 1. Most nurses were female (89.7%), with an average age of 49 years (SD=11.0). The majority of nurses held a Bachelor’s (52.3%) or Post-Graduate degree (18.5%). Further, most nurses had a Bachelor’s (52.8%) or Post-Graduate nursing degree (11.8%). Nurse respondents differed in their average tenure at their respective agencies, with nurses surveyed at Agency 1 being employed for significantly (<0.001) longer on average (M=12.0; SD=7.7) compared to those surveyed at Agency 2 (M=5.5;
SD=6.9). Further nurses surveyed at Agency 1 tended to have higher levels of education compared to those at Agency 2 (<0.001).

There was no difference in mean numeracy or graph literacy between nurses at the two agencies; the mean numeracy score for the total sample was 8.4 (SD 2.0) and graph literacy was 9.7 (SD 2.4). The average quality target comprehension score was 3.4 (SD=0.8) out of 4 possible points; this score also did not differ between the two agencies.

**Table 1.** Characteristics of Study Participants

<table>
<thead>
<tr>
<th></th>
<th>ALL N=195</th>
<th>AGENCY 1 N=129</th>
<th>AGENCY 2 N=66</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female Sex (%)</td>
<td>175 (89.7)</td>
<td>115 (89.2)</td>
<td>60 (90.9)</td>
<td>0.70</td>
</tr>
<tr>
<td>Mean Age in Years (SD)</td>
<td>49.0 (11.0)</td>
<td>48.3 (11.0)</td>
<td>50.5 (11.0)</td>
<td>0.19</td>
</tr>
<tr>
<td>Race/Ethnicity (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>American Indian/Alaska native</td>
<td>1 (0.5)</td>
<td>0</td>
<td>1 (1.5)</td>
<td></td>
</tr>
<tr>
<td>Asian</td>
<td>28 (14.4)</td>
<td>23 (17.8)</td>
<td>5 (7.6)</td>
<td></td>
</tr>
<tr>
<td>African American/Black</td>
<td>46 (23.6)</td>
<td>33 (25.6)</td>
<td>13 (19.7)</td>
<td></td>
</tr>
<tr>
<td>Native Hawaiian/Other Pacific Islander</td>
<td>2 (1.0)</td>
<td>2 (1.6)</td>
<td>0 (0.0)</td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>97 (49.7)</td>
<td>54 (41.8)</td>
<td>43 (65.2)</td>
<td></td>
</tr>
<tr>
<td>More than one race</td>
<td>11 (5.6)</td>
<td>9 (7)</td>
<td>2 (3.0)</td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>10 (5.1)</td>
<td>8 (6.2)</td>
<td>2 (3.0)</td>
<td></td>
</tr>
<tr>
<td>Education</td>
<td></td>
<td>&lt;0.001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Associate/Diploma Degree (%)</td>
<td>57 (29.2)</td>
<td>27 (20.9)</td>
<td>30 (45.5)</td>
<td></td>
</tr>
<tr>
<td>Bachelor’s Degree (%)</td>
<td>102 (52.3)</td>
<td>70 (54.3)</td>
<td>32 (48.5)</td>
<td></td>
</tr>
<tr>
<td>Post Graduate Degree (%)</td>
<td>36 (18.5)</td>
<td>32 (24.8)</td>
<td>4 (11.1)</td>
<td></td>
</tr>
<tr>
<td>Nursing Training Level</td>
<td></td>
<td>&lt;0.001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pre Bachelor’s (%)</td>
<td>69 (35.4)</td>
<td>34 (26.4)</td>
<td>35 (53.0)</td>
<td></td>
</tr>
<tr>
<td>Bachelor’s (%)</td>
<td>103 (52.8)</td>
<td>75 (58.1)</td>
<td>28 (42.4)</td>
<td></td>
</tr>
<tr>
<td>Post Graduate (%)</td>
<td>23 (11.8)</td>
<td>20 (15.5)</td>
<td>3 (4.6)</td>
<td></td>
</tr>
<tr>
<td>Mean Years as a Nurse (SD)</td>
<td>19.7 (11.4)</td>
<td>20.2 (11.4)</td>
<td>18.7 (11.4)</td>
<td>0.39</td>
</tr>
<tr>
<td>Agency Tenure (Years) Mean (SD)</td>
<td>9.8 (8.1)</td>
<td>12.0 (7.7)</td>
<td>5.5 (6.9)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Staff Nurse (vs. Per diem) (%)</td>
<td>150 (76.9)</td>
<td>101 (78.3)</td>
<td>49 (74.2)</td>
<td>0.53</td>
</tr>
<tr>
<td>Mean Total Numeracy Score (SD)</td>
<td>8.4 (2.0)</td>
<td>8.4 (2.0)</td>
<td>8.5 (2.0)</td>
<td>0.86</td>
</tr>
<tr>
<td>Mean Graph Literacy Score (SD)</td>
<td>9.7 (2.4)</td>
<td>9.5 (2.5)</td>
<td>9.9 (2.2)</td>
<td>0.34</td>
</tr>
<tr>
<td>Mean Target Comprehension Score (SD)</td>
<td>3.4 (0.8)</td>
<td>3.4 (0.7)</td>
<td>3.4 (0.9)</td>
<td>0.46</td>
</tr>
</tbody>
</table>

In bivariate and multivariate analyses mean graph literacy did not differ by sex, age, race (white vs. non-white), education level, nurse training level, whether the nurse was staff or per-diem, agency, experience (years as a nurse), or agency tenure. Mean numeracy was slightly and inversely associated with years as a nurse (each year of additional nursing experience was associated with a 0.03 lower numeracy score (p=0.014) in bivariate analyses. After adjusting for all demographic characteristics in multivariate analyses, years as a nurse was no longer a significant predictor of numeracy level (p=0.08).

We categorized nurses as having either high (score ≥ 8; N=147, 75%) or low (score < 8; N=48, 25%) numeracy and high (score ≥ 9; N=146, 75%) or low graph literacy (score < 9; N=49, 25%). Nurses were then further categorized into four groups based on their level of numeracy and graph literacy: 1) low numeracy and low graph literacy (n=24; 12%), 2) high numeracy and low graph literacy (n=24; 12%), 3) low numeracy and high graph literacy (n=25; 13%) and 4) high numeracy and high graph literacy (n=122; 63%).

Table 2 presents the results of three linear regression models examining associations between graph literacy and numeracy with comprehension of quality targets presented in a graphical format. Findings from Models 1 and 2 suggest that both high graph literacy and high numeracy, respectively, are positively and significantly associated with level of quality target comprehension. Model 3 explores the relationship of a categorical variable representing the interaction between graph literacy and numeracy with level of quality target comprehension. Results from this model suggest that, compared with nurses who scored as having both low graph literacy and low numeracy, greater comprehension of quality targets is associated with the combination of low graph literacy and high numeracy, high graph literacy and low numeracy, and high graph literacy and high numeracy.
Table 2. Linear Regression of Quality Target Comprehension on Graph Literacy and Numeracy Scores (N=195)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Model 1 Est (SE)</th>
<th>Model 2 Est (SE)</th>
<th>Model 3 Est (SE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>High Graph Literacy</td>
<td>0.68 (0.13)***</td>
<td>0.40 (0.13)**</td>
<td>1.00 (0.17)***</td>
</tr>
<tr>
<td>High Numeracy</td>
<td>1.06 (0.22)***</td>
<td>0.67 (0.22)**</td>
<td></td>
</tr>
<tr>
<td>Graph Literacy and Numeracy Group</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High Graph Literacy and High Numeracy</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High Graph Literacy and Low Numeracy</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low Graph Literacy and High Numeracy</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low Graph Literacy and Low Numeracy</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model Intercept</td>
<td>2.88 (0.11)</td>
<td>3.08 (0.11)</td>
<td>3.54 (0.68)</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.13</td>
<td>0.05</td>
<td>0.17</td>
</tr>
</tbody>
</table>

NOTE: * $p<0.05$; ** $p<0.01$; *** $p<0.001$

Conclusion

This study identifies the importance of numeracy and graph literacy levels among home care nurses for their comprehension of quality targets presented in graphical displays—a format that is increasingly used by healthcare organizations to share performance feedback with staff. We found variation in both numeracy and graph literacy in our sample population. Our sample of nurses had higher average numeracy and graph literacy skills than that identified in the wider US population (9). Compared to previous studies of nurses, respondents in our study scored higher on measures of numeracy and graph literacy (10). The findings also highlight the importance of considering numeracy and graph literacy when evaluating data presentation formats. Our results suggest that individuals with low numeracy and low graph literacy are less likely to comprehend quality targets compared with those who have high numeracy and high graph literacy. These findings may help to clarify prior research on the impact of clinical dashboards on patient outcomes, by raising the possibility that comprehension of quality targets is an additional factor that may be related to clinical practice (11). Individuals who have low numeracy and graph literacy may have difficulty effectively comprehending quality targets and other data presented in dashboards, potentially influencing their ability to make decisions related to patient care.

The findings of our study have important implications for the design of clinical dashboards that include quality targets. In general, there is an assumption that the presentation of data in a visualized format will improve data comprehension, through reduction in cognitive workload associated with processing that information (12, 13). However, our findings would suggest that consideration also needs to be given to individual variation in clinicians’ ability to understand that information. Although individuals with high graph literacy and numeracy may be able to comprehend quality targets presented in graphical formats, those with low graph literacy or numeracy may not. One unexpected finding was the lack of any significant relationships between demographic characteristics and graph literacy or numeracy level.

Strengths of our study include a large home care nurse sample recruited from two health care agencies, with variation in age, experience, and race/ethnicity, the use of a validated objective measure of numeracy rather than subjective measures used in prior research (4, 10), and a validated measure of graph literacy. Study limitations included that the sample was self-selected by choosing to respond to the survey invite. This sample excludes individuals who started but did not complete the survey. Analysis of partial responses to the survey revealed that those who did not complete the full survey tended to have significantly lower scores on the graph literacy scale, suggesting that they may have been more challenged by the questionnaire and decided not to continue with the survey.

As healthcare organizations expand use of complex data presentations for key performance indicators, the ability of end-users to comprehend that data will represent an important linchpin in efforts to improve healthcare quality. Further research is needed to explore numeracy and graph literacy in nursing populations across health care settings and to investigate the influence of these factors on clinical practice and patient outcomes. In addition, future research should investigate how best to provide support to clinicians to enable them to comprehend dashboards that include data visualizations such as quality targets. Researchers may also seek to examine whether providers in
References

Detection of Suicidality in Adolescents with Autism Spectrum Disorders: Developing a Natural Language Processing Approach for Use in Electronic Health Records

Johnny Downs, MRCPsych*1,2, Sumithra Velupillai*, PhD1,3, George Gkotsis*, PhD1, Rachel Holden, MSc1,4, Maxim Kikoler, MSc1,4, Harry Dean, MSc1, Andrea Fernandes, MSc1, Rina Dutta, PhD, FRCPsych1,2

1Department of Psychological Medicine, NIHR Biomedical Research Centre, Institute of Psychiatry, Psychology and Neuroscience, King’s College London, London, UK; 2South London and Maudsley NHS Foundation Trust, London, UK; 3School of Computer Science and Communication, KTH, Stockholm; 4University of Canterbury, Southborough, UK.

* authors contributed equally to this work

Abstract

Over 15% of young people with autism spectrum disorders (ASD) will contemplate or attempt suicide during adolescence. Yet, there is limited evidence concerning risk factors for suicidality in childhood ASD. Electronic health records (EHRs) can be used to create retrospective clinical cohort data for large samples of children with ASD. However, systems to accurately extract suicidality-related concepts need to be developed so that putative models of suicide risk in ASD can be explored. We present a systematic approach to 1) adapt Natural Language Processing (NLP) solutions to screen with high sensitivity for reference to suicidal constructs in a large clinical ASD EHR corpus (230,465 documents), and 2) evaluate within a screened subset of 500 patients, the performance of an NLP classification tool for positive and negated suicidal mentions within clinical text. When evaluated, the NLP classification tool showed high system performance for positive suicidality with precision, recall, and F1 scores all > 0.85 at a document and patient level. The application therefore provides accurate output for epidemiological research into the factors contributing to the onset and recurrence of suicidality, and potential utility within clinical settings as an automated surveillance or risk prediction tool for specialist ASD services.

Introduction

Over 1 in 6 young people with autism spectrum disorders (ASD) will contemplate or attempt suicide during childhood, making them 30 times more at risk than typically developing children.1,2 Why children with ASD have higher rates of suicidal behaviours is unclear. It is possible that risk factors for childhood suicidal behaviour found in typically developing children, such as depression or being bullied, are more prevalent or potentially have a greater negative impact in children with ASD.1 However, very little work has been conducted in ASD cohorts, and findings derived from non-ASD samples cannot be assumed to generalise to children with ASD.1 A growing number of studies have shown that putative risk factors (both environmental and genetic) for psychiatric outcomes can have different effects in children with neurodevelopmental disorders.4,5 Therefore individuals with ASD may express and manifest suicidal tendencies and behaviours in ways that differ from those observed in typical development.9

Given the widespread adoption of Electronic Health Records (EHRs) in primary and hospital care systems and the rapid growth of health informatics capabilities, longitudinal data from large samples of children with ASD can be used to develop and test new models of suicide risk behaviour. There is considerable potential to adapt EHR research methodologies used in recent epidemiological and risk factor studies1 and apply these approaches to address the evidence gap in ASD and other vulnerable adolescent groups. Although to capitalise on these developments for suicide research, accurate EHR data extraction systems need to be developed to capture data on those young people with ASD who present to public health services with suicidal thoughts or behaviours.
Information about suicidality in clinical documents is predominantly written in free-text. Haerian et al. showed that using only ICD-9 E-codes to detect patient-level suicide and suicide ideation from clinical text had the lowest positive predictive value (PPV): 0.55, while a combination of codes and Natural Language Processing (NLP) had the highest: 0.97, when applied on EHRs from the New York Presbyterian Hospital/Columbia University Medical Center. They used MedLEE (Medical Language Extraction and Encoding System) to generate Concept Unique Identifiers (CUIs) related to suicidality, and to filter out negated mentions as well as mentions not related to the patient. Anderson et al. applied a rule-based NLP approach to identify positive or negated mentions related to suicidality in the History of Present Illness (HPI) section of EHRs from a distributed health network of primary care organizations in the US, and found that suicidality information was predominantly recorded in free-text.

Because suicidality is routinely assessed in mental health care, the absence or negation of suicidal behaviour is also documented in EHRs. An NLP tool developed specifically for detecting negated mentions of suicide in mental health records using syntactic tree information has been developed in our group with high accuracy (91.9%) when evaluated on 6,000 sentences from mental health EHRs. The tool classifies each target mention (e.g. suicid*) as negated or positive using a set of negation terms and rules applied on the information provided from a syntactic representation (constituency-based parse tree) of the sentence in which the target term is found, using the Stanford Core NLP toolkit for preprocessing and syntactic representation.

The aim of this study was to extend, further develop and robustly evaluate an NLP approach which could accurately identify suicidality in ASD-patients’ clinical records, with the future goal that it may provide data to enable improved risk prediction for related major adverse events, such as suicide attempts. Using EHR documents, such as progress notes, risk assessments and medical correspondence, we examined whether negation detection methods could be used to accurately identify references to suicidality in the EHRs of adolescents with ASD presenting to clinical mental health services. We defined suicidality as either the reporting of the intention to engage in a potentially lethal act towards oneself, or undertaking such acts themselves. To achieve the study aim, we developed coding rules using expert consensus, to define explicit suicidality-related mentions for adolescents with ASD seen in specialist mental health clinics (inpatient and ambulatory). Based on these rules we extended our NLP tool to 1) identify documents containing suicide-related (SR) information (i.e. NLP tool to screen documents) and 2) identify positive and negated references of suicidality on a document and patient level (i.e. NLP to classify SR documents and patients as positive, SR-Pos, or negative, SR-Neg) across a large number of EHRs. We then compared the performance of the NLP tool against expert human-rater case note reviews.

Materials and Methods

Data resources

This study used data extracted from the anonymised, electronic clinical records of a sample of adolescents with ASD referred to South London and Maudsley NHS Foundation Trust (SlaM). This sample and clinical setting has been described elsewhere, but in brief SlaM provides specialist inpatient and outpatient ASD assessment and treatment services for young people from across the UK. Children and adolescents in this study were referred from primary care, child health, and educational and social care services, and typically underwent a multidisciplinary assessment by Child and Adolescent Mental Health Service (CAMHS) clinicians. Primary and secondary psychiatric disorders were diagnosed by CAMHS using the International Classification of Diseases, 10th Revision (ICD-10) multi-axial classification system.

The Clinical Record Interactive Search (CRIS) system was used to produce an anonymised EHR dataset to search on structured data and free text fields for all ASD patients. CRIS was established in 2008 to allow searching and retrieval of full but de-identified clinical information for research purposes with a permission of secondary data analysis, approved by the Oxfordshire Research Ethics Committee C (reference 08/H0606/71+5). The patients were part of an open clinical cohort (entering and leaving the study at different time points) and included children aged 3–17 years with a diagnosis of ASD (ICD-10 F84.0, F84.1, F84.5, F84.9) recorded between 1 January 2008 and 31 December 2013. Free text entries, correspondence and reports were available for this sample from their initial assessment until June 2016. The resulting cohort contained 3,642 unique patients (complete age range). For the purposes of this study, we selected the sample of adolescents who had at least one contact with CAMHS (i.e one free text document in CRIS) between the ages 14 and 18 years, totalling 1,906 patients.
Overall workflow

Figure 1 outlines the overall workflow of our study. There were three main phases. The first phase related to the definition of classification rules to identify suicidality-related information in EHR documents for adolescents with ASD (step 1 below). These rules were then applied in the second phase where a manual review of documents (step 2) was used to inform the development and evaluation of our NLP approach to screen for suicidality-related (SR) mentions in documents and filtering out documents with no mentions related to suicidality (NSR) – step 3 below. The NLP approach was then used to extract SR documents for the third phase (step 4). In the third phase, a manual review of documents was performed to annotate mentions of suicidality in SR documents as positive (SR-Pos), negative (SR-Neg) or uncertain (SR-U), step 5. Finally, the NLP approach was evaluated for its ability to correctly classify SR-Pos or SR-Neg in these documents and patients, step 6.

Step 1: Development of a set of classification rules to identify suicidality in adolescents with ASD.

Senior clinicians with expertise in the clinical management of neurodevelopmental disorders (JD) and suicidality assessment (RD) developed a set of rules to classify explicit mentions of suicidality in every document as either positive, negated or unknown. Positive mentions included text that referred to previous attempts, the presence of current or past plans of suicidal acts, command hallucinations related to carrying out a suicide attempt, a desire to be dead, researching methods, having ideas or describing plans of how to end their life or, a clinical opinion of the young person being at an elevated risk of attempting suicide. Negated terms included clinical opinions of the young person not being at elevated risk of suicide, recorded denial of suicidality by the young person (either directly or via third person report). Mentions were classified as uncertain, when aspects of suicidality were referred to, but did not appear to relate to risk of the young person being suicidal, for example references to dreams of being dead, or joking about death, or when references to suicidality were about other people (e.g. family members or friends).

Figure 1. Overall workflow of the study.

Step 2: Manual review of suicidality-related (SR) information and NLP screening tool development.

A randomly extracted subset of 100 patients and their corresponding documents were allocated to a training corpus, and another random selection of 100 patients was allocated to the test corpus. To generate a subset of patients with a
reasonable amount of documentation for manual review, the random sample was extracted based on documentation prevalence: each included patient had at least 7 documents (1st quartile) and at most 50 (3rd quartile), yielding a total of 2,445 (training set) and 2,433 (test set) documents in total. One clinically trained annotator (HD) was given these documents grouped for each patient. The annotator reviewed all documents for each patient, marked suicidality-related (SR) expressions, and labelled each SR-expression as either positive, negated or uncertain, according to the rules developed in step 1. However, for this phase, only annotations for SR information (regardless of polarity) were used for analysis.

**Step 3**: Extension and provisional assessment of the NLP approach for SR screening.

Results from the manual review were used to extend the NLP approach with the addition of new explicit SR expressions. Given the low frequency of the positive or negated SR mentions within the training set, we used the test set to assess precision, recall and F1-score of the tool detecting any SR content (regardless of polarity). Because the end goal is to address overall suicidality risk behaviour, the approach was evaluated on a document and patient level rather than on the mention level.

**Step 4**: NLP tool deployed to screen for SR documents

The NLP tool was then deployed to filter out documents without any SR mentions (positive or negative) from the original cohort (excluding the already annotated 200 patients). From 1706 patients (225,577 documents), 890 (52.2%) patients had at least one SR document, resulting in a total of 10,749 documents.

**Step 5**: Manual review of SR subset for identification of positive (SR-Pos) and negative (SR-Neg) suicidality mentions

Two manual coders (RH and MK) were randomly assigned 500 (56.2%) patients from the SR subset. Each annotator was given all documents (for each patient) that were detected by the NLP tool as containing a SR mention. The annotators were not given the NLP system output, but instead were asked to annotate explicit mentions of suicidality (same as above) and label these as positive (SR-Pos), negated (SR-Neg) or uncertain (SR-U). The documents were given to the annotators on a per-patient basis, and each patient was reviewed by one annotator. A subset (n=100) of randomly extracted documents was also used to calculate inter-rater agreement (measured with Cohen's κ and F1-score) on a document-level.

A majority rule was applied when evaluating document-level agreement: all mention-level annotations in each document were first counted, then, if the number of annotations labelled as positive for suicidality outnumbered or equalled the number of annotations labelled as negated, the document-level label was assigned “Positive for suicidality” – SR-Pos, otherwise it was designated “Negated for suicidality” – SR-Neg. To evaluate patient-level performance, priority was given to document-level outcomes: if the patient had at least one document labelled as SR-Pos using the majority rule, the patient-level label was assigned SR-Pos, irrespective of the number of previous or subsequent documents labelled as SR-Neg, i.e. each patient only required a single ‘positive suicidality’ document to be labelled SR-Pos.

**Step 6**: Final, comprehensive evaluation: NLP SR-Pos/SR-Neg classification

As a final step, the NLP approach was evaluated with precision, recall and F1-score against the manual annotations of the larger, filtered set of documents/patients with SR-Pos and SR-Neg labels, using the same heuristics for document- and patient-level classification assignments as above. Note that the evaluation is only performed on these two labels, i.e. SR-U annotations are not mapped to SR-Pos or SR-Neg. Thus, a false positive or false negative from the NLP approach could be due to an annotation marked as SR-U. A manual error analysis on cases of disagreements between the NLP tool and human annotation labels was also performed to gain a deeper understanding of the results.

**Results**

Table 1 shows the distribution of SR and non-SR (NSR) documentation and the individual level prevalence amongst the 100 adolescent patients with ASD in the final training set and the 100 patients in the test set. Manual review of both training and test documents revealed that only a small proportion of the corpus contained any SR information: <3% at the document level and around 22% at the patient level, with a similar distribution in the training and test.
set. Precision, recall, and F1 scores showed high system performance (> 0.8) for both SR and NSR in the test set (table 1).

The lexical markers of suicidality that were added to the NLP tool included *kill himself/herself/themselves/myself, end his/her/their life, take his/her/their own life, want to die, were dead*. Note that the NLP tool relies on lemmatised forms in both target expressions and document surface forms in order to achieve a more robust matching, e.g. different verb inflections of *want* will be matched with this approach.

**Table 1.** Confusion matrix: Screening for suicidality (SR) or non-suicidality (NSR), NLP tool compared to human annotation (A).

<table>
<thead>
<tr>
<th>NLP (Training)</th>
<th>Documents</th>
<th>Patients</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NSR</td>
<td>SR</td>
</tr>
<tr>
<td>NSR</td>
<td>2374</td>
<td>10</td>
</tr>
<tr>
<td>SR</td>
<td>5</td>
<td>56</td>
</tr>
<tr>
<td></td>
<td>2379</td>
<td>66</td>
</tr>
<tr>
<td>Precision</td>
<td>0.99</td>
<td>0.85</td>
</tr>
<tr>
<td>Recall</td>
<td>0.99</td>
<td>0.91</td>
</tr>
<tr>
<td>F1</td>
<td>0.99</td>
<td>0.88</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>NLP (Test)</th>
<th>Documents</th>
<th>Patients</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NSR</td>
<td>SR</td>
</tr>
<tr>
<td>NSR</td>
<td>2356</td>
<td>13</td>
</tr>
<tr>
<td>SR</td>
<td>8</td>
<td>56</td>
</tr>
<tr>
<td></td>
<td>2365</td>
<td>69</td>
</tr>
<tr>
<td>Precision</td>
<td>0.99</td>
<td>0.81</td>
</tr>
<tr>
<td>Recall</td>
<td>0.99</td>
<td>0.88</td>
</tr>
<tr>
<td>F1</td>
<td>0.99</td>
<td>0.84</td>
</tr>
</tbody>
</table>

Table 2 shows the distribution of negated and positive suicidality-related information (SR-Pos/SR-Neg) using our majority rule criteria in 4,911 pre-screened documents derived from 500 patients. Evaluation of the NLP tool (table 2) showed high system performance for SR-Pos with precision, recall, and F1 scores all > 0.83 at a document and patient level. SR-Neg performance measures were lower, especially in recall (0.75 on document level, 0.62 on patient level), but overall good levels of classification were produced (F1 = 0.79 on document level, 0.72 on patient level).

A manual error analysis on a random sample of ten documents where the NLP tool classified a document as positive (SR-Pos) but the human annotator as negated (SR-Neg) was performed to gain a deeper understanding of the reasons behind the lower recall results. The main themes involved:

1) classification of documents with only one suicide-related mention (annotator SR-Neg count = 1, NLP SR-Pos count = 1) due to missing negation term, e.g. ‘*Nil suicidal*’ or error in syntactic parsing due to e.g. badly formatted sentences.

2) cases where the majority heuristic is problematic and negation detection scope is erroneous, e.g. one document annotated with SR-Neg = 2, while the NLP output was: SR-Neg = 1, SR-Pos = 3 contained the following: ‘*XXX denied any recent sleep difficulties, excessive fatigue or guilt, changes in appetite or morbid or suicidal ideation*’, ‘*The risk of suicide is low, XXX denies suicidal ideation.*’

3) co-reference in combination with majority heuristics (annotator SR-Neg = 3, SR-Pos = 2, NLP SR-Neg = 1, SR-Pos = 2): ‘*XXX reported that XXX has had suicidal thoughts in the past but has no current plans on acting on them, co-reference*’ (sentence repeated twice in document), ‘*[clinician reporting] further stated that no evidence of psychosis, self-harming behaviour, suicidal thoughts, sleep or appetite …*’

4) clinically challenging cases and complex information given in the document. Three examples are:
Annotator: SR-U = 4, SR-Neg = 3, SR-Pos = 1; NLP tool output: SR-Neg = 1, SR-Pos = 2. ‘... reported fleeting suicidal thoughts... but strongly denied that XXX could act on these...’, ‘presenting for serious OD without suicidal...’, ‘firm denial of suicidal...’, ‘overdose’ marked as uncertain by annotator.

Annotator: SR-U = 2, SR-Neg = 1, NLP tool output: SR-Neg = 2, SR-Pos = 3 included sentences with information reported by external authorities such as the health care team and the school, references to the past, and includes a conclusive statement towards the end of the document: ‘we could not assess negated suicidal ideation as XXX left the room’, ‘unable to assess negated suicidal ideation’; ‘historically past has threatened self harm and disclosed suicidal ideation...’, ‘concerns from school about suicidal ideation’; ‘no suicidal ideation expressed’.

Annotator: SR-U = 2, SR-Neg = 1, NLP tool output: SR-Pos = 2: ‘I tried to assess XXX’s suicidal risk - XXX does not know if XXX wants to kill self’, ‘XXX does not have any specific plan’

Table 2. Confusion Matrix: Classification of positive and negative SR, document- and patient level assessments. SR-Neg = Suicidality-related (SR) mention is negated (Neg), SR-Pos = Suicidality-related mention is positive (Pos).

<table>
<thead>
<tr>
<th>Annotator</th>
<th>SR-Neg</th>
<th>SR-Pos</th>
<th>∑</th>
<th>SR-Neg</th>
<th>SR-Pos</th>
<th>∑</th>
</tr>
</thead>
<tbody>
<tr>
<td>SR-Neg</td>
<td>1379</td>
<td>463</td>
<td>1842</td>
<td>81</td>
<td>50</td>
<td>131</td>
</tr>
<tr>
<td>SR-Pos</td>
<td>273</td>
<td>2796</td>
<td>3069</td>
<td>14</td>
<td>355</td>
<td>369</td>
</tr>
<tr>
<td>∑</td>
<td>1652</td>
<td>3259</td>
<td>4911</td>
<td>95</td>
<td>405</td>
<td>500</td>
</tr>
<tr>
<td>Precision</td>
<td>0.83</td>
<td>0.86</td>
<td>0.85</td>
<td>0.87</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Recall</td>
<td>0.75</td>
<td>0.91</td>
<td>0.62</td>
<td>0.96</td>
<td></td>
<td></td>
</tr>
<tr>
<td>F1</td>
<td>0.79</td>
<td>0.88</td>
<td>0.72</td>
<td>0.92</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

In total, 100 random documents were double annotated (table 3). A document-level assessment using the majority rule yielded an average Cohen’s κ of 0.83, F1-scores for SR-Neg and SR-Pos document-level assessment were 0.89 and 0.94 respectively, indicating high agreement.

Table 3. Confusion Matrix: Inter-Rater Agreement on document level. SR-Neg = Suicidality-related (SR) mention is negated (Neg), SR-Pos = Suicidality-related mention is positive (Pos).

<table>
<thead>
<tr>
<th>Annotator 1</th>
<th>SR-Neg</th>
<th>SR-Pos</th>
<th>∑</th>
</tr>
</thead>
<tbody>
<tr>
<td>SR-Neg</td>
<td>32</td>
<td>2</td>
<td>34</td>
</tr>
<tr>
<td>SR-Pos</td>
<td>6</td>
<td>60</td>
<td>66</td>
</tr>
<tr>
<td>∑</td>
<td>38</td>
<td>62</td>
<td>100</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Annotator 2</th>
<th>SR-Neg</th>
<th>SR-Pos</th>
<th>∑</th>
</tr>
</thead>
</table>
Discussion and Conclusion

This is the first study to demonstrate that an NLP tool can be used to accurately capture a clinical construct as complex as suicidality within health records of young people with ASD. Our NLP tool identified suicidality-related (SR) mentions with high degrees of precision and recall from clinical free text held within EHRs. This NLP application provides powerful opportunities for surveillance work in adolescent ASD and in other clinical samples, with the potential to improve risk prediction for major adverse events, such as suicide attempts.

The development of this high performance NLP tool was achieved in several steps. First, owing to the potentially distinctive characteristics of the ASD clinical population, and their specialist mental health service provision, we began by building a suicidality terminology from a detailed note review of over 2000 random sets of clinical entries in 100 children with ASD, combined with expert clinical consensus. Because of the limited literature on suicidal terminology in ASD, we used a randomly extracted training and test set from all potential ASD EHR source data, rather than an enriched set filtered by restricted terms (e.g. “suicid” or ICD coding classifications). The rationale for this was to reduce selection bias and loss of sensitivity through the use of training and test data derived using restricted terms or coding classifications.

Random selection from the whole potential corpus also provided us with a better understanding of the overall distribution of suicidality-related information in documents and allowed us to refine and advise on additional terminology. During the training phases, it became clear that there was a low frequency of SR terms (less than 3% of all documents). A much larger corpus was required to conduct an adequate test of the NLP tool’s classification performance in discerning positive and negated SR mentions within the documents. We therefore configured the NLP tool to provide a SR screening step and identify a smaller, more feasible volume of suicidality-related documents for human annotation from the whole ASD corpus of c. 220,000 documents. We then tested the positive predictive value (precision) and sensitivity (recall) of the NLP screening tool against the EHR of another 100 children (approximately 2000 human rated documents) which were annotated for any references to suicidality. Finally, we began a comprehensive test of the accuracy of the tool against positive (SR-Pos) or negated (SR-Neg) mentions of suicide within the screened suicide related documents.

The abstraction of mention-level annotations and NLP system predictions to document- and patient-level assessments using simple heuristics (majority rule for document level and SR-Pos priority on patient level) showed that promising results can be obtained even though the NLP tool relies only on a relatively small number of suicidality-related and negation terms. This finding also shows that even though suicidality behaviour is documented with a variety of expressions (e.g. ‘took an excessive amount of pills’, ‘threw him/herself in front of a train’), indicative terms (mainly suicide in different forms) are typically also used at some point in the documentation and will thus be detected automatically.

Our motivation for applying a majority rule on document level assessments was based on the finding that the main source for false positive errors in our negation detection approach stemmed from cases of question forms (e.g. ‘I asked him if he feels suicidal’), references to the past, etc. Applying this rule was a way of smoothing this error rate. However, the error analysis showed that this approach might be a limitation. In future studies, we aim to compare results with NLP approaches such as ConText where variables relating to the past (‘historicity’) and subject (‘experiencer’) are encoded with target terms. We also aim to experiment with other abstraction heuristics, e.g. instead of majority rule, applying a priority hierarchy. In keeping with prior work, another alternative could be to define the annotation task on a document level. Longer term, we aim to compare the predictive validity of different heuristics within our NLP tool, and across other NLP approaches, for later adverse outcomes (i.e. significant suicide attempts or death by suicide), and seek external validity through replication in other EHR systems. Without these further steps, it is difficult to assess the potential clinical impact of differences in precision or recall across NLP tools.

The annotators expressed that it was sometimes challenging to assess suicidality risk based on one document at a time; single documents did not provide sufficient context in all cases. At the same time, given the rare prevalence of suicide-related content in all patient documents, defining a patient-level annotation task using this type of abundant clinical documentation would be very time-consuming. We plan to explore different ways of addressing this issue, one being a nested case-control study design similar to the one presented in Metzger et al.
A strength of this study is that we have not assumed that clinical terms used in more typically developing children or adults generalize to ASD populations. Assessing suicidality in adolescents with ASD often requires a different approach to other patient groups, which in our clinical experience was likely to be reflected in the clinical notes. Young people with ASD presenting to mental health services commonly have severe difficulties with interpersonal interactions, making for a more complex clinical assessment. Clinicians are likely to deliberate within the clinical notes on whether potential behaviours are driven by suicidal ideation. They may have a greater reliance on third person report – i.e. caregivers voicing concerns regarding the young person’s suicidality rather than direct accounts from the young person. Also, where a first person account is provided, clinicians will often write verbatim statements (i.e. He told me “I just want to end it”, and he “went to the car park to get it done”), providing more atypical clinical terminology for describing suicidality, and increasing the chance of NLP misclassification.

In addition, young people with ASD may not present with suicidality as a principle complaint, but through a behavioural change such as school refusal, with suicidal behaviour emerging through later clinician screening. This may change the emphasis and position within the patient’s clinical record relative to other populations where suicidal behaviour is the principle trigger during the first presentation to services. Testing these clinical assumptions empirically using an non-ASD control sample was beyond the scope of the current study, however future work is underway to examine the variability of the NLP tool’s accuracy across non-ASD child populations seen in mental health services. NLP applications are commonly validated using randomly extracted documents from EHRs covering a broad range of clinical contexts, seldom rarer clinical populations, such as young people with ASD. As mental health assessment and management needs to be tailored to the developmental needs of the young people in clinic, so should the validation of NLP data extraction tools.

The suicidality outcome data provided by this NLP extraction tool permits analyses of the complex interplay of ASD-specific traits on factors contributing to the onset and recurrence of suicidality. ASD specific mental health services are becoming increasingly available for child and adolescent populations in high-income countries. Although there is more work to be done before clinical application, we believe the NLP tool described provides a step forward in enhancing suicidality surveillance, risk prediction and treatment selection for children with ASD.

Acknowledgements

JD is supported by a Medical Research Council (MRC) Clinical Research Training Fellowship (MR/L017105/1), SV is supported by the Swedish Research Council (2015-00359), Marie Skłodowska Curie Actions. Cofund, Project INCA 600398, HD received salary support from the Foundation of Professional Services to Adolescents, RD is funded by a Clinician Scientist Fellowship (research project e-HOST-IT) from the Health Foundation in partnership with the Academy of Medical Sciences which also funded GG’s salary.

References

8 Downs J, Gilbert R, Hayes R, Hotopf M, Ford T. Linking health and education data to plan and evaluate services


Secure and Trustable Electronic Medical Records Sharing using Blockchain

Alevtina Dubovitskaya, MS1,2, Zhigang Xu, MD3, Samuel Ryu, MD3,
Michael Schumacher, PhD1, Fusheng Wang, PhD4

1University of Applied Sciences Western Switzerland (HES-SO), Sierre, VS, Switzerland;
2École polytechnique fédérale de Lausanne (EPFL), Lausanne, VD, Switzerland;
3Stony Brook Medicine, Stony Brook, NY, United States;
4Stony Brook University (SBU), Stony Brook, NY, United States

Abstract Electronic medical records (EMRs) are critical, highly sensitive private information in healthcare, and need to be frequently shared among peers. Blockchain provides a shared, immutable and transparent history of all the transactions to build applications with trust, accountability and transparency. This provides a unique opportunity to develop a secure and trustable EMR data management and sharing system using blockchain. In this paper, we present our perspectives on blockchain based healthcare data management, in particular, for EMR data sharing between healthcare providers and for research studies. We propose a framework on managing and sharing EMR data for cancer patient care. In collaboration with Stony Brook University Hospital, we implemented our framework in a prototype that ensures privacy, security, availability, and fine-grained access control over EMR data. The proposed work can significantly reduce the turnaround time for EMR sharing, improve decision making for medical care, and reduce the overall cost.

Introduction

Electronic medical records (EMRs) are critical but highly sensitive private information for diagnosis and treatment in healthcare, which need to be frequently distributed and shared among peers such as healthcare providers, insurance companies, pharmacies, researchers, patients families, among others. This poses a major challenge on keeping a patient’s medical history up-to-date. Storing and sharing data between multiple entities, maintaining access control through numerous consents only complicate the process of a patient’s treatment. A patient, suffering from a serious medical condition such as cancer, or HIV, has to maintain the long history of the treatment process and post-treatment rehabilitation and monitoring. Having access to a complete history may be crucial for his treatment: for instance, knowing the delivered radiation doses or laboratory results is necessary for continuing the treatment.

A patient may visit multiple medical institutions for a consultation, or may be transferred from one hospital to another. According to the legislation1–4, a patient is given a right over his health information and may set rules and limits on who can look at and receive his health information. If a patient needs to share his clinical data for the research purposes, or transfer them from one hospital to another, he may be required to sign a consent that specifies what type of data will be shared, the information about the recipient, and the period during which the data can be accessed by the recipient. This may be extremely difficult to coordinate, especially when a patient is moving to another city, region, or country and may not know in advance the caregiver or hospital where he will be receiving care later on.

Even if the consent is provided, the process of transferring the data is time consuming, especially if sending them by post. Sending the patients’ data via email over the Internet is not considered in most hospitals as this could impose security risk while the patient’s healthcare records are in transit1. Ecosystems for health information exchange (HIE) such as CommonWell Health Alliance aim to ensure that the data form patient electronic health record are securely, efficiently and accurately shared nationwide in US. This implies that once providers receives an access to the patient’s health information it is difficult to guarantee that a patient could receive independent opinions from different healthcare providers. Moreover, such ecosystems do not address the requirements in case of transferring data from one country to another.

Data aggregation for research purposes also requires the consent unless the data are anonymized. However, it has been shown that independent release of locally anonymized medical data corresponding to the same patient and originated from different sources (e.g., several healthcare institutions visited by the patient) could cause de-identification of the patient, and, therefore, violation of privacy31,32.

Relying on centralized entity that would store and manage the patients’ data and access control policies means having
single point of failure and a bottleneck of the whole framework. It also requires either conducting all the operations (such as search, or anonymization) over encrypted data, or choosing a fully trusted entity that will have access to sensitive information about the patients. The former still requires management of large amounts of memory and is not suitable for hospital environment. The latter was proven to be very difficult to put in practice. An example of GoogleHealth wallet has shown that patients are concerned about their privacy and aware of the potential risk that their sensitive data might be misused.

Having access to a ledger - shared, immutable, and transparent history of all the actions that have happened to all the participants of the network (such as a patient modifying permissions, a doctor, accessing or uploading new data, or sharing them for research) overcome the issues presented above. By providing the tool to achieve consensus among distributed entities without relying on a single trusted party, blockchain technology will guarantee data security, control over sensitive data, and will facilitate healthcare data management for the patient and different actors in medical domain. In the healthcare settings we can define a transaction as a process of creating, uploading or transferring EMR data that is performed within the connected peers. A set of transactions grouped at certain time is added to the ledger that records all the transaction and therefore represents the state of the network. The key benefits of applying the blockchain technology in healthcare are the following: verifiable and immutable transactions; tamper resistance, transparency, and integrity of distributed sensitive medical data. This is mainly achieved by employing consensus protocol and cryptographic primitives such as hashing and digital signatures.

The possibility of using blockchain for healthcare data management has recently raised a lot of attention in both industry and academia. However, only one functioning prototype of a system that uses blockchain for medical data management has been proposed. In our work, we focus on a practical implementation of a system that uses blockchain technology and can be integrated in clinical practice. We employ permissioned blockchain technology to maintain metadata and access control policy and a cloud service to store encrypted patients’ data. Combining these technologies allows us to guarantee data security and privacy as well as availability with respect to the access control policy defined by the patient.

The contribution of the paper is twofold. First, we propose multiple scenarios of blockchain applications in healthcare and analyze existing technology implementations that could be used to put the scenarios in practice. Second, we present a framework for blockchain based data sharing for primary care of oncology patients under cancer treatment. We developed a prototype in collaboration with the Department of Radiation Oncology in a major US hospital. Therefore, the functionality of the prototype is expected to meet the requirements from medical practice perspective.

1 Background on Blockchain

Blockchain is a peer-to-peer distributed ledger technology that was initially used in the financial industry. Based on how the identity of a user is defined within a network, one could distinguish between permissioned and permissionless blockchain systems. A permissionless system is one in which the identities of participants are either pseudonymous or even anonymous and every user may append a new block to the ledger. In contrast, in case of a permissioned blockchain, the identity of a user is controlled by an identity provider. The identity provider is trusted to maintain access control within the network and the user’s rights to participate in the consensus, or validate a new block. Next we introduce two most well-known implementations of the blockchain technology: Ethereum and Hyperledger.

1.1 Permissionless Blockchain Implementation

Ethereum is an implementation of a permissionless programmable blockchain that allows any user to create and execute the code of arbitrary algorithmic complexity on the Ethereum platform: Ethereum Virtual Machine (EVM). “Accounts” of two types could be created on EVM. Externally owned account (EOA) is an account controlled by a private key of a user. Contract account is the second type of accounts that can be seen as an autonomous agent that lives in the Ethereum execution environment and is controlled by its contract code: smart contract. Smart contract is used to encode arbitrary state transition functions, allowing users to create systems with different functionalities by transforming the logic of the system into the code.

Code execution in Ethereum must be paid. The transaction price limits the number of computational steps for the code execution in order to prevent infinite loops or other computational wastage. Users could participate in a consensus
process to obtain the tokens to be paid for transaction execution. In Ethereum, the consensus is achieved by using a proof-of-work (PoW) mechanism. PoW is based on “mining”: finding a nonce input to the algorithm so that the resulting hash of a new valid block satisfies certain requirements. These requirements set the difficulty threshold for the process of finding the nonce. The difficulty threshold impacts the amount of energy to be spent to find such nonce. For example, the amount of energy used by Bitcoin mining is comparable to the Irish national energy consumption. Existing PoW blockchains can achieve throughput of not more than 60 transactions per second without significantly affecting the blockchain’s security. These two findings show that PoW can negatively impact the system scalability and overall throughput.

Proof-of-Stake (PoS) and Proof-of-Burn (PoB), or virtual mining mechanisms, have been recently proposed as alternatives to PoW. Instead of having participants mine by exchanging their wealth for computational resources (which are then exchanged for mining rewards), in virtual mining, participants could exchange their wealth directly for the ability to append a new block to the ledger. For example, in PoS, selection of a participant that will create a new block is based on the amount of tokens owned by the participant, in PoB – based on the amount of tokens “burnt” (sent to an unspendable address). However, providing a rigorous argument for or against the stability of virtual mining remains an open problem.

1.2 Permissioned Blockchain Implementation

In the case of a permissioned system, users do not have an incentive to cheat as their identity is revealed to the identity server. Moreover, participation in consensus management is restricted to a predefined set of users. This opens a possibility to use a state machine replication algorithm (such as PBFT) as a consensus mechanism. Hyperledger – an implementation of a permissioned blockchain – is an open source blockchain initiative hosted by the Linux Foundation. Hyperledger has a modular architecture that allows plugging in different consensus mechanisms, including PBFT. Hyperledger services could be logically grouped in three categories: Membership services, Blockchain services, and Chaincode services.

Membership services manage identity, privacy, and confidentiality on the network. A user is assigned a username and a password that will be used to issue the Enrollment certificate (ECert) to identify every registered user. It is possible to use different Transaction certificates (TCert) associated with the same ECert for every transaction to ensure their unlinkability (a mapping between TCert and ECert are only known to the membership service). Blockchain services manage the distributed ledger through a peer-to-peer protocol built on HTTP/2. In Hyperledger, smart contracts are implemented by the chaincode. Chaincode services provide a secure way to execute smart contracts on validating nodes.

In Hyperledger, smart contracts are implemented by chaincode that consist of Logic and associated World State (State). Logic of the chaincode is a set of rules that define how transactions will be executed and how State will change. State is a database that stores the information in a form of keys and values that are arbitrary byte arrays. The State also contains the block number to which it corresponds. Ledger manages blockchain by including an efficiently cryptographic hash of the State when appending a block. This allows efficient synchronization if a node was temporary off-line, minimizing the amount of stored data at the node.

2 Potential Blockchain Applications in Healthcare

Blockchain provides a unique opportunity to support healthcare. In this section, we propose three scenarios: primary patient care, medical research, and connected health. Figure 1 shows a graphical representation of the combination of the aforementioned scenarios.

Scenario 1: Primary Patient Care. Using blockchain technology for primary patient care can help to address the following problems of the current healthcare systems:

- A patient often visits multiple disconnected hospitals. He has to keep the history of all his data and maintain the updates. This leads to the situation when required information may not be available.
- Due to the unavailability of the data, patient may have to repeat some tests for laboratory results. This is common when the results are stored in another hospital and can not be immediately accessed.
The healthcare data are sensitive and their management is cumbersome. Yet, there is no privacy-preserving system in clinical practice that allows patients to maintain access control policy in an efficient manner.

- Sharing data between different healthcare providers may require major effort and could be time consuming.

Next, we propose two approaches that can be implemented separately or combined to improve patient care.

- **Institution-based**: The network would be formed by the trusted peers: healthcare institutions or general practitioners (caregivers). The peers will run consensus protocol and maintain a distributed ledger. The patient (or his relatives) will be able to access and manage his data through an application at any node where his information is stored. If a peer is off-line, a patient could access the data through any other online node. The key management process and the access control policy will be encoded in a chaincode, thus, ensuring data security and patient’s privacy.

- **Case specific** (serious medical conditions, examination, elderly care): During a patient’s stay in a hospital for treatment, rehabilitation, examination, or surgery, a case-specific ledger could be created. The network would connect doctors, nurses, and family to achieve efficiency and transparency of the treatment. This will help to eliminate human-made mistakes, to ensure consensus in case of a debate about certain stage of the treatment.

**Scenario 2: Data Aggregation for Research Purposes.** It is highly important to ensure that the sources of the data are trusted medical institutions and, therefore, the data are authentic. Using shared distributed ledger will provide traceability and will guarantee patients’ privacy as well as the transparency of the data aggregation process. Due to the current lack of appropriate mechanisms, patients are often unwilling to participate in data sharing. Using blockchain technology within a network of researchers, biobanks, and healthcare institutions will facilitate the process of collecting patients’ data for research purposes.

**Scenario 3: Connecting Different Healthcare Players for Better Patient Care.** Connected health is a model for healthcare delivery that aims to maximize healthcare resources and provide opportunities for consumers to engage with caregiver and improve self-management of a health condition. Sharing the ledger (using the permission-based approach) among entities (such as insurance companies and pharmacies) will facilitate medication and cost management for a patient, especially in case of chronic disease management. Providing pharmacies with accurately updated data about prescriptions will improve the logistics. Access to a common ledger would allow the transparency in the whole process of the treatment, from monitoring if a patient follows correctly the prescribed treatment, to facilitating communication with an insurance company regarding the costs of the treatment and medications.

**Implementing the Scenarios.** In order to implement the three healthcare scenarios presented above, we must choose
between a permissionless and a permissioned blockchain implementations. Below we present the facts that favor a permissioned system implementation.

- The anonymity of users and impossibility to verify the identity of account holders (as in case of permissionless blockchain) could cause impersonalization and data misuse.
- Patients’ healthcare data are of high sensitive nature. Even monitoring communication between a patient and a specific clinician may reveal some sensitive data about the patient, therefore violating the privacy.
- Fast response of a system is required as any update of the information about a patient’s treatment could be crucial for the patient.
- The need to pay for transaction execution, for example, updating permissions for a medical doctor to access a piece of healthcare information or sharing some data for research could limit the usability of the system.

3 Application in Radiation Oncology: Sharing Clinical Data between Healthcare Providers

In this section, we present a prototype design and implementation of a system to support electronic medical record sharing for primary patient care (Scenario 1). More precisely, we focus on patients that are receiving a cancer treatment via ionizing radiation, which is usually performed in the Department of Radiation Oncology of a hospital. First, we describe a specific use-case scenario and the benefits of the system. Second, we present the architecture of the system and describe the data structure and functionality of the system. Finally, we discuss how privacy, security, and scalability are ensured within the proposed framework.

3.1 Use Case Scenario

Cancer is a serious medical condition that may require a long-lasting treatment and a life-time monitoring of a patient. Therefore, it is crucial for the patient to maintain his medical history and to be able to access or share his medical data during the treatment and post-treatment monitoring. Due to the mobility of a patient, the management of the data generated during every patient’s visit can be cumbersome especially given the sensitive nature of healthcare data. How to guarantee that the patient’s data are complete, stored securely, and can be accessed only according to the patient consent in a fast and convenient manner?

We tackle this problem by applying the blockchain technology to create a prototype of an oncology-specific clinical data sharing system. To present our solution, we take as an example an oncology information system, ARIA, which is widely used to facilitate oncology-specific comprehensive information and images management. ARIA combines radiation, medical and surgical oncology information and can assist clinicians to manage different kinds of medical data, develop oncology-specific care plans, and monitor radiation dose of patients. Different types of data stored in this system can be structured depending on the clinician’s request and exported in PDF format. The documents that contain the data such as history and physical exams, laboratory results, and delivered radiation doses are of the high importance for the clinicians and are most commonly used during the treatment.

Currently, if any of these data have to be transferred from Hospital 1 to Hospital 2, the following procedure takes place. First, the patient (or his official representative) has to sign a consent – a document that specifies the data to be transferred and contains the information about the recipient of the data (Hospital 2). Then, the information has to be printed and mailed to the recipient. Consent management and data transfer in this case can become complicated and inconvenient: the patient may need to contact the caregiver and sign a consent in the hospital from which he is not receiving care anymore. Data transfer can take time, and on receiving the hard copy of the patient data, a clinician will have to introduce them into the system again. Moreover, with this approach, it is very difficult for the patient to maintain any access control of his data and to have a complete view of the data.

By employing blockchain technology, our solution allows to facilitate the consent management and speed up data transfer. We developed a chaincode that allows a patient to easily impose fine-grained access control policy for his data and enables efficient data sharing among clinicians.

3.2 System Architecture

Figure 2 presents the architecture of our framework for the oncology-specific data management. The framework consists of the Membership service, Databases for storing healthcare data off-chain, Nodes managing consensus process,
and APIs for different user’s roles. Currently, we focus on Doctor and Patient, but the roles and their functionality could be extended depending on the scenario.

The main functionality of the Membership service is to register users with different roles (currently Doctor and Patient). The roles define the functionality of the chaincode that is available to the user. During the registration of a user as a Doctor, it is important to ensure that it is not a potential malicious user, but a qualified medical doctor. To verify this, the National Practitioner Data Bank could be consulted by the membership service. The membership service is also hosting a certification authority involved in the generation of a key pair for signing ($SK_{SU}$, $PK_{SU}$) and encryption key pair ($SK_{EU}$, $PK_{EU}$) for every user ($U$).

Patient ($P$) also generates a symmetric encryption key ($SK_{AESP}$) that will be used to encrypt/decrypt the data corresponding to the patient, $P$. This key will also be used to generate pseudonyms so that only authorized users could verify whether the ledger stores any information about the patient. When a patient ($P$) needs to share his data with a doctor ($D$), the patient could share this key, $SK_{AESP}$, using the encryption public key of the corresponding clinician ($PK_{ED}$). If the symmetric key $SK_{AESP}$ is compromised, the patient could generate a new one, run a proxy re-encryption algorithm$^{26}$ on the data stored in the cloud and then share a new key with the clinicians according to the desired access control policy.

The patient’s data are stored off-chain in the following Databases. First, a local database management system in the hospital that stores the oncology-related data (for example, ARIA in our use case). Second, a cloud based platform (Varian Cloud) that stores patient’s data organized based on the data category (for instance, according to the sensitivity level of the data, or their semantics), and encrypted with corresponding patient key, $SK_{AESP}$. A registered clinician could assess or upload the data in the cloud repository based on the access control policy defined by the patient and implemented in the chaincode Logic.

A custom chaincode is deployed on every Node that acts as a Hyperledger validating peer. Nodes receive all transactions submitted by the users through a role-based APIs. The Node, selected as a leader, organizes transactions in a block and initiates the PBFT consensus protocol. Transactions are executed by all nodes according to the implemented chaincode Logic. The State stores the information about patients in a key-value pair format. A Key – a Patient Id in the system – is a pseudonym of the patient that is generated as a hash of the concatenation of the symmetric key $SK_{AESP}$ and a Uniquely Identifiable Information of the patient ($UII_P$): $H(SK_{AESP} || UII_P)$. Combination of SSN (if applicable), date of birth, given names, and a ZIP code of the patient could be used as $UII_P$. A Value is a patient record stored as a byte array. Next we describe the data structure in detail.
3.3 Data Structure and System Functionality

Figure 3 shows how the patient’s data and metadata are organized: the patients’ data are stored off-chain: locally (in the clinician database) and in the cloud as presented in Figure 3 (a) based on their categories. Currently we use three categories in our prototype: History and physical exams, Laboratory results, and Delivered radiation doses. In the future, we plan to define categories based on both the semantics and the sensitivity level of the data. Data files related to the patient and uploaded by different clinicians are stored within the corresponding category. A patient could optionally store some private data or notes encrypted with the patient public key, $\mathcal{PK}^S_p$.

Figure 3 (b) presents the structure of the patient’s metadata that consists of the following blocks: Permissions, Clinical Metadata, and Patient Private Data (optional). Permissions block is organized as follows: every Permission corresponds to a Doctor Id, with which a clinician is registered in the system. Every permission specifies the timeframe (From: To:) during which a clinician has a Right to read the patient’s data that fall into a specific Data Category, upload them to the cloud repository (write), or share the patient’s data within a framework of a specific research study, Study Id. For the latter the patient could also use Anonymity tag to specify if the data must be anonymized before sharing or could be shared as they are. Timestamp makes every permission unique and allows a patient to update and track access control changes corresponding to the same Doctor Id.

Clinical Metadata are a block that contains information about all the data files uploaded to the cloud by different clinicians. The Metadata Items are categorized based on the semantics of the corresponding data files. Every Item contains an Id of the clinician that uploaded the data (Doctor Id), a pointer to the file that is stored in the cloud, Path to File, the Hash of the Data File, Hash(File), to ensure unforgeability of the data stored in the cloud, and the Timestamp of the event when the Data File was uploaded. Similarly to the Patient Private Data stored on the cloud, some private data could be added by a patient to be stored in the State associated with the chaincode (CC). The metadata are stored as a “value” part of a CC State. It can be accessed and modified using the functions that can be invoked on a CC.

To ensure a correct functioning of the developed chaincode we built a network that consists of a Membership service and four Nodes capable of running PBFT consensus protocol. Four is the minimum number of nodes needed to run the PBFT consensus protocol. We deployed CC on every node and issued a set of the “invoke” transactions (that trigger creation of a new patient metadata record, adding a permission, and uploading the metadata item), and “query” transactions to access the information from the State. Currently, a patient is able to create a metadata record on the chaincode, add permissions, and retrieve his up-to-date metadata record, and, thus, his data that are stored on the cloud. A user registered with a Doctor role is able to upload, access and share for research purposes the data in the cloud based on the permissions specified by the patient.

Verification of the access control rights (currently read, write or share) is done via Logic of a chaincode written in Go programming language\cite{27}. For instance, every time a clinician would try to add new data on the cloud repository,
a permission corresponding to this clinician has to be retrieved from the patient’s metadata record. Then, the validity of the permission with respect to the data category and timeframe is controlled. Similarly, sharing patient data for the research purposes can not be performed by clinician without patient’s agreement. This is guaranteed by the chaincode implementation. Interfacing our system with the existing clinical database management systems and conducting more experiments with the data of the real patients are next steps of our work.

4 Discussion

Next we discuss the privacy, security, and scalability of the proposed framework.

Privacy. A patient’s privacy is ensured by providing the patient with a possibility to specify fine-grained access control over his data via permissions. Permissions are enforced by chaincode logic and, therefore, can not be violated by any user, unless the consensus protocols fails. The latter could happen only if a fraction of the verifying nodes intentionally tries to damage network operations. Centralized membership service already protects against Sybil attacks. Moreover, in the permissioned network, the nodes identities are known, therefore, there is no incentive for malicious behavior. In the case if a node still behaves maliciously, access to the network could be promptly restricted for this node. Membership service also controls the identity of the users. Before registering a clinician, his identity is verified in the National Practitioner Data Bank. A patient is registered with his UII, but all his data are linked to the pseudonym generated using his secret key, SKAES. Therefore, Membership service does not have an access to the patient’s clinical data, yet guarantees authenticity of the users (via digital signature verification). If SKAES is compromised or lost, access to the network will be recovered using UII of a patient, a new key will be generated, and proxy reencryption will be used.

Security. Clinical data stored in the cloud repository are encrypted with a patient secret key, SKAES, to provide data confidentiality. Only the patient can share encryption key and set up the access control policy via permissions. Shared data from the cloud registry are hashed and signed with a secret key of a user (SKSU), before the data are uploaded. The hashes are stored as a part of a corresponding metadata item in the State. Transactions are also digitally signed, thus the data integrity is ensured.

Availability of the shared data is guaranteed by providing a cloud platform to store the data. Role-based APIs can be used at any node registered in the network to invoke or query the chaincode. As already mentioned, if a patient loses his credentials, access to the data stored on- and off-chain could still be recovered.

Scalability. Clinical data sharing requires scalability of the system in terms of both the number of users and the number of nodes. PBFT consensus protocol provides excellent scalability in terms of the number of users, but have not been well explored in terms of the number of Nodes (verified only up to few tens of Nodes). Possible scalability issues could be addressed by using hierarchical BFT protocols. Frequency of creating a block or number of transactions in a block (batch size) could be adjusted. System load is already minimized by storing patient’s clinical records off-chain. We plan to evaluate the system performance and scalability in clinical settings in future work.

5 Related Work

The potential of the applications built on top of the blockchain technology for healthcare data management has been recently discussed. Yue et al. claim to be the first to import blockchain into the design of a healthcare system. They presented the architecture of a healthcare data gateway application for easy and secure control and sharing of medical data between different entities that may use patient data. However, the system has not been implemented nor tested yet. A possibility of sharing the data for research purposes is only sketched in the paper, without any security or privacy evaluation. Jenkins et al. proposed to use blockchain technology for a multifactor authentication in a specific research scenario (medical large data analysis with functional biomarkers) that involves biometric and biomedical data.

MedRec is the first and the only functioning prototype that have been proposed until now. The authors presented a system based on Ethereum smart contracts for an intelligent representations of existing medical records that are stored within individual nodes on the network. Two incentivizing models for “mining” are also proposed in. Our prototype significantly differs from the framework in. First, MedRec is based on permissionless blockchain implementation.
and PoW, thus involves transaction fees, and requires involvement into “mining” and account management processes. In contrast, we have chosen permissioned blockchain implementation based on the requirements from the medical perspective. We justified our choice in Section 3.4. Second, in the patients data are stored locally at every node. We decided to use a cloud-based storage, and employ encryption and key-sharing to ensure availability of the data even if the hospital node is temporary off-line.

Conclusion and Future Work

In this paper, we proposed scenarios of blockchain technology application in different healthcare settings: primary care, medical data research, and connected health. We discussed how maintaining an immutable and transparent ledger, which keeps track of all the events happened across the network, could improve and facilitate the management of medical data.

Based on the constrains related to the healthcare context, we justified the choice of the permissioned blockchain technology for the implementation of the proposed scenarios. We also presented an architecture of the framework for the specific needs in case of radiation oncology data sharing and implemented a prototype that ensures privacy, security, availability, and fine-grained access control over highly sensitive patients’ data.

As part of future work, we would like to extend the structure of a patient record and its metadata, using the semantics of healthcare data, including the possibility of sharing radiology images, which is much more challenging. Since we work in collaboration with a hospital, we plan to test our system with the data of the real patients. Our long term goal is to explore other scenarios proposed in the paper (such as connected health and medical data research) and apply them in practice to enhance the current healthcare data management.

References

4. EU Directive. 95/46/ec of the european parliament and of the council of 24 october 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data. Official Journal of the EC, 23(6), 1995.
21. GEM [Internet]. Available from: https://gem.co/health/.
Evaluation of Clinical Text Segmentation to Facilitate Cohort Retrieval

Tracy Edinger, ND, MS1, Dina Demner-Fushman, MD, PhD2, Aaron M. Cohen, MD, MS 1, Steven Bedrick, PhD 1, William Hersh, MD1
1Department of Medical Informatics and Clinical Epidemiology, Oregon Health & Science University, Portland, OR, USA
2National Library of Medicine, National Institutes of Health, Bethesda, MD, USA

Abstract

Objective: Secondary use of electronic health record (EHR) data is enabled by accurate and complete retrieval of the relevant patient cohort, which requires searching both structured and unstructured data. Clinical text poses difficulties to searching, although chart notes incorporate structure that may facilitate accurate retrieval. Methods: We developed rules identifying clinical document sections, which can be indexed in search engines that allow faceted searches, such as Lucene or Essie, an NLM search engine. We developed 22 clinical cohorts and two queries for each cohort, one utilizing section headings and the other searching the whole document. We manually evaluated a subset of retrieved documents to compare query performance. Results: Querying by section had lower recall than whole-document queries (0.83 vs 0.95), higher precision (0.73 vs 0.54), and higher F1 (0.78 vs 0.69). Conclusion: This evaluation suggests that searching specific sections may improve precision under certain conditions and often with loss of recall.

Introduction

The use of electronic health records (EHR) in inpatient and outpatient clinical facilities has increased rapidly over recent years. Incorporating EHRs has facilitated clinical practice, resulting in widespread access to patient information and potentially better care. Because of the quantity of data in an EHR and the number of patients in a typical health-care system, using this data for secondary purposes can be of enormous benefit both to the local health care system and nationally.

EHRs are used primarily for direct patient care and for billing purposes, and they are designed to maximize the ability to provide these functions. However, in this process, the EHR becomes a repository of a vast amount of clinical data, both structured and text. This data can be used for many purposes beyond clinical care and billing: research, quality measures, disease surveillance, operational improvement, and administrative applications. All of these uses of EHR data require the ability to find desired information with a high degree of accuracy. In many cases, this means that we need to be able to identify specific patient cohorts. For example, researchers may want to look at records for all patients with a particular disease or all patients who received a specific treatment. For accurate disease surveillance, we must be able to identify patients who had that disease, either by locating specific diagnosis codes or mentions of the disease, or by locating surrogate indicators such as treatments or symptoms. Alternatively, hospital administrators may want to count the number of patients who received a particular treatment, or the number of procedures performed in a clinic. In some cases, conditions or symptoms may be documented only in clinical notes, and only available through a text search.

The way data is stored in an EHR determines how we need to search for specific patients. Some of the data is stored in structured fields, which are relatively easy to access. This data is recorded in consistent ways as specific codes, including international classification of disease (ICD) or other codes, or as predetermined phrases. Although structured data is relatively easy to search, several studies suggest that a search relying solely on structured data will not retrieve all patients relevant to the topic1,2,3,4, and other studies suggest that combining structured and unstructured data will improve cohort retrieval5. A great deal of EHR data, however, is stored as unstructured text in clinical notes, history and physical exam notes, and reports. This unstructured text is very difficult to access on a large scale. Medical text contains several features that compound this difficulty, including frequent use of abbreviations, lack of standardization of abbreviations, context-dependent differences in word meanings and abbreviations, negation of symptoms or diseases, and documentation of history not directly related to the current visit. Because of the difficulty in accurately extracting data from text, most non-research use of EHR data utilizes structured data only. Clinical notes contain highly valuable information not found in structured fields, and they
document clinical thinking and decision making. Further, relying solely on structured data for analytic functions results in a greater data-entry burden on providers. Because of these factors, improving retrieval accuracy from text would have great value.

The National Institute of Standards and Technology (NIST) has sponsored challenges to evaluate and improve the ability to accurately retrieve patient data from medical text for secondary uses. In NIST's 2011 and 2012 Text Retrieval Conference Medical Records Track (TRECMed), participants were given a list of clinical topics and a set of de-identified textual medical records. Participants developed search systems and algorithms to retrieve records relevant to each topic. Retrieved visits were then judged for relevance to the topic. An analysis of incorrectly retrieved visits identified several key problems in accurate retrieval of patient cohorts6. Factors in retrieving non-relevant visits included terminology similarities, negation of the desired term, and mention of the desired term as a past or future occurrence. Relevant visits were overlooked when the chart notes used different terminology or described rather than named the condition.

The Electronic Medical Records and Genomics Network (eMERGE) Consortium has also evaluated the ability to retrieve specific patient cohorts from EHRs7. In this study, EHR data from five different sites were used to identify patients with at least one of five diseases (dementia, cataracts, peripheral arterial disease, type 2 diabetes, and cardiac conduction defects). Patients were identified with a high level of accuracy when the data were stored in structured fields; however, in some cases, target information was stored only in clinical text. The use of natural language processing (NLP) tools increased retrieval significantly—129% more cases were identified by including the use of NLP tools rather than through using structured data and string matching alone. In evaluating results from one site, use of the same terms to mean different things within one document was an issue in correct retrieval of patients; for example, 'potassium' can be a medication or the name of a lab value, and a drug name can be listed as an allergy or as a prescribed medication8.

Several approaches may be utilized to overcome these issues and facilitate the retrieval process: the query can be constructed to yield a more accurate response, and the original text can be manipulated to make it more searchable. Clinicians are trained to write medical records in a highly structured fashion. Physicians' chart notes are also divided into sections that indicate the source and purpose of the information, in a structure referred to as SOAP (Subjective-Objective-Assessment-Plan). Within these sections, a typical chart note for a first encounter with a patient includes the chief complaint, a history of the present illness, a review of systems, past medical history, family history, and social history. Manipulating and searching the text according to these sections would allow the construction of searches targeted to the appropriate section, avoiding or minimizing some of the issues found in previous work. Several tools and strategies9,10,11,12,13,14 have been documented that segment clinical records. Although the effectiveness of each segmentation strategy has been evaluated, no studies could be located that demonstrate whether segmenting improves the performance of searching clinical text.

Temporality is an information retrieval (IR) issue that is particularly relevant to medical text, which often documents the current illness as well as previous illnesses. Clues to temporality can be found in identifying the section of the medical record: a description of the chief complaint is likely the current issue, whereas a condition listed in the past medical history is something that has resolved or is not the focus of the current visit.

Subject identification can also complicate retrieval of medical information. Chart notes may document illnesses of other family members as well as those of the patient. Identifying who has the disease improves accuracy of recall by avoiding retrieval based on someone else's disease status. The ability to separate sections of the medical record may facilitate retrieval accuracy by identifying the family history section and allowing that section to be searched only when applicable.

The hypothesis of this study is that searching for patient cohorts by looking in specific, relevant sections of clinical text will improve retrieval accuracy.

**Methods**

**Data** This project used de-identified clinical records developed by the Massachusetts Institute of Technology (MIT), Philips Medical Systems, and Beth Israel Deaconess Medical Center15. The Multiparameter Intelligent Monitoring in Intensive Care (MIMIC-II) data is a publicly available dataset containing more than 30,000 intensive-care unit (ICU) patients. The MIMIC-II data is stored in a relational database containing structured data and unstructured
textual discharge summaries, MD notes, radiology reports, and nursing notes. This project used all four types of text
documents for the search corpus.

**Search Engine** Queries for this project were run using the Essie 4 search engine developed by the National Library
of Medicine. Essie 4 maps terms to the UMLS and allows differential weighting of terms to alter the order of
documents retrieved. Mapping terms to the UMLS allows comparison of different but equivalent terms; for example,
'myocardial infarction' and 'heart attack' refer to the same concept using different words. Because Essie 4 maps to
the UMLS, equivalent terms are found without being listed explicitly, allowing the queries to focus on detecting the
difference in retrieval when using sections, rather than including an exhaustive list of all possible synonyms.

When a query is run, Essie 4 returns a list of documents retrieved for that query. Because of term weighting, each
document is assigned a rank that indicates the relative likelihood that it is relevant to the query. Weights are given
values greater than 0 and less than 1.

**Segmentation** To segment the documents, we examined a subset of each type of document to identify the most
common section headings, recording all terminology, spelling, and punctuation variations. We then created a text
file containing these heading variations and the corresponding XML heading tags for the new sections. Table 1
shows examples of the headings inserted for the indicator text listed.

<table>
<thead>
<tr>
<th>Inserted Heading</th>
<th>Indicator Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>AssessmentAndPlan</td>
<td>disposition/plan treatment/plan: overall assessment and plan:</td>
</tr>
<tr>
<td>Course</td>
<td>clinical course in the emergency department: ed course: institution course:</td>
</tr>
<tr>
<td>CDCDisposition</td>
<td>transferred to: dispo:</td>
</tr>
<tr>
<td>LabRadResults</td>
<td>cta chest: important diagnostics and labs: radiographs-</td>
</tr>
<tr>
<td>FinalDiagnosis</td>
<td>diagnoses at the time of discharge: final discharge diagnosis: diagnosis on transfer:</td>
</tr>
</tbody>
</table>

The documents were searched for exact matches to the heading variations listed in the text file, and the appropriate
heading tags were inserted at those locations. Two tags were inserted for each heading, one at the start of the section
and the other at the end. When a new heading was located, an opening tag was inserted for that heading, and an
ending tag for the previous heading was inserted just before it. Tags were set up in XML format; for example,<AdmissionDiagnosis> at the start of the section and </AdmissionDiagnosis> at the end of the section.

Prior to having the section tags inserted, the original documents contained a single block of text surrounded by
opening and closing tags, as shown in this example:

```xml
<text>
DATE: [**3305-8-7**] 1:51 PM
CHEST (PORTABLE AP)
Reason: CHECK ETT TUBE PLACEMENT
?PNA, CHF
REASON FOR THIS EXAMINATION:
CHECK ETT TUBE PLACEMENT
?PNA
CHF
UNDERLYING MEDICAL CONDITION:
85 y/o male s/p acute mi and catheterization now
</text>
```
in ccu with cardiogenic shock.

FINAL REPORT


Comparison is made to previous study of one day earlier. An endotracheal tube is present, in satisfactory position. A Swan-Ganz catheter terminates in the proximal left pulmonary artery and has been withdrawn in the interval. An intraaortic balloon pump terminates about 3.3 cm below the superior aspect of the aortic knob, and a nasogastric tube terminates in the region of the gastroduodenal junction.

Cardiac and mediastinal contour are stable in the interval and pulmonary vascularity is within normal limits for technique. There has been improvement in the left retrocardiac opacity and there remains a patchy right basilar opacification which is slightly increased. A small amount of fluid is seen in the minor fissure.

IMPRESSION:
1) Lines and tubes in satisfactory position, as detailed above, with no evidence of pneumothorax.
2) Improved left retrocardiac opacity and worsened right lower lobe opacity likely due to atelectasis.

After segmenting this text, the document is broken into blocks, with the preamble, indication, condition, procedure details, and study impression separated by XML tags:

The segmented documents were indexed in Essie 4. Search queries utilized the XML tags to locate text in specific sections of the documents.

Queries A set of clinical topics to be retrieved from the text were developed from topics in TREC Med 2012, a project which used a set of de-identified clinical records for hospital and emergency-department patients. This set contains fifty clinical topics drawn from the Institute of Medicine's clinical comparative effectiveness priorities (16 topics), meaningful use clinical quality measures (12 topics), and the OHSUMED literature retrieval test collection (22 topics). Because the original query topics were developed for a wider range of patients, not all topics were relevant to ICU patients. A subset of 22 topics was used, modified as necessary to fit the current clinic population.
Queries were developed in an iterative fashion, refining the search details to maximize the number of relevant visits returned. An initial query was run against the data. The text of a subset of the returned visits was examined to determine if any new terms or operators needed to be added to the query, or if any query components needed to be removed. Although information from the returned visits was used to guide query development, details that appeared to be site-specific were not used. Once a set of queries was developed, they were run on the data, and the retrieved visits were recorded. The queries were then revised to utilize the sections, and visits retrieved with the revised queries were recorded. Below are examples of two clinical topics and the two queries for each topic:

**Topic: Patients with esophageal cancer who develop pericardial effusion**
Base Query: \textit{esophageal cancer AND "pericardial effusion"}
Query by sections: \textit{esophageal cancer AND [AREA][FinalDiagnosis] "pericardial effusion" OR AREA[Course] "pericardial effusion" OR AREA[LabRadResults] "pericardial effusion" OR AREA[AssessmentAndPlan] "pericardial effusion"}

**Topic: Adults with Alzheimer's disease with pressure ulcers discharged to nursing homes**
Base Query: \textit{alzheimers AND EXPAND[concept] (bed sore OR pressure ulcer) AND (NOT home OR facility OR "nursing home" OR "extended care" OR "assisted living") AND NOT expired}
Query by sections: \textit{alzheimers AND EXPAND[concept] (bed sore OR pressure ulcer) AND AREA[DCDisposition] (NOT home OR facility OR nursing OR extended) AND AREA[DCDisposition] NOT expired}

**Query Analysis** A subset of retrieved documents was examined to understand the effect of segmenting on retrieval performance. Three sets of documents were examined; two of these sets were the ten highest ranked documents retrieved only with the base query or only with the query by sections. The third set compared results when both queries retrieved the same document; the difference in rank assigned to those documents by each query was used to decide which to examine. The difference was calculated by subtracting the rank assigned by the query by sections from the rank assigned by the base query. When this value was highly positive, the base query had assigned a much greater rank than the query by sections; ie, the document was placed much lower in the list of ranked retrieved results. When this value was highly negative, the query by section had assigned a much greater rank than the base query. The ten documents with the largest positive difference in rank and the ten documents with the largest negative difference in rank were examined. Because the queries retrieved different numbers of documents, the actual number analyzed for each topic was usually lower than 40. Table 2 below shows the topics and the number of documents analyzed for each topic.

**Table 2. Query topics and number of documents analyzed.**

<table>
<thead>
<tr>
<th>Number of Documents Analyzed</th>
<th>Query Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>40</td>
<td>Patients with dental caries</td>
</tr>
<tr>
<td>40</td>
<td>Patients with thyrotoxicosis treated with beta-blockers</td>
</tr>
<tr>
<td>40</td>
<td>Patients with acute vision loss</td>
</tr>
<tr>
<td>40</td>
<td>Patients with left lower quadrant abdominal pain</td>
</tr>
<tr>
<td>40</td>
<td>Patients with low back pain who had imaging studies</td>
</tr>
<tr>
<td>38</td>
<td>Patients treated for the post-partum problems depression, hypercoagulability or cardiomyopathy</td>
</tr>
<tr>
<td>30</td>
<td>Patients who developed disseminated intravascular coagulation in the hospital</td>
</tr>
<tr>
<td>30</td>
<td>Patients who have gluten intolerance or celiac disease</td>
</tr>
<tr>
<td>30</td>
<td>Patients with colon cancer receiving chemotherapy</td>
</tr>
<tr>
<td>29</td>
<td>Patients with ventilator-associated pneumonia</td>
</tr>
<tr>
<td>22</td>
<td>Patients with delirium, hypertension, and tachycardia</td>
</tr>
<tr>
<td>20</td>
<td>Patients admitted to the hospital with end-stage chronic disease who are discharged on hospice care</td>
</tr>
<tr>
<td>20</td>
<td>Patients who had a carotid endarterectomy during this admission</td>
</tr>
<tr>
<td>20</td>
<td>Patients who received pneumonia vaccination during this admission</td>
</tr>
<tr>
<td>20</td>
<td>Patients with HIV/AIDS who develop pancytopenia</td>
</tr>
<tr>
<td>20</td>
<td>Patients with diabetes mellitus who also have thrombocytosis</td>
</tr>
</tbody>
</table>
For each document, several observations were recorded. Retrieved documents were evaluated for relevance to the topic. Non-relevant documents were examined to determine the reason for retrieval; specifically, the documents were examined to identify occurrences of the search terms. We assessed the reason for performance differences between the two queries, and the reason for performance difference when querying by sections. Codes to reflect these assessments were developed in an iterative fashion, with new codes added as necessary to capture new reasons.

**Results**
A total of 574 documents were examined. Of those, 344 were relevant to the given topic. Both queries retrieved 247 of these documents, querying by sections retrieved an additional 20 documents, and querying the whole document retrieved an additional 77 documents. The remaining 230 non-relevant documents included 146 that were not at all relevant, 53 that were relevant to portions of the topic but not the entire topic, and 18 that were possibly relevant.

Eighty-one non-relevant documents were retrieved by both queries, querying by sections retrieved an additional six documents, and querying the whole document retrieved an additional 143 documents. Figure 1 illustrates the comparative retrieval rates for queries of specific sections versus querying the whole document.

---

**Table 3.** Recall, precision, and F1 measures for both query types.

<table>
<thead>
<tr>
<th>Query Type</th>
<th>Recall</th>
<th>Precision</th>
<th>F1 Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Whole-document Queries</td>
<td>0.95</td>
<td>0.54</td>
<td>0.69</td>
</tr>
<tr>
<td>Section Queries</td>
<td>0.83</td>
<td>0.73</td>
<td>0.78</td>
</tr>
</tbody>
</table>

Examination of the documents revealed that not all sections were identified correctly. This was not surprising, given our method of section identification. Because sections were ended only when new ones were identified, unidentified
sections were erroneously included in the previous section. Our identification rules specified text patterns that
signified new sections, so sections that were implicitly identified by extra spaces or changes in wording or content
were overlooked. In all, we identified 44 documents in which relevant sections were not correctly tagged, factoring
into incorrect retrieval. Eleven of these were relevant documents not retrieved when querying by section; accurate
segmenting would have enabled retrieval of this set.

Of the relevant documents, 77 were retrieved by the base query only, and 20 were retrieved only by the query by
section. The remaining 247 were retrieved by both queries. Querying by section returned 78% of the relevant
documents examined, and the base query returned 94%. Designing queries to examine only specific sections of the
text led to overlooking a number of relevant documents: fifty-two relevant documents were not retrieved because of
overly restrictive queries of specific sections of the document.

Looking at the 20 documents retrieved when querying by section revealed two topics these queries handled
especially well. One topic was "Patients with low-back pain who had imaging studies." The query for this topic
included the word *lumbar* as a synonym for *low back*. Adding this term resulted in retrieving a number of
documents in which patients had lumbar punctures, so the base query was adapted to eliminate documents with any
mention of that procedure. Querying by section eliminated documents that mentioned lumbar puncture only in the
lab and radiology results, allowing this query to place additional relevant documents higher in the ranked list. The
other topic handled well by querying by section was "Patients taking atypical antipsychotics without a diagnosis of
schizophrenia or bipolar depression." The base query looked for documents without these diagnoses, leading to
omission of documents in which non-relevant mention was made in the family or social history. The other query was
able to specify the absence of these in sections of the chart pertaining to the patient, while ignoring sections
mentioning diagnoses of relatives or other family members.

Queries of specific sections returned only 38% of the non-relevant documents examined, whereas the base queries
returned more than 97%. Only six non-relevant documents were retrieved only by querying specific sections, and
143 were retrieved by whole-document queries only. The remaining 81 were retrieved by both queries. Reasons for
returning non-relevant documents included mention of conditions that had been denied or ruled out, past conditions
or medications, future or possible conditions, non-relevant references to conditions (for example, in the family
history, or one word with multiple meanings), and, in one case, a procedure that was aborted prior to completion.
Because of their ability to search in specified sections of the chart, these queries were able to avoid retrieval of many
non-relevant documents.

Matthews correlation coefficient (MCC) was used to evaluate performance of individual queries. MCC is used in
machine learning as a measure of the quality of binary classifications. It was chosen for this project because it yields
reliable results with small samples and can measure both increases and decreases in performance. To calculate
MCC, retrieval results were first classified as true and false positives and negatives, indicating relevance to the topic
and score differential. Table 4, below, illustrates the classification of documents according to these criteria.

<table>
<thead>
<tr>
<th>Document relevant</th>
<th>Querying by section has higher score than querying whole document</th>
<th>Querying by section has lower score than querying whole document</th>
</tr>
</thead>
<tbody>
<tr>
<td>True Positive</td>
<td>False Negative</td>
<td>False Negative</td>
</tr>
<tr>
<td>False Positive</td>
<td>True Negative</td>
<td>False Positive</td>
</tr>
</tbody>
</table>

In some cases, only one type of query retrieved a document. For example, a document may be retrieved by the base
query but not the query of specific sections. In this case, the score assigned by the base query was used, and a score
of zero was used for the section query. Because of this, the MCC values do not reflect a difference between
documents that were retrieved and documents that were not retrieved. Possible values for MCC range from -1 to 1.
If querying by sections produces an overall performance decrease, MCC will be less than zero. A score of -1
indicates that querying by sections yields only false positive and false negative results. If querying by sections
increases performance, MCC will be greater than zero. An MCC of 1 indicates that querying by sections yielded
only true positive results. MCC was calculated for each topic as follows (scores listed in Table 5):
MCC could not be calculated for six topics because the sum of true negative and false negative or the sum of true positive and false positive was zero, resulting in a denominator of zero. The average of all MCC values was 0.422, which is significant at \( p<0.01 \). Although there was great variability in the values of MCC, the queries using sections for eight topics performed very well, showing statistical improvement over the base queries using Fisher’s exact test for significance, and no topics had statistically significant decreased performance. The queries of sections for Patients who had a carotid endarterectomy during this admission and Patients with diabetes mellitus who also have thrombocytosis performed very well with MCCs of 0.905 (\( p<0.001 \)). The sectioned query for Patients who develop thrombocytopenia in pregnancy had an MCC of 0.853 (\( p<0.001 \)). Other high performers were Patients with acute vision loss (MCC=0.756, \( p<0.001 \)), Patients who received pneumonia vaccination during this admission (MCC=0.734, \( p<0.01 \)), and Patients who have gluten intolerance or celiac disease (MCC=0.666, \( p<0.001 \)).

**Table 5.** Matthews correlation coefficients for each topic and Fisher's exact test for significance.

* \( p<0.05 \), ** \( p<0.01 \)

<table>
<thead>
<tr>
<th>Query</th>
<th>True Positive</th>
<th>False Positive</th>
<th>True Negative</th>
<th>False Negative</th>
<th>MCC</th>
<th>Fisher's exact</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patients who had a carotid endarterectomy during this admission</td>
<td>9</td>
<td>1</td>
<td>10</td>
<td>0</td>
<td>0.905**</td>
<td>0.0000595</td>
</tr>
<tr>
<td>Patients with diabetes mellitus who also have thrombocytosis</td>
<td>9</td>
<td>1</td>
<td>10</td>
<td>0</td>
<td>0.905**</td>
<td>0.0000595</td>
</tr>
<tr>
<td>Patients who develop thrombocytopenia in pregnancy</td>
<td>4</td>
<td>1</td>
<td>10</td>
<td>0</td>
<td>0.853**</td>
<td>0.00366</td>
</tr>
<tr>
<td>Patients with acute vision loss</td>
<td>10</td>
<td>0</td>
<td>16</td>
<td>4</td>
<td>0.756**</td>
<td>0.0000333</td>
</tr>
<tr>
<td>Patients who received pneumonia vaccination during this admission</td>
<td>10</td>
<td>0</td>
<td>7</td>
<td>3</td>
<td>0.734**</td>
<td>0.00155</td>
</tr>
<tr>
<td>Patients who have gluten intolerance or celiac disease</td>
<td>9</td>
<td>1</td>
<td>16</td>
<td>4</td>
<td>0.666**</td>
<td>0.00405</td>
</tr>
<tr>
<td>Patients with inflammatory disorders receiving TNF-inhibitor treatments</td>
<td>8</td>
<td>2</td>
<td>7</td>
<td>3</td>
<td>0.503*</td>
<td>0.0322</td>
</tr>
<tr>
<td>Patients with thyrotoxicosis treated with beta-blockers</td>
<td>8</td>
<td>2</td>
<td>6</td>
<td>4</td>
<td>0.408</td>
<td>0.0750</td>
</tr>
<tr>
<td>Patients with colon cancer receiving chemotherapy</td>
<td>10</td>
<td>0</td>
<td>10</td>
<td>15</td>
<td>0.400*</td>
<td>0.0178</td>
</tr>
<tr>
<td>Patients with delirium, hypertension, and tachycardia</td>
<td>8</td>
<td>2</td>
<td>6</td>
<td>6</td>
<td>0.311</td>
<td>0.130</td>
</tr>
<tr>
<td>Patients who developed disseminated intravascular coagulation in the hospital</td>
<td>7</td>
<td>3</td>
<td>11</td>
<td>9</td>
<td>0.236</td>
<td>0.139</td>
</tr>
<tr>
<td>Patients with low back pain who had imaging studies</td>
<td>20</td>
<td>0</td>
<td>2</td>
<td>18</td>
<td>0.229</td>
<td>0.244</td>
</tr>
<tr>
<td>Patients with dental caries</td>
<td>7</td>
<td>3</td>
<td>6</td>
<td>7</td>
<td>0.164</td>
<td>0.252</td>
</tr>
<tr>
<td>Patients with ventilator-associated pneumonia</td>
<td>7</td>
<td>2</td>
<td>5</td>
<td>15</td>
<td>0.0300</td>
<td>0.358</td>
</tr>
<tr>
<td>Patients who have cerebral palsy and depression</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>5</td>
<td>-0.149</td>
<td>0.367</td>
</tr>
<tr>
<td>Patients with esophageal cancer who develop pericardial effusion</td>
<td>5</td>
<td>5</td>
<td>3</td>
<td>7</td>
<td>-0.204</td>
<td>0.240</td>
</tr>
<tr>
<td>Patients taking atypical antipsychotics without a diagnosis of schizophrenia or bipolar depression</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Adults with Alzheimer's disease with pressure ulcers discharged to nursing homes</td>
<td>6</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Patients treated for the post-partum problems depression, hypercoagulability or cardiomyopathy</td>
<td>0</td>
<td>0</td>
<td>26</td>
<td>12</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Patients with HIV/AIDS who develop pancytopenia</td>
<td>0</td>
<td>0</td>
<td>19</td>
<td>1</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Patients admitted to the hospital with end-stage chronic disease who are discharged on hospice care</td>
<td>0</td>
<td>0</td>
<td>7</td>
<td>13</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Patients with left lower quadrant abdominal pain</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>0.000</td>
<td>0.000</td>
</tr>
</tbody>
</table>
Discussion
In the documents examined, searching for information in specific sections of the document did provide some improvement. Overall, queries of specific sections retrieved only about half the number of non-relevant documents as compared to the base queries. However, they also retrieved only about 80% of the relevant documents. This suggests that querying specific sections will not retrieve as large a set of relevant documents, but it is more likely to avoid retrieving non-relevant documents.

Expanding the queries to look at more sections for the desired information would result in retrieval of more documents, both relevant and non-relevant. Because Essie returns documents that match the search criteria, queries developed for this project were designed to retrieve documents that were most likely to be relevant. The queries utilizing the section headings were written to look only in the relevant section, resulting in two different queries for each topic. Some NLP search engines will identify each relevant item of information as well as the document section in which it was found. While the methodology in the present project may reflect real-world practice, using an NLP tool like this would allow the use of one set of queries and an evaluation that could focus more closely on the use of looking in specific sections. Improving retrieval of relevant documents could also be achieved through greater accuracy in labeling section headers. In the sample of documents validated for this project, better section detection may have allowed retrieval of 13 additional relevant documents, about 16% of the overlooked relevant documents.

Although these methods may improve the chance of retrieving relevant documents, it is important to keep in mind that the documents themselves are not perfect. Clinical text is a tool used to communicate medical information, and is often created in high-stress situations. The current examination found several cases where information was documented in non-typical sections. For example, one document noted the patient's Alzheimer's disease only in the social history when describing the living condition. Additionally, the sections themselves may vary slightly between institutions or clinicians. The current data set contained problem lists in some documents, while other documents listed ongoing problems only in the past medical history.

For some topics, the query must look for documents that do not reference a specific condition or medication, as in the topic "Patients taking atypical antipsychotics without a diagnosis of schizophrenia or bipolar depression." Searching for documents that do not reference these eliminates those that mention the conditions in other people. Being able to search specific sections improves recall in this situation by avoiding sections likely to contain false positives.

Future work should take several approaches. First, a formal evaluation of the section identification in these documents would provide a baseline to help interpret the results of the query evaluation. An alternative approach would be the use of a validated sectioning algorithm, which may provide greater accuracy in labeling section headings. The tool chosen should have the ability to identify section changes that are not explicitly identified by headings. Using regular expressions or other pattern matching algorithms would improve section detection and accuracy, which would improve overall performance.

Second, some of the common retrieval issues in medical text, such as negation, should be identified using a published tool. Next, a set of queries can be run, and a quantitative analysis of the results can be done to provide greater insight into the effectiveness of segmenting documents on retrieval.

Another avenue of future work is to develop algorithms that combine the two search types to predict document relevance. Because of the differences in recall and precision between the two types of searches, knowing which documents are retrieved by both searches provides a clue to predict relevance. Future work in this area would develop prediction models based on retrieval results.

The queries used in this project contained multiple components. Breaking down the cohort criteria into individual concepts would allow a more focused evaluation. For example, in the search for patients who have cerebral palsy and depression, one set of queries could look for cerebral palsy and another set for depression. This would allow an assessment of which individual concepts may be more accurately identified in specific sections of the document, and which can be accurately identified when searching the whole document.

Because of the high value of clinical text and the documentation burden on clinicians, pursuing further work in this area would provide great value to providers, healthcare organizations, and patients. Identifying and searching
specific sections of clinical text significantly improves precision with only modest decreases in recall, resulting in greater usability of this data source.

References


An Automated System for Categorizing Transthoracic Echocardiography Indications According to the Echocardiography Appropriate Use Criteria

Aaron S. Eisman, BS1, Rory B. Weiner, MD2, Elizabeth S. Chen, PhD1, Paul C. Stey, PhD1, Rishi K. Wadhera, MD, MPhil3, Aaron P. Kithcart, MD, PhD3, Indra Neil Sarkar, PhD, MLIS1

1Center for Biomedical Informatics, Brown University, Providence, RI, 2Massachusetts General Hospital, Boston, MA 3Brigham and Women’s Hospital, Boston, MA

Abstract

The Echocardiography Appropriate Use Criteria (EAUC) are a set of indications for transthoracic echocardiography (TTE) developed to guide physician decision making around ordering of TTE. In this study, an automated rule-based method for processing “indications” listed within TTE reports and classification into one of the major EAUC categories was developed and validated against a clinician-annotated reference standard. The system performed at a comparable level to trained physicians allowing for the automated classification of more than 30,000 TTE indications from a public database in less than ten minutes. The most common indication for TTE was Valvular assessment closely followed by General, Hypertension/Heart Failure/Cardiomyopathy, Acute, and Cardiac Structure assessment each contributed more than ten percent within this patient population. These results suggest potential for automated approaches for tracking appropriate use of TTE, as well as guide the development of systems for prospectively identifying when TTE use is recommended.

Introduction

Medicare alone spends more than $1B annually on transthoracic echocardiography (TTE) and the per capita growth rate in the performance of TTE increased steadily by 8% per year from the mid 1990s to early 2000s, a rate that exceeded any plausible increase in commensurate clinical value1,2. TTE is a powerful, non-invasive, low-cost imaging tool for evaluating cardiac structure and function. Trends towards evidence-based, targeted diagnostic testing highlight the potential for optimization of transthoracic echocardiography utilization. In response to this trend, a joint task force published a set of consensus echocardiography appropriate use criteria (EAUC) to define indications for TTE in 2007 and revised criteria were published in 20113. In the years since the EAUC were revised their use has been a component of echocardiography laboratory accreditation, necessary for insurance reimbursement, and also a way to promote quality improvement. However, the guidelines for how EAUC should be used are non-specific, only requiring that labs be aware of the EAUC, have access to them, and make concerted effort to improve the rates of appropriate testing while maintaining the caveat that the EAUC are “expert opinion” that are non-exhaustive in scope4. Changes to prior authorization requirements implemented with the Affordable Care Act were in part responsible for a 17% decline in spending between 2009 and 2013, however there is little evidence that this decline resulted in a higher rate of appropriate testing and there still exists an 8.5 fold variation in testing rates by geographic region in the United States5,6. All of this suggests the need to characterize physician TTE ordering practices for improving rates of appropriate clinical testing.

There have been concerted efforts to characterize TTE ordering behavior using the EAUC since they were published7. These studies have revealed insights into the rates of inappropriate testing and variability across centers, as well as the relative effectiveness of different interventional methods for impacting physician behavior and reducing rates of inappropriate testing. However, these studies have been limited by the need for manual classification of TTE appropriateness. This is a labor intensive process which is time consuming and limits the sample size available for analysis. The largest appropriate use TTE study consisted of approximately 3,000 TTEs classified in a manual fashion8.

To date, TTE appropriate use studies have relied on manual annotation of patient medical records to determine EAUC classifications; however, natural language processing (NLP) has been successfully applied to extract information from the medical record in other related ways. For example, one study analyzed the medical record for
necessary information required to calculate the Framingham Risk Score\textsuperscript{9}. Others have developed methods to extract clinical findings from echocardiogram reports\textsuperscript{10,11}. Tools like these have been shown to be able to be applied across health systems demonstrating their broad utility beyond the electronic health record system within which they were developed\textsuperscript{12}. Applying such technology to the study of the appropriate use of TTE could have far reaching implications as it would allow for classification of TTEs on a much larger scale and provide the opportunity to more readily study clinical practice patterns and the impact of educational and other interventions.

This study aimed to develop an automated rule-based method for processing the “indication” listed in a TTE report and classifying it into one of the broad, major EAUC categories. The technique was evaluated relative to a manually annotated reference standard derived from a publicly available dataset.

**Method**

TTE indications were extracted from a publically available dataset. The indication text was processed and then matched against a thesaurus of terms to determine the EAUC indication category for the TTE. This method was validated on a sample of expertly adjudicated indication categorizations and then applied to the entire dataset. Most of the processing was implemented using the Julia (v.0.5) programming language, with statistical calculations done using R (v.3.3.1).

![Image](image.png)

**Figure 1. Overview of Study**

**MIMIC-III Database**

The Medical Information Mart for Intensive Care III (MIMIC-III) is a publicly available de-identified database of critical care unit patient health information from a tertiary care hospital\textsuperscript{13}. MIMIC-III data range from provider notes and charts to imaging reports and survival information on approximately 60,000 critical care unit admissions at Beth Israel Deaconess Medical Center from 2001 to 2012. This study focused on patients who had at least one transthoracic echocardiogram report available in the “noteevents” table and associated demographic information from the “patients” table. TTEs were identified using the “category” column from “noteevents” where values containing “Echo” AND either the phrase “TTE (Complete)” OR “TTE(Complete)” were selected for analysis. Free text indications for performing each TTE were extracted from the reports as the text between the standardized header “Indication:” and special character marking the end of the line “\n.”

**Manual Classification of TTEs**

A reference standard was developed to evaluate the performance of the automated classification program. Using the EAUC and specific annotation guidelines (shown in Table 1), two cardiology fellows with prior experience researching TTE ordering practices performed independent classification of a random sample of TTE indications. Three overlapping training rounds were used to ensure consistency between reviewers. Discrepancies were discussed and adjudicated by a cardiology attending physician, and the annotation guidelines were updated after each training round to reflect consensus categorization and a Cohen’s Kappa statistic was calculated. Once the training rounds were performed, each reviewer independently categorized 450 TTEs to complete the reference set (n = 1200).
### Table 1. Annotation Guidelines

<table>
<thead>
<tr>
<th>Category</th>
<th>Scenario</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) General</td>
<td><strong>EAUC:</strong> Suspected Cardiac Etiology [or] Arrhythmias [or] Lightheadedness/Presyncope/Syncope [or] Evaluation of Ventricular Function [or] Perioperative Evaluation [or] Pulmonary Hypertension</td>
</tr>
<tr>
<td></td>
<td><strong>Additional Guidance:</strong> Ventricular Function as the only indication</td>
</tr>
<tr>
<td>(2) Acute</td>
<td><strong>EAUC:</strong> Hypotension or Hemodynamic Instability [or] Myocardial Ischemia/Myocardial Infarction [or] Evaluation of Ventricular function after ACS [or] Respiratory Failure [or] Pulmonary Embolism [or] Cardiac Trauma</td>
</tr>
<tr>
<td></td>
<td><strong>Additional Guidance:</strong> Mention of Chest Pain [or] MI [or] Pulmonary Embolism</td>
</tr>
<tr>
<td>(3) Valvular</td>
<td><strong>EAUC:</strong> Murmur or Click [or] Native Valvular Stenosis [or] Native Valvular Regurgitation [or] Prosthetic Valves [or] Infective Endocarditis</td>
</tr>
<tr>
<td></td>
<td><strong>Additional Guidance:</strong> Any mention of valve disease (incl. murmur) [or] Endocarditis</td>
</tr>
<tr>
<td>(4) Cardiac Structure</td>
<td><strong>EAUC:</strong> Suspected Cardiac Mass [or] Source of Embolism [or] Pericardial Effusion [or] Guidance for Percutaneous Coronary Intervention</td>
</tr>
<tr>
<td></td>
<td><strong>Additional Guidance:</strong> Stroke [or] Pericardial Effusion [or] TIA/Stroke/Embolism</td>
</tr>
<tr>
<td>(5) Aortic</td>
<td><strong>EAUC:</strong> Evaluation of the ascending aorta</td>
</tr>
<tr>
<td>(6) HTN/HF/CM (Hypertension/Heart Failure/Cardiomyopathy)</td>
<td><strong>EAUC:</strong> Hypertension [or] Heart Failure [or] Device Evaluation [or] Ventricular Assist Device and Cardiac Transplantation [or] Cardiomyopathies</td>
</tr>
<tr>
<td></td>
<td><strong>Additional Guidance:</strong> Chemotherapy or other Cardiotoxic Agent [or] Myocardiitis</td>
</tr>
<tr>
<td>(7) ACHD</td>
<td><strong>EAUC:</strong> Initial Evaluation or ongoing Surveillance of Adult Congenital Heart Disease</td>
</tr>
<tr>
<td>Cannot be Categorized</td>
<td>Does not clearly fit into one of the categories outlined by the EAUC</td>
</tr>
</tbody>
</table>

**Automated Classification of TTEs**

A rule-based computer program was developed in the Julia programming language to automatically classify TTEs into the categories outlined in the EAUC for echocardiography. Phrases and acronyms likely to indicate TTE ordering reason were generated using a combination of sampling MIMIC-III TTE indications (different from the reference set), consulting clinical experts, and SNOMED-CT concepts. These phrases were then associated with
EAUC indication categories and collectively make up a “TTE thesaurus” of 355 phrase-category associations, which were mapped to 179 clinical concepts. TTE indications were pre-processed and compared to each phrase in the thesaurus using a string matching algorithm, which were consequently mapped to one of the clinical concepts contained in the thesaurus.

Pre-processing prior to string matching consisted of converting text to lowercase and removing non-alphabetic characters (except for hyphens, which are commonly used in clinical text). Finally, “stop words” (common indefinite articles and prepositions such as “the”, “at”, or “on”) were also removed.

Processed TTE indications were then compared against the TTE thesaurus to determine likely matches (demonstrated in Table 2). For each processed TTE indication string, the matching algorithm employed a moving window of \( m \) words, where \( m \) is the number of words in the thesaurus phrase string (note that for the purposes of this study, an abbreviation is considered a “word”). The window was moved across the TTE indication phrase to compare the potential thesaurus phrases against substrings of the TTE indication. Comparisons of processed TTE strings to thesaurus substrings were done using Damerau-Levenshtein distance to evaluate the similarity of strings. The similarity threshold for declaring a match was dependent on the length of the strings being compared, with shorter strings requiring near-perfect similarly and longer strings allowing for some disagreement. A constraint was also imposed such that a match could not be declared unless the first character of each word in one string matched that of its positional counterpart in the other.

**Table 2. Sample of TTE Program Categorizations**

<table>
<thead>
<tr>
<th>Sample TTE Indications</th>
<th>Processed String ( \Rightarrow ) Thesaurus Match</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Mitral stenosis. Aortic regurgitation.”</td>
<td>“mitral stenosis” ( \Rightarrow ) Valvular</td>
<td>Valvular</td>
</tr>
<tr>
<td>“aortic regurgitation” ( \Rightarrow ) Valvular</td>
<td></td>
<td></td>
</tr>
<tr>
<td>“endocarditis”</td>
<td>“endocarditis” ( \Rightarrow ) Valvular</td>
<td>Valvular</td>
</tr>
<tr>
<td>“NSTEMI Myocardial infarction. Shortness of breath.”</td>
<td>“nstemi” ( \Rightarrow ) Acute</td>
<td>Acute</td>
</tr>
<tr>
<td></td>
<td>“myocardial infarction” ( \Rightarrow ) Acute</td>
<td></td>
</tr>
<tr>
<td></td>
<td>“shortness breath” ( \Rightarrow ) General</td>
<td></td>
</tr>
<tr>
<td>“Left ventricular function. Pericardial effusion.”</td>
<td>“left ventricular function” ( \Rightarrow ) General</td>
<td>Cardiac Structure</td>
</tr>
<tr>
<td></td>
<td>“pericardial effusion” ( \Rightarrow ) Cardiac Structure</td>
<td></td>
</tr>
<tr>
<td>“Evaluate for Left ventricular function/wall motion abnormalities/diastolic function.”</td>
<td>“left ventricular function” ( \Rightarrow ) General</td>
<td>HTN/HF/CM</td>
</tr>
<tr>
<td>Coronary artery disease. Hypertension. Diastolic Congestive heart failure.”</td>
<td>“wall motion” ( \Rightarrow ) General</td>
<td></td>
</tr>
<tr>
<td></td>
<td>“diastolic function” ( \Rightarrow ) General</td>
<td></td>
</tr>
<tr>
<td></td>
<td>“coronary artery disease” ( \Rightarrow ) General</td>
<td></td>
</tr>
<tr>
<td></td>
<td>“hypertension” ( \Rightarrow ) HTN/HF/CM</td>
<td></td>
</tr>
<tr>
<td></td>
<td>“congestive heart failure” ( \Rightarrow ) HTN/HF/CM</td>
<td></td>
</tr>
<tr>
<td>“Left ventricular function. New onset tachycardia. EKG changes.”</td>
<td>“left ventricular function” ( \Rightarrow ) General</td>
<td>General</td>
</tr>
<tr>
<td></td>
<td>“tachycardia” ( \Rightarrow ) General</td>
<td></td>
</tr>
<tr>
<td></td>
<td>“EKG changes” ( \Rightarrow ) General</td>
<td></td>
</tr>
</tbody>
</table>

*Typographical errors are highlighted*

TTE indication matches were then reduced to distinct clinical concepts using the TTE thesaurus and frequency of concept-category matches for each TTE indication were tabulated. TTE indications were then assigned an EAUC associated category based on the highest number of matches with the following exceptions. The General category was only chosen when no more precise category matches were found and due to the specificity of valvular indications, any mention of valve disease or endocarditis were annotated as “Valvular.” The automated rule-based method for classifying TTE indications according the EAUC was evaluated compared to the expert derived reference set. Level of agreement was quantified using Cohen’s Kappa. Precision, recall, and F-measure were all
calculated for each classification category. The program was then used to categorize all 30,320 TTE reports in MIMIC-III database.

Comparison Against a Machine Learning Classifier

The rule-based classification approach was compared to a random forest classifier. The target variable was the human-generated, adjudicated TTE classifications. The features were engineered using a bag-of-words approach on individual words from the indication text after the removal of "stop words" (same "stop words" as above). Term frequency–inverse document frequency (TF*IDF) was used for normalization. This yielded 603 sparse features. The 1200 adjudicated TTE classifications were split into training and test sets (80% and 20%, respectively). Five-fold cross validation was used to select the following meta-parameters: number of features to sample at each node, proportion of training set to be sampled for each tree. All trials used 1000 trees.

Results

Data for 30,320 TTEs from 18,487 different patients were analyzed with a mean of 1.6 (95% CI: [1-2]) TTEs per patient. Less than half (38%) of the TTEs were performed inpatient. Patients had a mean age of 68 years (95% CI: [56-79]), 56% were female, with a mean height of 67 inches (95% CI: [64-70]), weighing 170 lbs (95% CI: [144-201]). Mean patient BMI was 26.8 (95% CI: [23.4 - 31.3]).

Classification validation results are reported in Table 3. Two physicians classified a sample of TTE indications according to the EAUC across three training rounds in order to ensure consistency between them. Once consistency was achieved they each independently classification an additional 450 TTE indications. Collectively, 1200 expertly classified TTE indications comprise a reference set. The automated method was then tested against a small sample of the expertly adjudicated classifications before being applied to the balance of the reference set. Physician v. automated categorization validation round discrepancies are shown in Table 4. Precision, recall, and f-measure were calculated for each category and are reported in Table 5.

Table 3. Validation Statistics

<table>
<thead>
<tr>
<th></th>
<th>Comparison</th>
<th>Weighted Kappa</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Training Round 1 (n = 100)</strong></td>
<td>Physician v. Physician</td>
<td>0.76 (0.64 - 0.89)</td>
</tr>
<tr>
<td><strong>Training Round 2 (n = 100)</strong></td>
<td>Physician v. Physician</td>
<td>0.78 (0.65 - 0.90)</td>
</tr>
<tr>
<td><strong>Training Round 3 (n = 100)</strong></td>
<td>Physician v. Physician</td>
<td>0.84 (0.72 - 0.96)</td>
</tr>
<tr>
<td><strong>Program Testing (n = 50)</strong></td>
<td>Physician v. Computer</td>
<td>0.90 (0.79 - 1.00)</td>
</tr>
<tr>
<td><strong>Validation Round (n = 1150)</strong></td>
<td>Physician v. Computer</td>
<td>0.89 (0.86 - 0.91)</td>
</tr>
</tbody>
</table>
### Table 4. Validation Round Discrepancies

<table>
<thead>
<tr>
<th>Automated Categorization</th>
<th>General</th>
<th>Acute</th>
<th>Valvular</th>
<th>Cardiac Structure</th>
<th>Aortic Disease</th>
<th>HTN, HF, or CM</th>
<th>ACHD</th>
<th>Cannot be Categorized</th>
</tr>
</thead>
<tbody>
<tr>
<td>General</td>
<td>317</td>
<td>2</td>
<td>0</td>
<td>5</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Acute</td>
<td>30</td>
<td>126</td>
<td>0</td>
<td>8</td>
<td>0</td>
<td>12</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Valvular</td>
<td>4</td>
<td>7</td>
<td>325</td>
<td>7</td>
<td>1</td>
<td>14</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Cardiac Structure</td>
<td>8</td>
<td>5</td>
<td>0</td>
<td>111</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Aortic Disease</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>HTN, HF, or CM</td>
<td>69</td>
<td>4</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>124</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ACHD</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Cannot be Categorized</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
</tr>
</tbody>
</table>

### Table 5. Precision, Recall, and F-measure

<table>
<thead>
<tr>
<th></th>
<th>General</th>
<th>Acute</th>
<th>Valvular</th>
<th>Cardiac Structure</th>
<th>Aortic Disease</th>
<th>HTN, HF, or CM</th>
<th>ACHD</th>
<th>Cannot be Categorized</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>0.98</td>
<td>0.72</td>
<td>0.91</td>
<td>0.87</td>
<td>0.50</td>
<td>0.63</td>
<td>0.17</td>
<td>0.57</td>
</tr>
<tr>
<td>Recall</td>
<td>0.73</td>
<td>0.88</td>
<td>1.00</td>
<td>0.82</td>
<td>0.25</td>
<td>0.81</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>F-measure</td>
<td>0.84</td>
<td>0.79</td>
<td>0.95</td>
<td>0.84</td>
<td>0.33</td>
<td>0.71</td>
<td>0.29</td>
<td>0.73</td>
</tr>
</tbody>
</table>

A random forest classifier was generated to benchmark the rule-based method. The final model used 1000 trees, randomly sampled 250 features for splitting at each node, had a maximum of five observations in leaf nodes, and drew bootstrap samples from 70% of the training set for each tree. This model achieved a classification weighted Cohen’s Kappa of 0.89 (0.83 - 0.95).

After the automated classification program was validated against the reference set, it was applied to the entire MIMIC-III database. Frequency of each TTE category is reported in Table 6, showing that more than half of the TTEs were ordered for General (27%) or Valvular (30%) reasons. HTN, HF, or CM (17%), Acute (15%), and Cardiac Structure (10%) accounted for most of the remaining TTEs with Aortic Disease and ACHD at less than 1% combined. Less than 1% of the TTE indications could not be categorized automatically.

To assess the effect of accepting fuzzy matches in the algorithm, the program was rerun to only accept exact matches. Out of the entire dataset of 30,320 TTE indications, there were 4,511 with missed thesaurus phrase matches when fuzzy matching was not implemented. This resulted in 364 changes to classification categories, indicating redundancy in the information available in the TTE indication despite typographical errors.
Table 6. TTE Categories

<table>
<thead>
<tr>
<th>TTE Category</th>
<th>N (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) General</td>
<td>8334 (27%)</td>
</tr>
<tr>
<td>(2) Acute</td>
<td>4420 (15%)</td>
</tr>
<tr>
<td>(3) Valvular</td>
<td>8967 (30%)</td>
</tr>
<tr>
<td>(4) Cardiac Structure</td>
<td>3100 (10%)</td>
</tr>
<tr>
<td>(5) Aortic Disease</td>
<td>114 (&lt;1%)</td>
</tr>
<tr>
<td>(6) HTN, HF, or CM</td>
<td>5042 (17%)</td>
</tr>
<tr>
<td>(7) ACHD</td>
<td>132 (&lt;1%)</td>
</tr>
<tr>
<td>Cannot be Categorized</td>
<td>211 (&lt;1%)</td>
</tr>
</tbody>
</table>

Discussion

An automated rule-based method was developed in this study for classifying TTEs into one of seven categories outlined by the EAUC for echocardiography. The most common indication for TTE was Valvular assessment closely followed by General, HTN/HF/CM, Acute, and Cardiac Structure assessment each contributed more than ten percent within this patient population. The system performed at a comparable level as trained physicians, allowing the classification of more than 30,000 TTE indications from a public database in less than ten minutes. This task would have taken the same physicians more than 45 hours.

Validation testing elucidated that areas of disagreement between the computer program and reference set established by trained physicians were mostly limited to four out of 56 possible discrepancy types, accounting for more than 65% of discrepancies. Two of these included the computer selecting the more specific categories of Acute or HTN/HF/CM, whereas a trained physician selected the General category. The remaining common discrepancy areas included the computer selecting Acute or Valvular where the trained physician classified them as HTN/HF/CM. The tight clustering of discrepancies demonstrates known overlaps in the EAUC categories. For example, determining whether the assessment of ventricular function belongs in the General or HTN/HF/CM depends on whether or not there is associated clinical suspicion of heart failure. This information is not always available from the TTE report indication and therefore would require additional associated clinical data to make this classification with certainty.

It was found that a machine learning classifier (built using Random Forests) performed comparably to the rule-based method. However, machine learning approaches may be difficult to fully interpret and may therefore be viewed as being less transparent (i.e., the rules for a particular classification are not readily discernable). In addition, rule-based approaches may be more computationally efficient than machine learning approaches, which often require large training sets. For example, the Random Forest approach used in this study required 80% of the adjudicated data set for training. The success of the rule-based approach demonstrated in this study, suggests the potential scalability to development of similar methods for use in other clinical contexts.

Categorization of TTE indications according to the EAUC have previously involved relatively small sample sizes largely due to the time required for manual classification. Research to date has utilized the EAUC in order to track rates of appropriate testing or as an outcome measure for quality improvement. These studies have reported that a significant number (15-25%) of TTEs are ordered inappropriately, with the majority of rarely appropriate TTEs being ordered in the outpatient setting. Randomized control trials of various combinations of educational intervention, clinical decision support, and audit/feedback mechanisms have demonstrated the ability to improve rates of appropriate use with limited sustained effect. Applying an automated tool such as the one developed in the current study would allow for the evaluation of appropriate use of TTE to occur on a much larger scale. This would
likely have several important clinical implications, including more accurate characterization of current practice patterns in various settings (e.g., academic, community, and inpatient/outpatient) and determining the impact of educational and other interventions.

Generalizability of the breakdown of TTE ordering indications reported in this study are limited by the patient population. All TTE data within the MIMIC-III database were either obtained during or subsequent to a hospitalization in which the patient received care in an ICU. Other limitations to this study are a function of the data being used and how the EAUC are written. TTE indications are brief notes that lack clinical context provided by access to a complete medical record. Accurately distinguishing between some overlapping indications would require this additional information when TTE indications are non-specific.

Future research will use this classification algorithm to guide automated EAUC classification using additional clinical data including physician notes, discharge summaries, and lab testing results to determine appropriateness. The development of a system to automatically determine TTE appropriateness according to the EAUC will enable sustainable audit/feedback intervention that can be applied to any electronic medical record system at the point of care. Finally, an automated tool may facilitate appropriate use tracking for individual laboratories and improve the ability to comply with standards set by accrediting bodies.

Conclusion

More than 30,000 TTEs were classified according to the EAUC for echocardiography using a computer program that was validated against trained physicians. The promising results suggest that there is potential for an automated approach to track appropriate use of TTE, as well as guide the development of systems for prospectively identifying when TTE use is recommended. Automated assessment of EAUC adherence may enable effective and sustainable intervention to improve appropriate use rates and efficiency in the application of diagnostic echocardiography in the care of patients.

Acknowledgments

This work was supported in part by the Scholarly Concentration Program at The Warren Alpert Medical School of Brown University.

References

Hybrid Semantic Analysis for Mapping Adverse Drug Reaction Mentions in Tweets to Medical Terminology

Ehsan Emadzadeh, Ph.D1, Abeed Sarker, Ph.D2, Azadeh Nikfarjam, Ph.D3, Graciela Gonzalez, Ph.D2

1Department of Biomedical Informatics, Arizona State University, Scottsdale, AZ; 2Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, Philadelphia, PA; 3 Department of Biomedical Informatics, Stanford University, Stanford, CA.

Abstract

Social networks, such as Twitter, have become important sources for active monitoring of user-reported adverse drug reactions (ADRs). Automatic extraction of ADR information can be crucial for healthcare providers, drug manufacturers, and consumers. However, because of the non-standard nature of social media language, automatically extracted ADR mentions need to be mapped to standard forms before they can be used by operational pharmacovigilance systems. We propose a modular natural language processing pipeline for mapping (normalizing) colloquial mentions of ADRs to their corresponding standardized identifiers. We seek to accomplish this task and enable customization of the pipeline so that distinct unlabeled free text resources can be incorporated to use the system for other normalization tasks. Our approach, which we call Hybrid Semantic Analysis (HSA), sequentially employs rule-based and semantic matching algorithms for mapping user-generated mentions to concept IDs in the Unified Medical Language System vocabulary. The semantic matching component of HSA is adaptive in nature and uses a regression model to combine various measures of semantic relatedness and resources to optimize normalization performance on the selected data source. On a publicly available corpus, our normalization method achieves 0.502 recall and 0.823 precision (F-measure: 0.624). Our proposed method outperforms a baseline based on latent semantic analysis and another that uses MetaMap.

Introduction

Pharmacovigilance is defined as “the science and activities relating to the detection, assessment, understanding and prevention of adverse effects or any other possible drug-related problems”.1 The primary focus of pharmacovigilance is the monitoring of adverse drug reactions (ADRs). Due to the various limitations of pre-approval clinical trials, it is not possible to assess all the consequences of the use of a particular drug before it is released.2 Therefore, ADRs caused by prescription drugs is currently considered to be a major public health problem and various ADR monitoring mechanisms are currently in place, such as voluntary reporting systems, electronic health records, and, relatively recently, social media.3

Social media has emerged as an important source of information for various public health monitoring tasks. The increasing interest in social media is largely because of the abundance of data in the multitude of social networks—data that is directly generated by a vast number of consumers. It is estimated that about 75% of all U.S. adults have a social network account and about 50% worldwide. Data from social networks have been used in the past for a variety of tasks such as studying smoking cessation patterns on Facebook,4 identifying user social circles with common medical experiences (like drug abuse),5 and monitoring malpractice.6 In addition, recent research has utilized social media for the monitoring of ADRs from prescribed medications.7 From the perspective of pharmacovigilance, social media could be a platform of paramount importance, since it has been shown in past research that users discuss their health-related experiences, including use of prescription drugs, side effects and treatments on a regular basis. Users are known to share their personal experiences over social media sources, and as such, a large amount of health-related knowledge is generated within the realm of social media.7 However, while significant progress have been made in ADR text classification, and ADR mention extraction,7,8 the normalizing of user posted ADR mentions into a predefined set of concepts is still a largely unaddressed problem.

In this paper, we address the task of normalizing distinct ADR mentions to standardized concepts. This is an essential task given that the same ADR concept may have multiple lexical variants (e.g., “high blood pressure”, and “hypertension”). Therefore, following automatic ADR extraction approaches, automatic normalization techniques
must be applied to obtain realistic estimates for the occurrences of ADRs. For social media data, this is particularly important because users often tend to express their problems using non-standard terms. For example, consider the following user posts:

"<DRUG NAME> makes me having the sleeping schedule of a vampire.",
"<DRUG NAME> evidently doesn't care about my bed time",
"...wired! Not sleeping tonight. #<DRUG NAME>"

In the above examples, all the posts are referring to sleeplessness caused by the intake of a specific medication. Each expression is unique and non-standard, although referring to the same ADR concept. In the corpus that we use, each ADR concept is encoded using the Unified Medical Language System (UMLS) concept IDs. All standard and non-standard lexical variants of an ADR are mapped to the most appropriate UMLS entry. The target of our approach, formally, is to predict the UMLS concept ID of a text-based ADR mention.

**Concept normalization**

The task of normalization of ADRs involves assigning unique identifiers to distinct ADR mentions, with different lexical variants of the same concept. The IDs are derived from any lexicon or knowledge base with sufficient coverage. In the case of our research, we use the UMLS concept identifiers to uniquely specify each ADR concept. The UMLS provides a vast vocabulary of medical concepts and the semantic groups into which the concepts can be classified. Each UMLS concept is assigned a unique ID, which represents all the lexical variants of the concept. From the previously mentioned example, all synonyms of the concept hypertension (e.g., hypertensive disorder, high blood pressure, high bp and so on) are assigned the ID c0020538. The UMLS Metathesaurus, due to its comprehensive coverage of medical terminologies, has been used to build corpora specialized for normalization in the past.9–11

The task of medical concept normalization can be regarded as a sub-field of biomedical named entity recognition (NER). Due to the abundance of text based medical data available, NER and concept normalization have seen growing research in the medical domain primarily through challenges such as BioCreative,12 BioNLP,13 TREC,14 and i2b2.15 Building on from these initiatives, the problem of concept normalization has seen substantial work for genes and proteins. Majority of the research on concept normalization relies on some variants of dictionary lookup techniques and string matching algorithms. Machine learning techniques have recently been employed, but mostly in the form of filtering techniques to choose the right candidates for normalization.16 A number of approaches17 rely on the use of tools/lexicons such as MetaMap18 as a first step for the detection of concepts. Due to the advances in machine learning techniques and also the increasing availability of annotated data, recent approaches tend to apply learning based algorithms to improve on banal dictionary lookup techniques. Very recently, Leaman et al.19 applied pairwise learning from a specialized disease corpus for disease name normalization. Prior works have involved list-wise learning, which learn the best list of objects associated with a concept and return the list rather than a single object, for tasks such as gene name normalization,19,20 graph-based normalization,21 conditional random fields,22 regression based methods,23 and semantic similarity based techniques.24 Semantic similarity or relatedness is a measure that shows how similar two concepts are. Such measures are often used for word sense disambiguation,25 where the term and its context information are utilized to assign a meaning to it. A number of techniques for computing semantic relatedness among medical entities have been proposed and compared in the past,26 some of which are mentioned in the next section. However, to the best of our knowledge, measures of semantic relatedness have not been previously used for normalizing ADR mentions.

**Social media text normalization**

While the task of normalization of medical concepts is itself quite challenging, in our case, the problem is exacerbated by the fact that our data originate from social media. Social media data is notoriously noisy.27 And while this hampers the performance of natural language processing (NLP) techniques, it is also the primary motivation behind the implementation of techniques for automatic correction and normalization of medical concepts in this type of text. Typos, ad hoc abbreviations, phonetic substitutions, use of colloquial language, ungrammatical structures and even the use of emoticons make social media text significantly different from texts from other sources.27
Past work on normalization of social media text focused at the lexical level, and has similarities to spell checking techniques with the primary difference that out of vocabulary terms in social media text are often intentionally generated. Text messages have been used as input data for normalization models, and various error models have been proposed, such as Hidden Markov Models and noisy channel models. Similar approaches targeted purely towards lexical normalization have been attempted on social media texts as well. For the research task we describe in this paper, although the primary goal is to perform concept level normalization, we apply several preprocessing techniques to perform lexical normalization before the application of our concept normalization pipeline.

Methods

The goal of this normalization task is to find the UMLS concept ID related to a text segment in a tweet that is pre-tagged as an ADR. For example, in the tweet: "had 2 quit job: tendons in lots of pain," the phrase "tendons in lots of pain" is tagged as an ADR. The goal of our system is to normalize the annotated text to a concept in UMLS, which in this example is "c0231529-tenalgia". Figure 1 shows the overall pipeline of the proposed normalization system. The system consists of syntactic and semantic matchers, synonym normalization and evaluation components. The pipeline is sequential, and so, as soon as a matching module finds a match between a lexical component and a UMLS concept, the remaining matchers in the pipeline are skipped and the flow goes to the synonym normalization and evaluation components.

For evaluation, we use a publicly available, annotated corpus of 2008 tweets mentioning drugs and adverse reactions. The corpus was generated by using Twitter API to search for tweets that contain the names of selected drugs. The dataset includes 1544 annotations using 345 unique concepts, of which 1272 are ADRs, 239 are indications/symptoms and 32 are medications. The annotations were performed by two trained biomedical informatics annotators, and all disagreements were resolved and the final corpus was validated by a pharmacology expert. In this work, we did not differentiate between annotation types (e.g., ADR vs. indication) and attempted to normalize all types using the same pipeline. More information about the corpus and annotations can be found in the publication associated with this dataset.

Figure 1. Overall architecture of the Hybrid Semantic Analysis technique.
**Syntactic match**

The first step in our normalization pipeline involves syntactic or lexical matching with concept names in UMLS. This part of the pipeline involves two steps: *exact match* and *definition match*. An exact match happens when an ADR mention in a user post exactly matches a UMLS concept name (i.e., when the user uses a standard lexical expression for a concept). This simple matching technique can detect many easy matches as standard terminologies are often used by the users. However, in many cases in informal text, ADR mentions are misspelled, and exact matches are not possible. Some of these misspellings can be caught by simple pre-processing techniques. For example, unnecessary character repetition can be removed, as in the tweet "I feel siiiiiiiiiiiiick", "siiiiiiiiiiiiick" is matched with UMLS concept "c0231218-sick".

The next step in syntactic matching utilizes the formal definitions of UMLS concepts. The UMLS metathesaurus provides one or more definitions for each concept. The definition is a passage that describes the concept in plain English. We use this information, in the semantic similarity component later in the pipeline, to create semantic vectors and calculate the similarity values. In the syntactic matching module, we check if the mention appears in the definition of a single concept only in UMLS. If it does, the mention is normalized to the concept. In most of the cases, a phrase appears in the definition of many concepts and no conclusion can be made.

**Semantic match**

ADR concepts that are not normalized by the syntactic matching components are passed on for semantic matching. The primary task of this component is to compute the similarities of potential ADR concepts with the UMLS concepts. We experiment with two Measures of Semantic Relatedness (MSR) methods. MSR methods or kernels are functions that accept a pair of phrases/words as input, and return a numeric value representing the relatedness score of the inputs. In this module, an MSR method is used to find semantic similarity of a mention and a subset of concepts in UMLS. The most similar concept, with a similarity above a specified threshold, will be chosen as the concept of the mention. We evaluated Latent Semantic Analysis (LSA), and our proposed hybrid method. In the semantic matching modules, only the UMLS concepts that are used in the annotations are considered for the prediction.

**Latent semantic analysis**

Latent Semantic Analysis (LSA) uses a term-document frequency matrix to estimate semantic similarity of two segments of text. LSA then harvests the matrix using Singular Value Decomposition (SVD), by selecting the $k$ best SVD values. More details about LSA technique and various weighting techniques can be found in past publications. In our system, for the first step, the term vector space is generated from a corpus of plain text documents. Then this vector space is used to find a representative vector for each UMLS concept. The UMLS concept names are used to search for term vectors in the vector space. We evaluate some of the corpora, which are listed in Table 1 for creating the representative vectors for the UMLS concepts. After finding a representative vector for each UMLS concept, we search for a representative vector for each annotated text in the same vector space. The cosine similarity of each concept's vector and the annotated text’s vector is computed. The concept with the highest similarity to the ADR is chosen as the normalized concept if the cosine similarity is above a certain threshold ($\geq 0.8$).

**HSA**

HSA uses machine learning to find the optimal combination of semantic relatedness function scores for each context, based on a set of calculated features. Using different free text sources, semantic representations for the concepts are learned, and then the different MSR scores are computed for the social media based lexical representations and the standard lexical representations of the concepts. Since different resources can be used, each MSR can return different values when applied to different resources. For example, we can apply PMI, and distinct resources like PMI-GENIA and PMI-I2B2ClinicalNotes. Each MSR method returns different scores when trained on different resources, and these scores are combined in a regression function as features.

For each pair of words/phrases, the Feature Calculator component of the system computes feature scores, which are the returned values from each MSR consisting of different corpora combinations. For example, one feature can be semantic relatedness returned for a pair by LSA-I2B2ClinicalNotes. After feature calculation, the regression model (SVM) is trained, and the model is evaluated against the test set.
Since the MSR and information resources are dynamic, and can be added or removed from the system, the feature set for the regression function is also dynamic and can be varied depending on the task. The output of regression function is a semantic relatedness score of two concepts in the given context, and the regression function is optimized using the labeled training data. The method is designed to easily adapt to new knowledge sources or corpora and adjust parameters accordingly. Also, it can be trained for a new text type or entity type.

For training the regression model, we prepared the training set from a subset of annotation (50% of the annotation). For each annotated text we created training examples for the annotated text and the UMLS concept names of the assigned concept with expected similarity of 100. For each annotation we generated 10 negative examples from the annotated text to random concepts in UMLS with expected similarity of 0. Figure 2 shows an example how HSA training examples are generated.

The ratio of negative to positive examples can affect how the HSA regression model is trained. We used SVM with a linear kernel as the regression model, and trained HSA with the resources listed in Table 1 and LSA as the only MSR. SVM (SVMLight) was used to create the regression model but other models such as neural network can be used and explored. We refrained from adding additional MSR methods as the intent of this experiment to study the effect of using the regression model with a single MSR and various additional resources. These resources are described in the next section.

After HSA is trained, the regression model is used to calculate the similarity of annotated phrases as ADR (which are the input to our system) to UMLS concept names. First, testing instances between the phrase and a set of selected UMLS concepts names are created. To limit the search space, UMLS concepts appeared in the training set annotations with frequencies of three or more are used for creating the test instances for HSA. Following that, for each example, the features are calculated. The features are all possible MSR and resource combinations defined in the system setup (e.g., LSA with PubMed). Next, the regression model is run on the test instances to calculate the similarity of the annotated text and each UMLS concept. The concept with the highest similarity and above a certain threshold (\(\geq 90\)), note that the maximum and minimum similarities in the training set are 0 and 100) is chosen as the normalized concept. Since the method has to calculate several semantic similarities for normalizing each annotated text, the process is slower than using a single MSR. The output of the trained regression is not normalized to any boundary and can be any real value.

**Corpora**

The two semantic matching techniques discussed above require data from suitable corpora to generate their models. We used three textual corpora generated from three different queries on PubMed (provided as special queries: http://www.nlm.nih.gov/bsd/special_queries.html): Dental Journals (PubMed query: “(jsubsetd[text]))”, Nursing Journals (PubMed query: “(jsubsetn[text])”) and Systematic Reviews. We filtered out articles that do not have publicly available abstracts. Table 1 shows the number of documents in each corpus. We are also interested in evaluating additional corpora instead of only those generated from PubMed. HSA uses all of the corpora matched with LSA as features to train the hybrid model. When using LSA independently for evaluation, without HSA, only one corpus is used for each run.

For the semantic similarity match step, we evaluate the following different settings:

1. Most similar concept returned by LSA using each of the corpora listed in Table 1.
2. Most similar concept returned by HSA
Evaluation criteria

For strict evaluation, we consider a prediction correct when the predicted concept is exactly the same as the expected concept. In contrast, in the relaxed evaluation mode, before calculating the evaluation metrics, we change the predicted class to the expected class if the predicted class has any of these relationships in the UMLS: "synonym", "is-a", "mapped-to" relations with the expected class. This means that if the system predicts a concept which is, for example, the synonym, child or parent of the expected concept, we consider it as a true positive. Considering the size of UMLS graph, we only do this normalization by distance of 2—meaning that if a concept “A” has an ‘is-a’ relation with a concept “B”, and the concept “B” has a “mapped-to” relation with a concept “C”, the concept “A” and “C” would be considered the same for the evaluation purpose. The following list shows some other examples of match in the relaxed evaluation:

- A –(is-a)→ B –(is-a)→ C: A will match with C
- A –(is-a)→ B –(mapped-to)→ C: A will match with C
- A –(mapped-to)→ B –(mapped-to)→ C: A will match with C
- A –(mapped-to)→ B –(synonym)→ C: A will match with C

<table>
<thead>
<tr>
<th>Term Count</th>
<th>Document Count</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>PubMed Dental Journals</td>
<td>182641</td>
<td>236767</td>
</tr>
<tr>
<td>PubMed Nursing Journals</td>
<td>74000</td>
<td>72494</td>
</tr>
<tr>
<td>PubMed Systematic Reviews</td>
<td>219656</td>
<td>214252</td>
</tr>
<tr>
<td>BioNLP Corpus</td>
<td>9483</td>
<td>908</td>
</tr>
<tr>
<td>Reuters Corpus</td>
<td>105675</td>
<td>694335</td>
</tr>
<tr>
<td>ADR-Tweets Corpus</td>
<td>6205</td>
<td>2008</td>
</tr>
<tr>
<td>UMLS Definitions</td>
<td>103933</td>
<td>188647</td>
</tr>
</tbody>
</table>

Table 1. Resources used by HSA for the experiments described in this paper.

Results

From the perspective of evaluation, each UMLS concept is considered to be a class. We compute the final precision, recall, and F-measure as the micro-average of all the classes. For each class true positive (TP), false positive (FP), true negative (TN) and false negative (FN) are defined as below: TP is when the expected class is equal to the predicted class and the evaluated class. FP is when the predicted class is equal to the evaluated class but not equal to the expected class. FN is when the expected class is equal to the evaluated class but the predicted class is not equal to the expected class.
class. TN is when both predicted and expected classes are not equal to the evaluated class. The following table illustrates an example for the evaluation strategy. The micro-averaged precision and recall are calculated using the following formula:

\[
\text{Precision} = \frac{\sum_{c \in \text{Classes}} TP_c}{\sum_{c \in \text{Classes}} (TP_c + FP_c)}
\]

\[
\text{Recall} = \frac{\sum_{c \in \text{Classes}} TP_c}{\sum_{c \in \text{Classes}} (TP_c + FN_c)}
\]

\[
F\text{—}measure = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}
\]

<table>
<thead>
<tr>
<th>Mention</th>
<th>Expected Class</th>
<th>Predicted Class</th>
<th>Evaluated class</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1</td>
<td>Class1</td>
<td>Class1</td>
<td>TP</td>
</tr>
<tr>
<td>M2</td>
<td>Class1</td>
<td>Class2</td>
<td>FN</td>
</tr>
<tr>
<td>M3</td>
<td>Class2</td>
<td>Class1</td>
<td>FP</td>
</tr>
<tr>
<td>M4</td>
<td>Class2</td>
<td>Class2</td>
<td>TN</td>
</tr>
</tbody>
</table>

Table 2. Illustration of the evaluation technique.

Table 3 shows the results for syntactic matcher, LSA using different corpora and the proposed hybrid model. HSA yielded the best F-measure of 62.37 and the best recall of 50.20. The next best precision after syntactic match is achieved by LSA with UMLS definitions corpus. Among LSA with various corpus, ADR-Tweets resulted in the best F-measure. In the investigated normalization problem, the ADR-Tweets corpus yielded the best performance for LSA method. Syntactic matcher has the highest precision, which was expected. Adding LSA-ADR-Tweets matcher on top of syntactical matcher decreases the precision but increases the recall resulting in a higher F-measure. Using HSA instead of LSA decreases the precision slightly more than LSA but the gain on recall is higher and results in a higher F-measure. MetaMap, which is designed for public medical literature data, suffers from very poor recall and therefore overall F-measure. We used the relaxed evaluation method in all of the reported results.

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Syntactic</td>
<td>88.0</td>
<td>35.7</td>
<td>50.8</td>
</tr>
<tr>
<td>LSA-PubM-Dental</td>
<td>83.6</td>
<td>38.2</td>
<td>52.4</td>
</tr>
<tr>
<td>LSA-PubM-Nursing</td>
<td>83.1</td>
<td>38.6</td>
<td>52.7</td>
</tr>
<tr>
<td>LSA-UMLS-Defs</td>
<td>86.5</td>
<td>40.3</td>
<td>55.0</td>
</tr>
<tr>
<td>LSA-Reuters</td>
<td>81.5</td>
<td>44.9</td>
<td>57.9</td>
</tr>
<tr>
<td>LSA-PubM-Systematic</td>
<td>83.6</td>
<td>44.4</td>
<td>58.0</td>
</tr>
<tr>
<td>LSA-ADR-Tweets</td>
<td>84.6</td>
<td>47.7</td>
<td>61.0</td>
</tr>
<tr>
<td>MetaMap</td>
<td>82.6</td>
<td>18.7</td>
<td>30.5</td>
</tr>
<tr>
<td>HSA</td>
<td>82.3</td>
<td>50.2</td>
<td>62.4</td>
</tr>
</tbody>
</table>

Table 3. Results obtained by our system using the proposed pipeline and the relaxed evaluation technique.

Discussion

Figure 3 shows the sources of false positives and true positives. Semantic match generates most of false positives followed by exact match. Exact match returns majority of true positives followed by semantic match. As expected, when only the syntactic matching module is employed, we obtain high precision but very low recall. Searching for exact match in definition helps to find alternative representation of the concept. For example, “urge to vomit” is normalized correctly to “c0027497-Nausea” when we search the definition of “c0027497”: “unpleasant sensation in the stomach usually accompanied by the urge to vomit”. Exact match fails when the words in a phrase are expressed
in complex orders and another concept matches exactly with the annotated phrase. For example, in the following tweet: "dreams have taken a terrifying turn.", "dreams" is annotated as "c0857051-bad dreams" but exact match matches the phrase with "c0028084-dreams".

![Figure 3. Sources of correct and incorrect predictions.](image)

In contrast to syntactic matching, semantic methods are designed to compute estimates of similarity, and match concepts that are not necessarily the same, but are similar. As such, they are expected to have high recall. In our experiments, the semantic matchers LSA and HSA have the higher numbers of false positives but yield higher recalls than syntactic match. This was expected since most of the hard to normalize concepts reach the semantic matchers modules. Most of the errors are caused by concepts with very similar meanings. For example, "anti-depressant" in a tweet is tagged as "c0011570-mental depression," but LSA returns "c0005586-manic depression" as the most similar concept.

Table 4 shows examples of correct and incorrect predictions by HSA. The hybrid model is very good at normalizing when the same word is represented in a different variation ("antidepressant" vs. “depression”) or match similar words which appear frequently in corpora (“fewer” vs. “loss”, “increase” vs. “gain”). In contrast, HSA performance is limited to the information in the provided resources and MSR technique. Since in this experiment we only used LSA, HSA would perform solely based on co-occurrences of terms in the resources. If there are not enough numbers of co-occurrences of two terms in the provided resources, we expect to have a very low similarity of the terms. In addition to using larger corpora, adding more diverse techniques that can leverage other resource types (such as graph-based techniques) can significantly boost this limit.

**Conclusion**

In this work, we proposed a natural language processing pipeline for the problem of normalizing extracted mentions of ADRs from colloquial texts to UMLS concepts. We compared two semantic similarity techniques: LSA and a proposed hybrid approach (HSA). The hybrid approach shows improvement over a single similarity technique (LSA). The proposed hybrid approach is supervised and benefits from training data while LSA is unsupervised and does not have any training. Tweets, like other informal texts, required heavy pre-processing and cleaning. The errors of the system could be reduced by applying more advanced pre-processing like spelling correction. This is the first effort towards the ADR normalization from social media or other noisy text sources, and can provide a baseline for future work.

In the future, we will utilize our tool to perform normalization on a larger set of mentions. We will also incorporate more complex NLP preprocessing techniques, such as negation detection, and perform comparisons of our approach with a larger number of semantic similarity measurement approaches. In the recent past, approaches using distributed representations of words have become very popular, and the use of such representations along with deep neural networks have outperformed past benchmark systems in a variety of natural language processing tasks. Therefore, we will attempt to utilize annotated data to develop and evaluate such neural network based systems.
against ours. Because of the modular implementation of our system, if such methods provide promising results, we
will incorporate them as a module in our concept normalization pipeline.

<table>
<thead>
<tr>
<th>Annotated Phrase</th>
<th>Expected</th>
<th>Predicted</th>
</tr>
</thead>
<tbody>
<tr>
<td>Antidepressant</td>
<td>c0011570-Depression</td>
<td>c0011570</td>
</tr>
<tr>
<td>increase my weight</td>
<td>c0043094-Weight gain</td>
<td>c0043094</td>
</tr>
<tr>
<td>gain so much weight</td>
<td>c0043094-Weight gain</td>
<td>c0043094</td>
</tr>
<tr>
<td>fewer hours sleep</td>
<td>c0235161-Sleep loss</td>
<td>c0235161</td>
</tr>
<tr>
<td>feel like need to throw up</td>
<td>c0027497-Nausea</td>
<td>c0917799-Hypersomnia</td>
</tr>
<tr>
<td>just eat, and eat</td>
<td>c0232461-Apetite increase</td>
<td>c0015672-Fatigue</td>
</tr>
<tr>
<td>falling asleep every day</td>
<td>c0541854-Daytime sleepiness</td>
<td>c0917801-Insomnia</td>
</tr>
<tr>
<td>it's 4:30am. at this point ima just throw out a big</td>
<td>c0917801-Insomnia</td>
<td>c0917799-Hypersomnia</td>
</tr>
<tr>
<td>&quot;f*** you&quot;</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4. Examples of correct and incorrect predictions by HSA. The first four rows are correct predictions followed
by three rows of incorrect predictions. The last row is correct based on the relaxed evaluation criteria.

References


Word-of-Mouth Innovation: Hypothesis Generation for Supplement Repurposing based on Consumer Reviews

Jung-wei Fan, PhD1,2, Yves A. Lussier, MD1,2,3
1Department of Medicine; 2Center for Biomedical Informatics & Biostatistics; 3BIO5 Institute, The University of Arizona, Tucson, Arizona, USA

Abstract

Dietary supplements remain a relatively underexplored source for drug repurposing. A systematic approach to soliciting responses from a large consumer population is desirable to speed up innovation. We tested a workflow that mines unexpected benefits of dietary supplements from massive consumer reviews. A (non-exhaustive) list of regular expressions was used to screen over 2 million reviews on health and personal care products. The matched reviews were manually analyzed, and one supplement-disease pair was linked to biological databases for enriching the hypothesized association. The regular expressions found 169 candidate reviews, of which 45.6% described unexpected benefits of certain dietary supplements. The manual analysis showed some of the supplement-disease associations to be novel or in agreement with evidence published later in the literature. The hypothesis enrichment was able to identify meaningful function similarity between the supplement and the disease. The results demonstrated value of the workflow in identifying candidates for supplement repurposing.

Introduction

For more than a decade, there has been fervent interest in drug repurposing (or repositioning) due to the expensive and time-consuming process of drug development. It was estimated that on average it takes 10 years and at least $1 billion to bring a drug to market. On the other hand, finding a new use for an old drug has advantages such as shortened time in verifying toxicity and higher probability of government approval. Renowned examples include sildenafil for erectile dysfunction and thalidomide for erythema nodosum leprosum. Interestingly, a relatively overlooked territory along this repurposing movement is dietary supplements, which are widely available off the shelf and may offer great variety of functions.

According to a 2015 survey by the Council for Responsible Nutrition (CRN)2, about 68% of Americans use dietary supplements. Among these supplement users, 69% said their doctor talked to them about the benefits of taking supplements. Additionally, most supplement users aged 18-34 (66%) anticipate an increased use over the next 5 years. Dietary supplements come in diverse categories: minerals, vitamins, amino acids, enzymes, herbals and botanicals. These facts imply an extensive influence of dietary supplements over population health and well justify them as a rich source for discovering novel repurposing candidates. In fact, there has been ongoing “secondary repurposing” effort that searches for analogues of already repositioned chemicals (e.g., sildenafil3,4) in dietary supplements. There are tremendous opportunities for informatics to systematically facilitate the hypothesis generation process.

In our previous work of mining health-related issues in 1.3 million Amazon.com consumer reviews5, we found that about 40% of the reviews described certain health benefits from using a grocery food product. The Amazon online store is one of the largest retailers in the United States, with over 300 million users and net sales of $136 billion (in 2016)6. As of 2017 March 6, its numbers of supplements sold by categories were: 19,619 Multi & Prenatal Vitamins, 16,048 (specific) Vitamins, 10,255 Minerals, 35,584 (non-herbal) Supplements, 42,519 Herbal Supplements, and 12,370 Weight Loss products. Given the huge collection of supplements and user base, the consumer feedback may actually serve as an ideal data source for automated surveillance. Currency is another merit of the data, as the reviews are constantly growing and closely reflect the products that people use.

Although there are issues with using consumer reviews (e.g., credibility and sparseness), we believe it is worth developing solutions that can bring any meaningful signal to our attention for advancing science. In this study, we proposed and exercised a dry run of a discovery workflow. Text mining was applied to screen for unexpected benefits of dietary supplements in the consumer reviews, followed by manual curation of the candidate reviews, and hypothesis enrichment by linking to external biological databases. The goal is two-fold: 1) assess content of the data for the target use in repurposing; and 2) execute the prototype workflow and identify issues for improvement.
Methods

We obtained a subset of 2,982,326 Amazon reviews on health and personal care products. The reviews were made available courtesy of McAuley et al.7,8, who batch-fetched in previous work and shared the dataset with the research community. Our methods consisted of three stages, as elaborated below:

1. Text mining of reviews that mention an unexpected benefit: A java program was implemented to extract fields of interest (e.g., product ID, review text, review date) from the JSON file with ~2 million reviews and matched each review text with the following (case-insensitive) regular expression pattern:
   \b((unexpected|unintended|unanticipated|surprising)\s+(effects?|benefits?))\b

2. Manual curation/analysis of the mined reviews: The regex-matched reviews were manually curated by the first author (JF) to determine which ones indeed described a certain unexpected benefit of a dietary supplement (or called the “true positive” reviews). The true positives were further analyzed for characterizing the types of the supplements and the benefits. The ones that did not meet the true positive criterion were also examined and categorized.

3. Hypothesis enrichment for specific findings of interest: To test the feasibility of generating richer hypotheses by knowledge integration, we selected a true positive review where the supplement and unexpected benefit (syndrome relief) were both unambiguous and searched them in biological databases. For the syndrome, we manually identified the corresponding trait in the Genome-Wide Association Studies (GWAS) Catalog9 and its associated genes. As symptoms are also covered in the GWAS traits, this method is not limited to only diseases. For the supplement, the Comparative Toxicogenomics Database (CTD)10 was manually looked up to find its associated genes. In addition, functional similarity of the genes between the supplement and the syndrome was then assessed by using an information-theoretic measure11 to pinpoint possible biological explanation.

Results

Manual curation of the candidate reviews

Out of the 2,982,326 reviews, the regex pattern matched 169 candidate reviews. By reading through the candidate reviews, four categories were induced and summarized in Table 1. Less than half of the reviews (45.6%) contained unexpected benefits of interest (True positive) – with analysis elaborated in the next subsection. Among the true positives, there were five substance-effect associations supported by more than one review. For example, unexpected weight loss from using products with vitamin B2. A comparable portion (45.0%) of the reviews was about “non-dietary” health and personal care products (e.g., heart rate monitor and electric toothbrush), which were not filtered out in implementation of the screening program. The neutral effect category includes reviews with health-irrelevant property (e.g., enhanced flavor) or of debatable benefit (e.g., loss of appetite). The bottom minor category (Unspecified effect) is reviews with non-specific description such as “if you do a little research you may find some surprising benefits to xylitol”. Although infrequent, those irrelevant cases indicate the regex-based approach may still pick up some noise.

Table 1. Categories identified from manually annotating the regex-matched candidate reviews

<table>
<thead>
<tr>
<th>Category</th>
<th>Definition</th>
<th>Counts (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>True positive</td>
<td>Dietary supplement with unexpected benefit indicated in the review</td>
<td>77 (45.6%)</td>
</tr>
<tr>
<td>Non-dietary</td>
<td>Product is health-related by not for dietary use</td>
<td>76 (45.0%)</td>
</tr>
<tr>
<td>Neutral effect</td>
<td>The effect from the dietary supplement is not apparently beneficial</td>
<td>10 (6.0%)</td>
</tr>
<tr>
<td>Unspecified effect</td>
<td>Description of specific effect is missing or unclear</td>
<td>6 (3.4%)</td>
</tr>
</tbody>
</table>
Qualitative analysis of the unexpected benefits

By inspecting the true positive reviews and product information, we found that about a half of the consumer-claimed “unexpected” benefits might actually be expected. Given the subjective nature, it is not uncommon that the reviewers were surprised due to lack of domain knowledge or just did not pay attention to the product labels. Nonetheless, there remained a good amount of informative reviews, from which we listed some notable repurposing contexts in Table 2. The examples demonstrate diversity of the potential uses, ranging from relief of symptoms, mental function enhancement, modification of substance use behavior, to deterring of infection vectors.

Table 2. Examples of potential supplement repurposing based on the reviews

<table>
<thead>
<tr>
<th>Product ID</th>
<th>Product type</th>
<th>Review excerpt</th>
<th>Review year</th>
</tr>
</thead>
<tbody>
<tr>
<td>B00008CQTS</td>
<td>Recombinant human growth hormone (hGH)</td>
<td>…very noticeable increase in memory and cognitive skills</td>
<td>2006</td>
</tr>
<tr>
<td>B0000537A7</td>
<td>Hair regrowth tablets</td>
<td>…everyone around me was being bit and I wasn't bit once. The room I stayed in at night had a lot of mosquitos too. I killed over 20 of them in the room but I never suffered even one bite!</td>
<td>2011</td>
</tr>
<tr>
<td>B001DYKCJQ</td>
<td>Creatine monohydrate (for athlete performance enhancement)</td>
<td>…amazed how it controlled the acid reflux for me.</td>
<td>2012</td>
</tr>
<tr>
<td>B0058GXIYG</td>
<td>Ashwagandha extract capsules (marketed as a general healthy herb supplement)</td>
<td>I experienced an unexpected benefit by taking away the negative symptoms of IBS (with constipation) of which I have suffered for 33 years</td>
<td>2012</td>
</tr>
<tr>
<td>B0013OUKPC</td>
<td>Inositol powder (marketed for liver function support)</td>
<td>…a significant reduction in the number of heart palpitations (premature atrial contractions) I have.</td>
<td>2013</td>
</tr>
<tr>
<td>B002900RUS</td>
<td>Resveratrol (marketed as general healthy juice capsule)</td>
<td>A rather stubborn bout of eczema on my feet cleared up after being there for two years.</td>
<td>2013</td>
</tr>
<tr>
<td>B0087Q8GZA</td>
<td>Probiotics tablets (marketed as supplement for gut health)</td>
<td>…an unexpected benefit was the improvement in my sleep.</td>
<td>2013</td>
</tr>
<tr>
<td>B00EIW6NZC</td>
<td>L-arginine alpha-ketoglutarate (amino acid supplement)</td>
<td>A surprising effect is I quit smoking due to losing my desire for cigarettes since I started arginine.</td>
<td>2014</td>
</tr>
</tbody>
</table>

Hypothesis enrichment by linking to biological databases
From the true positive examples in Table 2, IBS-ashwagandha association (4th row under the heading) was utilized for the enrichment exercise, as the supplement and relief of the condition (IBS, irritable bowel syndrome) are both unambiguous according to the review. Ashwagandha (*Withania somnifera*) is a perennial shrub that has been used as medicinal herb in some cultures. To our knowledge, there is not any literature on the use of ashwagandha for IBS. We were able to identify their associated genes from the GWAS Catalog and CTD respectively (see Table 3). By further looking into the GO annotations, moderate functional similarity between the genes suggests possible biological underpinning of the observed effect. For example, Figure 1 illustrates a couple of the GO Biological Process terms between SOD1 (ashwagandha) and PER2 (IBS), with information-theoretic similarity scores computed using our previously published method. For GO Molecular Function terms, we also observed overlaps such as protein binding (GO:0005515) shared between PER2 (IBS) and HSPA9 (ashwagandha).

Table 3. Genes identified as associated with IBS and ashwagandha in the databases

<table>
<thead>
<tr>
<th>IBS</th>
<th>ashwagandha</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene ID: 166378 Symbol: SPATA5</td>
<td>Gene ID: 847 Symbol: CAT</td>
</tr>
<tr>
<td>Gene ID: 8864 Symbol: PER2</td>
<td>Gene ID: 3306 Symbol: HSPA2</td>
</tr>
<tr>
<td>Gene ID: 11014 Symbol: KDELR2</td>
<td>Gene ID: 4684 Symbol: NCAM1</td>
</tr>
<tr>
<td>Gene ID: 6647 Symbol: SOD1</td>
<td></td>
</tr>
</tbody>
</table>

Figure 1. Similar biological processes between two genes associated with ashwagandha (SOD1) and IBS (PER2)

Discussion

Consumer reviews as innovation driver

Our results suggest that mining massive consumer reviews may enable (timely) discovery of useful incidents of unexpected dietary benefits. Enhanced with an integrative approach, rich hypotheses can be generated to spark novel research. Below we discuss several noteworthy cases from Table 2:

1. In the enriched analysis of Figure 1, the possible involvement of circadian rhythm and ovarian follicle development offers a biological hint that aligns with evidence of IBS gender disparity as reported in the
We believe systematically linking consumer experience to formal databases will shed light on potentially fruitful pharmacogenomics studies.

2. The product review on improvement in cognitive skills after taking recombinant human growth hormone was written in 2006, while the earliest literature we found regarding such association was published in 2013\textsuperscript{15,16}. It is likely that screening from a large consumer population can help detect useful signals a few years ahead of other formal studies.

3. In 2013, a consumer reported eczema was resolved by dietary resveratrol, which does not seem to have been reported in literature. Interestingly, a US patent was filed in 2001 for external use of resveratrol for treating exfoliative eczema, acne, and psoriasis\textsuperscript{17}. Due to less stringent regulation, this example suggests that supplement-based intellectual property could be nimbly registered as soon as a promising effect is found.

4. In 2013, a consumer reported unexpected sleep improvement by using probiotics tablets. Until 2014 and 2016 respectively, such beneficial effect on mice and human began surfacing in the literature\textsuperscript{18,19}. Since many of the supplements are food-based, we foresee that mining the relevant consumer feedback would expedite hypothesis generation for the cutting-edge frontiers of microbiomics and broadly nutrigenomics.

### Potential model of citizen science

The concept of citizen science\textsuperscript{20,21} has been promoted in various domains. The basic idea is letting motivated lay people participate in research studies and perform tasks such as collecting samples or data. We believe that supplement repurposing is an ideal area to engage the general public into a crowd-sourced style of “observational trial”. The advantages of leveraging massive review data include: diverse product types, large consumer base, and constant monitoring as integrated into daily life. To facilitate this mission, informatics should be able to contribute on vocabulary harmonization and streamlining the integration of data sources. An interesting observation was how some consumers described in reviews that they tested the effect via self-controlled experiment, i.e., verifying the effect by stopping use and resuming again. This suggests that, with moderate scientific training, we could further improve the quality of data collected from general consumers.

### Limitations

As a proof-of-concept study, the review curation was conducted by only one annotator (JF) and without referring to a tested guideline. Due to the novelty in many of the findings, we could not find a proper reference standard for the evaluation. Some product reviews can be anecdotal/subjective by nature and of varying quality dependent on the consumer’s background. It is not uncommon to find inconsistent comments on the supplement-effect from other Internet sources (e.g., the relation between creatine monohydrate and acid reflux). Those inconsistencies could be accounted by justifiable genetic variation or could be just noise from uncontrolled covariates. Multi-ingredient supplements also pose challenges in pinpointing the exact active element that caused the effect. A relevant argument is that the sense of “repurposing” actually becomes indistinct for supplements that are marketed as “all-purpose”. In comparison to approved drugs, supplements also tend to conceal risks that are not well studied.

### Future work

With observing a considerable amount of unwanted non-dietary product reviews, we will try to include a filter based on the finer category labels available in the detailed product information. A more rigorous annotation process will be needed, especially with a formal guideline and execution by multiple annotators. For higher sensitivity in detecting the semantics of unexpected benefits, we will explore natural language processing (NLP) and machine learning techniques. To automate the discovery workflow, a pipeline will be developed by incorporating the NLP and function similarity programs based on our previous work. The throughput can be boosted by implementing batch-lookup with the NCBI E-utilities\textsuperscript{22} and directly processing downloaded GWAS Catalog and the CTD database. These enhancements should jointly contribute to a more accurate estimation of the yield (in terms of generating useful hypotheses) of the overall solution. Lastly, we still need to figure out a suitable reference standard for improving the validity and scalability of the evaluation.
Conclusion

To explore useful information for repurposing dietary supplements, we proposed a workflow of mining unexpected benefits mentioned in massive consumer reviews. The proof-of-concept study involved: 1) automatic matching of clue phrases to identify candidate reviews, 2) manual categorization/analysis of the candidate reviews, and 3) manually linking a novel supplement-disease pair to relevant biological databases for hypothesis enrichment. We found 45.6% of the 169 matched candidate reviews contained user-claimed unexpected benefits of dietary supplements, and about half of them were likely novel. By browsing relevant literature, some of the cases did demonstrate potential value for driving repurposing innovation. The hypothesis enrichment also derived informative functional associations between the supplement and the disease. More rigorous evaluation and validation approaches will be needed. The results show meaningful content in consumer reviews as well as the feasibility of a workflow to facilitate supplement repurposing.

Acknowledgement

We thank Colleen Kenost for assistance in editing the manuscript.

References

17. Pelliccia MT, Giannella A, Giannella J. Use of resveratrol for the treatment of exfoliative eczema, acne and


Improving Common Ground Development in Surgical Training through Talk and Action

Yuanyuan Feng, MS, and Helena M. Mentis, MS PHD
University of Maryland Baltimore County, Baltimore, MD, USA

Abstract

In surgical training, senior surgeons and residents rely on more than just verbal utterances to share information and coordinate their work practices – their actions also contribute to and shape the development of common ground. However, the function of actions in the grounding process and how that is interdependent with verbal utterances have not been made explicit. We have investigated actions and utterances using a dialogue act coding scheme that highlights the communicative functions of each act towards common ground development during a laparoscopic surgery. We show that utterances provide detailed information for surgeons to develop content common ground, whereas actions contribute to process common ground development. Thus, utterances and actions, by providing different forms of information, depend on each other to develop and maintain common ground. Based on this understanding, we discuss opportunities in transforming actions into perceivable knowledge on the laparoscopic display that supports effective communication and surgical training.

Introduction

The operating room is acknowledged to be a key learning environment where surgical trainees achieve surgical expertise by collaborating on surgical procedures with senior surgeons. This mentor-apprenticeship model requires timely and accurate information sharing between senior surgeons and surgical trainees – not only do they have to coordinate their actions, but the senior surgeon also has to assess the trainee’s level of knowledge of the case and determine how much direction he or she needs to perform the procedure. Any failures in communication lead to inevitable technical and judgment errors that have a great impact on patient safety. Recently, there has been interest in the design of tools to augment shared displays in supporting communication in the operating room. For instance, procedural updates, expert gaze, and gestures have been found to be critical sources of information that facilitates team information sharing and common ground development. To build tools that visualize such supplemental information, we need a better understanding of the mechanisms through which these sources of information are shared and used in common ground development.

Although there has been extensive effort in the investigation of the contributions of verbal communication made in the grounding process, we know little of how information is shared through physical interactions in collaborative tasks in a complex and information rich environment. In surgical training, senior surgeons and residents rely on more than just verbal utterances to share information and coordinate their practices – they use actions, such as gestures, jointly to explore the operative field, as well as illustrate how to proceed with a procedure. Moreover, actions have been shown to enhance the grounding of communication by replacing utterances – in distributed instructional collaboration tasks (where there is a local worker and a remote helper), workers used actions to provide evidence of comprehension of instructions. Thus, actions can provide necessary communicative and coordinative cues to minimize the collaborative effort in establishing common ground.

In this paper, we investigate the communicative functions of actions and explore how these functions change in the course of in situ laparoscopic training. We utilized a coding schema for identifying an action’s communicative function that has been shown to demonstrate the development of common ground in collaborative decision making tasks. Our analysis exposes not only how a senior surgeon and resident pair communicates through actions compared to utterances, but also how the communicative information provided by actions changes in response to the grounding process. Our findings illuminate opportunities to transform actions into perceivable knowledge that supports effective communication and surgical training.

Background

Communication is more efficient when people share greater amounts of common ground - mutual knowledge, beliefs, and assumptions. People from similar professional backgrounds may have already established some initial common ground. For example, surgeons with similar medical education and surgical training may have developed a shared idea on the general steps of a procedure and the basic motor skills in conducting these steps. In addition, a team constructs and expands their common ground within the on-going experience of the joint activity, as
well as adapts their understanding and acts based on the perceptual elements from the current context\textsuperscript{17, 18}. Thus, common ground is incrementally built on the previous joint actions between team members. As joint experience on a task accumulates, a team becomes more efficient in exchanging information and applying the knowledge gained to their tasks\textsuperscript{10, 19}.

Convertino et al. explained the distinction between content common ground and process common ground in the context of cooperative work, where the coordination of content depends on a shared understanding of the subject (know that), and the coordination of process depends on a shared understanding of the rules, procedures, timing, and manner in which the interaction will be conducted (know how)\textsuperscript{10}. Over the course of a joint task, content common ground is developed throughout a task, whereas process common ground is developed in the early stages and maintained for efficient work\textsuperscript{10, 11}.

Fulfilling communication goals, such as sharing information and repairing understanding, requires both spoken and unspoken communication, such as utterances, actions, gestures, gaze, and facial expressions\textsuperscript{20, 21}. Hand gestures uniquely communicate significant information in cooperative work, such as in assisting in the storing of information, expressing ideas, and mediating interaction in the communication between team members\textsuperscript{22}. Being able to see one’s actions has been shown to replace a worker’s utterances by serving as a tacit acknowledgement in grounding a given instruction\textsuperscript{15}. And a remote helper embodied through a pointing-drawing tool leads to representational gestures, such as directions of the actions and shape of the task objects, thus facilitating task communication and performance time\textsuperscript{23}.

Due to the evidence to date of the integral nature of actions in communication and collaboration, it is of interest to further articulate the relationship between actions and utterances as a function of common ground development and maintenance. In the following study, we investigated the communicative functions of actions and their relationship with verbal utterances within the grounding process. Capitalizing on common ground as an orienting theory, our research extends previous studies in identifying fine purposeful actions and understanding their functions in relation to utterances. With this knowledge, our motivation is to determine how to further support common ground development by facilitating gestural interaction of laparoscopic displays during in situ laparoscopic training.

**Study Design and Data Collection**

The study was based on videos recorded during fieldwork in the operating rooms of the Cambridge Hospital. Human Subject’s approval was granted through the Cambridge Health Alliance IRB. A total of ten laparoscopic cholecystectomy cases were analyzed. The cases included three different surgeons and three different PGY4 residents in different combinations of the two groups. The details of the cases have been published in our previous paper on conversation analysis\textsuperscript{24}.

**Coding Scheme and Procedure**

In this study, the communication content was examined in both dialogues and actions between the attending surgeons and residents. The utterances and actions were coded with the same dialogue act coding scheme, which was developed to understand the development of common ground among interdependent team members in emergency management tasks to provide implications for the design of collaborative systems\textsuperscript{12} (Table 1). This method emphasizes the use of communicative functions to interpret the development of common ground as opposed to linguistic or semantic meaning. We used this coding scheme to understand the functions of actions and utterances in the development of common ground in a complex, collocated task around a shared view. Using the same coding scheme facilitates the comparison of the two forms of communication and their patterns of use.

The first author and a research assistant coded the utterances and actions. For the verbal communication coding, the coders first viewed the videos and coded the transcripts independently. After the first independent coding session, the inter-rater reliability for the coders was found to be Kappa = 0.64. They negotiated for any conflicting codes and then coded again and achieved an agreement of Kappa = 0.87. This was deemed high agreement\textsuperscript{25} and so the remainder of the cases were coded. The inter-rater reliability was analyzed using the Kappa statistic for each case. For any case with Kappa less than 0.7, the two coders negotiated the difference to achieve higher agreement. In the end, the average Kappa score is 0.78, ranging from 0.71 to 0.87.

For the coding of actions with the same dialogue act coding scheme, the two coders were able to disambiguate the task actions from the communicative actions with guidance from the laparoscopic cholecystectomy task analysis\textsuperscript{26}, the laparoscopic surgeons from the Simulation to Advance Innovation and Learning Center at the Anne Arundel Medical Center, and the PI’s (last author) extensive knowledge and experience within laparoscopic surgery. The two
coders first viewed one case together to identify the actions surgeons made to collaborate in the surgery and discuss the functions behind every action. Then, the two coders coded the case separately and compared the codes. For the first round, the inter-rater reliability for the coders was found to be Kappa = 0.63. They viewed the video together and negotiated any conflicts in the code. The two coders coded the case again and the Kappa score was increased to 0.69. A large portion of the conflicts were due to the timing of the actions. The timing of an action was further clarified to be the start of each action. Then, the two coders coded the case separately for the third time and achieved a high inter-rater reliability, Kappa = 0.8. The remainder of the cases were then coded independently by the two coders.

Table 1. Dialogue act coding scheme for utterances and actions. Relationship of codes to content or process common ground are from Convertino et al.10.

<table>
<thead>
<tr>
<th>Class</th>
<th>Dialogue Act</th>
<th>Description</th>
<th>Relationship with Common Ground</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transfer Info</td>
<td>Add Info (AI)</td>
<td>Provides new information, not elicited.</td>
<td>Content Common Ground</td>
</tr>
<tr>
<td></td>
<td>Query (Q)</td>
<td>Question used to elicit new information.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Reply (R)</td>
<td>Reply to query to provide new information.</td>
<td></td>
</tr>
<tr>
<td>Check Understanding</td>
<td>Check (CH)</td>
<td>Verify own understanding of information previously presented by others.</td>
<td>Content Common Ground</td>
</tr>
<tr>
<td></td>
<td>Align (AL)</td>
<td>Verify partner's understanding of information previously presented to others.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Clarify (CL)</td>
<td>Clarifies or restates information already presented.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Acknowledge (AC)</td>
<td>Signals receipt of information, understanding.</td>
<td></td>
</tr>
<tr>
<td>Manage Process &amp; Decision</td>
<td>Manage (MN)</td>
<td>Instruction, command, direct or indirect request for action; orchestrating strategy, how to do the work.</td>
<td>Process Common Ground</td>
</tr>
<tr>
<td></td>
<td>Summarize (SA)</td>
<td>Summarizes information previously presented.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Judge (J)</td>
<td>Individual judgment, opinion, or preference.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Confirm (CO)</td>
<td>Requests partners' agreement on a proposed decision.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Agree (AG)</td>
<td>Indicates approval for a prior judgment or decision.</td>
<td></td>
</tr>
</tbody>
</table>

Phase Identification

As the grounding of communication is an ongoing process, we needed to look at the distribution of utterances and actions over the course of surgery. We identified the phases of laparoscopic cholecystectomy based on a hierarchical task analysis of the laparoscopic cholecystectomy procedure26, which includes surgical steps, sub-steps, tasks, sub-tasks and motions. In our study, we used the surgical steps to define the main phases of laparoscopic cholecystectomy: prepare patient, isolate gallbladder, remove gallbladder, and cleanup. The communicative events were grouped into the phases of surgery by matching the event’s occurrence to the phase time stamps.

Data Analysis

The data analysis consists of three parts – the frequency comparison of dialogue acts between utterances and actions, the sequential analysis for the two-chain communicative events, and the Poisson regression for the trend of communication in the grounding process.

In the frequency comparison, the frequencies of the dialogue acts were counted and normalized by the total counts in each case. The frequency of a communicative function was calculated as the sum of the frequencies of the dialogue acts in the corresponding category. Two-tailed Wilcoxon signed rank tests were conducted for the comparison of the frequencies of utterances and actions in each communicative functions and dialogue act.

In the sequential analysis, the two consecutive communication events with repeated codes across dialogue acts for utterances and actions were statistically identified27 and trimmed to meet the assumptions of the chi-square test28. Chi-square tests were conducted to infer any significant patterns among the communicative events. Since we only focused on the patterns with high frequency, a two-chain communicative event was deemed significant if its adjusted residual was greater than 2. The conditional probabilities were used to compare the frequencies between different significant patterns.

Poisson regression models with repeated measures were developed to identify the trends of communicative functions of utterances and actions over the four phases of the surgery. The counts of dialogue acts in frequency comparison were used with a different normalization – they were normalized based on the total counts in each phase, instead of each case. In this way, we can compare the frequencies between phases within each code. The frequency of communicative function was the combination of the dialogue acts that fall into its category.
Results

Of the ten laparoscopic cholecystectomy cases in this study, 657 minutes of operative time were assessed. In total, there were task-related 3207 utterances and 2414 actions taking place between attending surgeons and residents.

The Communicative Function of Utterances and Actions

To begin to uncover the communicative functions of actions compared to utterances in the grounding process, we first compared frequencies at the dialogue act level.

Information sharing through the direct providing of information (AI) or through a query-reply process (Q-R), was overwhelmingly verbal in nature. Utterances were dominantly used for adding information (AI) ($Z = 2.803, p = 0.005$), queries (Q) ($Z = 2.803, p = 0.005$), and replies (R) ($Z = 2.803, p = 0.005$) (Figure 1A). However, the effort the team spends in explaining the presented information to ensure all the team members reach the same understanding tended to engage actions throughout the collaborative task. We can see that surgeons preferred verbally verifying their own understanding (CH) ($Z = 2.805, p = 0.005$), but they engaged actions in clarifying the information (CL) ($Z = -2.497, p = 0.013$) or showing their acknowledgement (AC) ($Z = -2.803, p = 0.005$) (Figure 1B). For instance, when dissecting Calot’s triangle, the resident verified his understanding by saying, “so this is the junction, right?” (CH). The attending surgeon then verbally clarified the anatomy and pointed to the common bile duct with his instrument (CL). With this information, the resident inserted the instrument and dissected above the common bile duct (AC). And managing the process itself further engaged actions to a large extent above that of verbal utterances. Instructions (MN) were usually provided in both utterances and actions ($Z = 1.478, p = 0.139$), while judgment (J) ($Z = -2.191, p = 0.028$) and agreement (AG) ($Z = -2.191, p = 0.028$) were shown in actions (Figure 1C). For instance, when removing the gallbladder from the liver bed, the resident pulled out the instrument to show his judgment that the dissection was completed on the current side of the liver bed (J). The attending surgeon agreed with him by turning the gallbladder over (AG).

Thus, the comparison of frequencies in communicative functions between actions and utterances reflects that utterances are often used in explicitly presenting selective information, while actions are more frequently taken to negotiate how to proceed with the procedure. For example, in ligating the cystic duct, the attending surgeon verbally identified the duct (AI) and then the resident moved his instrument toward the structure to show his acknowledgement (AC). The attending surgeon then verbally directed the resident to the distal of the duct (MN). The resident pointed his instrument at the clipping location he selected based on the attending surgeon’s instruction (J) and waited for the attending surgeon’s agreement (CO). By holding the structure still, the attending surgeon tacitly approved the resident’s selection (AG) and the resident proceeded with the clipping action (AC).

According to Convertino et al.¹⁰, as content common ground is the selective sharing of information and process common ground is the management of the procedures and process, the observed distinction between actions and utterances indicates that the actions are predominantly a function of process common ground. On the contrary, the utterances tended to serve to share information that was required by the team and thus are a function of content common ground.

Sequential Frequency Analysis of Utterances and Actions

Frequency analysis provided a high-level understanding of the predominant function of utterances versus actions; however, utterances and actions are not siloed phenomenon. There is a clear relationship between these two mechanisms of coordination. We conducted a sequential frequency analysis in order to articulate the relationship between utterances and actions with regards to communicative function. Lag sequential analysis allows us to examine the immediate response, either by utterances or actions, after a given communicative event. We included
repeated codes in our analysis. Thus, in total, we received 5559 two-chain communicative events across the dialogue acts for utterances (22 codes) and actions (22 codes).

We first constructed a contingency table with all codes for both the given and target events. Due to the high dimensions, there were 1682 cells (91.0%) with an expected value less than 5. In order to meet the assumptions for chi square test, we did two steps for trimming the data. First, we identified and deleted the codes that led to a cell with an expected value less than 1. Next, we compared the total counts for the codes that had cells with expected values less than 5. The codes with smaller counts were further deleted. The trimmed data consisted of nine dialogue act codes for the given and target events, with 15 cells (18.5%) having expected value less than 5 and the minimum expected count of 1.63. The chi square test over this data set shows that the target events significantly depend on the given events \( \chi^2(64) = 5856.79, p < 0.0001 \). The adjusted residual of each cell allowed us to identify 15 significant two-chain communicative events with adjusted residuals greater than 2 among all others. In order to make sense of these pairs, we created a diagram with the relationships between the events.

As shown in Figure 2, the significantly occurring communicative pairs grouped into two clusters that are related to transferring information (on the left) and managing process (on the right). In the transferring information cluster, information was typically added by an attending surgeon through utterances (Att_AI_u) accompanied by an action (Att_AI_a) and then the resident verbally sharing information (Res_AI_u). This pattern was often used in identifying anatomy. For instance, the attending surgeon would first verbally describe the structure (Att_AI_u), and then move the tissue (Att_AI_a), grasping and exposing embedded structures. Through these actions coupled with the utterances, the resident perceives the described anatomy. Although actions can present explicit information to the residents, they were not typically decoupled from utterances. Furthermore, we observed that the actions for adding information were taken depending on the role of the communicator. The residents seldom used actions to describe or reference to any structure, however, the attending surgeons did in abundance.

In the managing process cluster, attending surgeons either verbally instructed the residents or physically guided the process. This is evidenced by the two loops observed in the cluster – the action loop on the bottom right and the utterance loop on the top right of the diagram in Figure 2. In the action loop, the attending surgeon used actions to request the resident to perform a movement (Att_MN_a), which was always indicated as understood by the residents with the action of acknowledgement (Res_AC_a). After fulfilling the requested task, the residents presented their own judgments through actions (Res_J_a), which were recognized and agreed on by the attending surgeons (Att_AG_a). This loop was often observed when the residents were dissecting the anatomy. For instance, the attending surgeon grasped the tissue around the cystic artery (Att_MN_a) and the resident inserted the instrument dissecting it (Res_AC_a). When the task was done, the resident stopped dissecting and pulled back the instrument (Res_J_a). The attending surgeon would then agree with the outcome by releasing the tissue and grasping another part of the anatomy for the next step of dissection (Att_AG_a).

In the utterance loop, the attending surgeons often would verbally inform the residents of the next step of the process (Att_MN_u). Sometimes, the verbal instructions were accompanied with an attending surgeon’s clarification action (Att_CL_a) when a precise movement was required. This loop was often observed when the team would move from one task to the next. For example, the attending surgeon informed the resident to dissect the back of the gallbladder by saying, “Alright. Now let’s see what’s on the back” (Att_MN_u), before he turned the gallbladder over (Att_CL_a). With the instruction, the resident stopped dissecting and pulled back his instrument to leave enough

Figure 2. Significant communicative patterns among ten cases – information sharing cluster on the left and process management cluster on the right. (Att: attending surgeons, Res: residents, u: utterances, a: actions. The frequency of each pair is shown as numbers in the diagram.)
space for the turning action (Res_AC_a). The clarification actions happened when a movement should be exactly executed. For example, the attending surgeon first verbally requested the resident to pull the gallbladder to the upper left (Att_MN_u), and then he identified the most appropriate position for the gallbladder. So the attending surgeon would point to the anatomy with his instrument (Att_CL_a) showing the resident where to grasp and pull.

We are intrigued to observe a link between these two clusters – the residents’ acknowledgement and the attending surgeons’ adding of information. This link indicates that an agreement in process management, i.e. process common ground, leads to an update of the task content information sharing, i.e. content common ground. Thus, although we learned from the prior analysis that predominately utterances are a function of content common ground and actions are a function of process common ground, from this sequential analysis we understand that the actions taken to manage process common ground are intimately tied back to the successive task’s need for continual management of the content common ground.

Changes in Utterances and Actions Throughout the Grounding Process

Our third set of analyses were to uncover how the grounding process changes the communicative needs for utterances and actions. As we know that common ground increases as a team collaborates over a task over time, we present the comparison of frequencies in utterances and actions over the course of surgery.

As shown in Figure 3, utterances and actions followed different patterns across the entire surgery. Phase 1 begins with the first incision followed by the surgeons putting the laparoscope into the abdominal cavity. This was the first time surgeons saw the anatomical structures that they would operate on. They needed to describe their understanding of the structures to their partner. They continued to verbally add information regarding the context until they achieved a mutual understanding of the operative field. This continuous stream of adding information increased in Phase 2 with the delicate task of dissecting and exposing the anatomy to be clipped and cut, but then we see a decrease in the verbal adding of information in Phase 3. At this point, the content common ground had been established and surgeons shifted their attention to negotiate how to perform the tasks and to develop a mutual understanding of the process. After taking out the gallbladder, the surgeons moved onto Phase 4, which required surgeons to inspect both the operative field and the surrounding area and cleanup any remaining fluids and address any residual bleeding. The surgeons needed to verbally update their understanding of anatomy in order to continue to coordinate on the procedure itself. Thus, the trend for information transfer increased again and process management decreased.

Actions displayed a different pattern than utterances when looked at across the different phases of the surgery. Compared with utterances, process management and checking understanding through actions had a high frequency in Phase 1, when the initial common ground of the surgical practice had already been developed before the work, and decreased in Phase 2, when the case based information was transferred in abundance. Then, as the amount of new information that needed to be understood and shared decreased during Phase 3, more actions were used by the surgeons to manage the process and ensure that they are on the same page to proceed. As the surgeons moved onto Phase 4, which involved new tasks, the actions for managing process decreased, while the actions for checking understanding continued to grow. This shows actions were a fundamental aspect of coordinating the process and checking understanding after they verbally transferred the information.

To assess the significance of these trends, we fit the Poisson regression model with repeated measures using normalized counts as the dependent variable and phases, communicative functions and their interactions as independent variables. For utterances, the frequency of verbally transferring information is significantly higher in Phase 1 (\( \beta = 0.618, p = 0.006 \)) and lower in Phase 3 (\( \beta = -0.417, p = 0.012 \)), compared to managing process. For actions, the frequencies of actions in information transfer is significantly greater in Phase 2, compared to checking information (\( \beta = 1.433, p < 0.0001 \)) and managing process (\( \beta = 1.392, p = 0.001 \)).

![Figure 3](image_url). Frequencies of transferring information, checking understanding and managing process over the phases.
We then divided the two-chain communicative event data based on the identification of the four surgical phases. For each phase, we trimmed the data to meet all of the assumptions for a chi-squared test. With the trimmed datasets, the dependency is significant between the given and target event for all four phases ($p < 0.001$). We used adjusted residual ($>2$) to identify the significant pairs. Figure 4 shows the change of communication patterns over the course of surgery. Each surgery started with attending surgeons adding information, which led to a series of acts in process management (Figure 4A). When a team started to collaborate, their common ground was relatively low compared to the rest of the phases. A team first needed to quickly build up enough mutual understanding of the task and anatomical structure that was specific to the patient before them. Thus, in Phase 1, collaboration was driven by verbally adding information from the attending surgeons (Att_AI_u), which has been shown to be the most efficient mechanism for transferring information.

Figure 4. Two-chain communicative events for each phase of the surgery. (Att: attending surgeon, Res: resident, u: utterances, a: actions. Frequency of each pair is shown as the numbers in the diagram.)

As a surgery proceeds to Phase 2, information transfer and process management become two separate efforts without any links between them. Information was added by both the residents and the attending surgeons via utterances or actions, while process management was driven by the verbal instructions from the attending surgeons (Figure 4B). As the content common ground accumulates, the teams had adequate background information for them to develop strategies in performing the task. For example, the team had already identified the location and position of the gallbladder, the liver and the duodenum. Now their task was to expose and dissect the gallbladder and its accessory structures. Thus, they did not need to wait for more information in order to negotiate where and how to dissect. Meanwhile, the team kept adding information and updated their understanding of the content to make sure they were working on the same page.

In Phase 3, as shown in Figure 4C, most arrows direct from process management to information transfer, which indicates new information was put forward when there was a need identified in managing the process. The link between the attending surgeons’ verbal addition of information (Att_AI_u) and the residents’ physical demonstration of their own judgment (Res_J_a) further explains the usage of the information – to aid residents’ decision making in planning and performing the surgery.

Finally, in Phase 4 (Figure 4D), information transfer and process management became separate clusters again. Since Phase 4 contains a different set of surgical skills and visual interpretation – cleanup and suture – we observed that
the communicative patterns follow that of Phase 2: the surgeons verbally added information to develop the content common ground, while physically exchanging information for the process common ground.

Discussion

In this paper, we presented a series of analyses investigating the communicative functions of actions between an attending surgeon and a resident. Observing the collaboration between attending surgeons and residents in real operations allows us to identify and explain different ways actions are used in communicating information, as well as helps us to understand how the communicative functions change as common ground grows. This in turn provides guidance for gestural interaction of laparoscopic displays during in situ laparoscopic training.

We found that utterances and actions provide different communicative functions in discourse. Utterances were frequently used to explicitly add information. On the other hand, actions were frequently used to indicate one’s understanding. The actions provide two types of information – the task field and the physical activity. When the residents directed the instrument to arrive at the target location as indicated by the attending surgeons, they showed that they understood the instructions without spending additional effort in articulating their understanding verbally. This reflects the assumptions made by residents that the attending surgeons would directly observe the actions, situate the actions in the task field, and interpret the actions according to the instructions. In other words, the physical activity is first recognized and the task field is the context that provides meaning to the physical activity.

We also see in our third set of analyses that the communicative function of utterances and actions complement each other as the procedure progresses – actions are predominantly a function of process common ground and utterances are a function of content common ground. And, as demonstrated in Figure 4, the content common ground and process common ground are interdependent. Thus, the development of content and process common ground was bound together for surgeons to efficiently coordinate the procedure. Correspondingly, the utterances and actions were used interdependently in the grounding process.

The understanding of the communicative functions, rather than the semantic form, of actions and utterances and the way they jointly contribute to the communication in the process of grounding are crucial for designing systems to support collaboration on physical tasks via a shared view. Our findings show that utterances and actions work interdependently with each other, in the way that utterances are used as a function of content common ground and actions are used as a function of process common ground. This interdependency highlights the needs to not only see actions in the course of collaborative work, but also understand the meanings behind the actions in the shared view.

Seeing the Actions

What we have seen is that actions tacitly convey information to the collaborators. We found that the attending surgeons used actions to not only inform the residents of which task to perform, but also to control how the task was accomplished. Thus, markings and gestures on the shared display can show which target should be acted on, what motor skills should be taken, in which directions the actions should be conducted, and to what the end point in the procedure. In this case, providing explicit marks, such as arrows, measures, and tools, is useful in making the information perceivable and persistent along with pointing or drawing gestures. For example, to perform a suture, a surgeon can use the pointing or drawing gestures to explicitly locate an exact target to which the worker should attend. The surgeon can also add an arrow and adjust its direction to illustrate which direction the first stitch should be.

Providing such tools is also in agreement with prior work that has shown that mediated communication is more effective with annotation and gesturing tools. For instance, when provided with drawing/pointing tools, we see that gestures and annotations lead to shorter, more efficient communication and quicker task performance, pointing is preferred over speaking, verbal interactions often include gestures such as circles and arrows, and often implicit gesturing is preferred over explicit drawing. Thus, as we move towards building new systems for telestration over surgical video, we are inspired to focus our attention on such gestures and drawing that facilitate process common ground in order to further explore whether preferences for drawing tools is associated with more efficient communication around process management.

Knowledge in the Actions

For most surgical training, the purpose is not merely completing the task at hand. Rather, the aim is to help the residents to develop adequate knowledge, so that they can be able to work independently in the future. Thus, understanding why the actions should be taken is a necessary part of the collaboration. The knowledge behind the actions is embedded in the development of the content common ground. As shown in our study, the content common

703
ground is achieved mostly through verbal communication, which serves for adding new information. Compared to utterances, actions have their limitations in presenting a piece of information. On the other hand, we found both actions and utterances could convey the instructions to coordinate the process. Thus, design for team collaboration should avoid the use of gestures to explain the content, while encourage the use of gestures to guide the process to release the grounding effort in verbally communicating the meaning behind the actions.

In this case, designers should consider the temporality feature of the grounding process. Our study suggests that content common ground is first developed, followed by process common ground. Thus, at the beginning of a task, the effort should focus on verbal communication for a team to efficiently achieve content common ground. As the task moves on, more effort is expected to be on process common ground, wherein the gestures can be most useful.

Limitations

In our study, we investigated the spoken and unspoken communication between an attending surgeon and a resident during in-situ training of a laparoscopic cholecystectomy. The strength of these cases is that most surgical actions taken by surgeons are shown in the laparoscopic video. Thus, it provided a rich collection of communicative actions in a shared view for collocated settings. However, we acknowledge that knowledge from the entire surgical team, including nurses, anesthetists, and technicians, is of importance for the success of the surgery. Thus, our findings need to be examined for the design of groupware for the entire surgical team. Our study is also limited to laparoscopic cholecystectomy cases. Although laparoscopic cholecystectomy is an essential and fundamental procedure in the laparoscopic training for general surgery residents, our results need to be further validated to be applicable on a larger scale, especially for urgent cases or intraoperative complications. In addition, the surgeons in our study have completed professional training in general surgery. Before collaborating on a case, they shared a mutual understanding of the guidelines for surgery. Thus, we observed a quick assessment of ability at the beginning of the surgery. However, in cases, where team members come from different backgrounds, it may take a longer time for the team to develop a shared understanding.

Conclusion

In this study, we investigate the communicative ability of action in common ground development between a resident and an attending surgeon in laparoscopic surgical training. We found that utterances provide detailed information for surgeons to develop content common ground, whereas actions contribute to process common ground development. In this, we showed utterances and actions depend on each other in the development and maintenance of common ground. Based on our findings, we provided design guidelines supporting the seeing of actions on a shared display and the educational needs of understanding the meaning behind the actions.

Acknowledgements

The authors would like to acknowledge Christopher K. Wong and Veeha Khanna for their supports in coding the data and Amine Chellali for his support in data collection. This work was sponsored by NSF Grant IIS – 1422671.

References

Routine self-tracking of health: reasons, facilitating factors, and the potential impact on health management practices

Mayara Figueiredo, MS1,†, Clara Caldeira, BS1,†, Yunan Chen, PhD1, Kai Zheng, PhD1
1University of California, Irvine, Irvine, Irvine, CA

Abstract

Despite a growing interest in self-tracking of one’s health, what factors lead to self-tracking routinely (i.e., collecting data at regular intervals), and the effects of this behavior, remain largely understudied. Using data from the Pew Survey on Tracking for Health, we examined the patterns of self-tracking activity to understand reasons for this behavior and its impact on health management practices. We tested multiple logistic regression models to assess the influence of different predicting variables, and to find whether routine self-tracking leads to positive change to one’s approaches to health management. Our results suggest that recent visits to emergency care and the type(s) of tracking tools used are significant predictors of routine self-tracking activities. Further, the results suggest that routine self-tracking, as opposed to occasional, event-triggered tracking, is more likely to result in positive changes to health management approaches. Our findings also highlight barriers to and opportunities for designing useful and usable tools to facilitate self-tracking and empower patients to become more proactive in managing their own health.

Introduction

According to a 2012 Pew Research Center survey1, 69% of U.S. adults use tracking to manage their health or that of a loved one. Self-tracking is defined as the practice of repeatedly recording information such as behaviors, thoughts, and feelings about oneself. It encompasses collecting data and reflecting on it in order to acquire knowledge or achieve a goal, such as behavior change5. Self-tracking health information has been practiced for several decades using methods such as pen and paper, or just memory12. This practice has recently gained increased attention with the popularity of mobile technology and its ability to facilitate recording health information5. Self-tracking with the use of modern technology, such as smartphones and wearable devices, has been the subject of several recent studies8–11. The practice of tracking and the knowledge that the data tracked may provide can bring awareness and support health behaviors, thus helping in improving quality of life4. It can also improve patient engagement in their own care, since it allows patients to play a more active role in their disease management by better understanding their health conditions and coping with treatment and communicating with providers5,6. Therefore, self-tracking is also expected to improve patient-provider communication7.

Many self-tracking studies focus on tracking diet or exercise. These studies approach self-tracking either for preventative health management, for personal curiosity, or for learning and promoting a healthier lifestyle. For example, Miller and Mynatt13 developed StepStream, a school-based pervasive social fitness system, to encourage adolescents to improve their attitudes and perceptions towards physical activities. In another study, Cordeiro et al.14 analyzed the challenges people face when tracking food intake in order to explore opportunities to improve the support of diet tracking activities.

Other studies have approached self-tracking by focusing on its ability to assist in chronic disease management, since chronic illnesses usually require long-term treatment and management activities. The collection and reflection on the collected data can help people suffering from these conditions in their disease management by improving their knowledge about the illness, identifying triggers, and controlling health indicators such as glucose levels or blood pressure. Studies focusing on these aspects generally aim to control or mitigate symptoms and to prevent or delay disease progression. For example, Mamykina et al.15 developed and performed a deployment study of MAHI (Mobile Access to Health Information), a distributed mobile application to assist newly diagnosed diabetes patients in learning about their condition. Other similar studies focused on cancer16,17, heart conditions8,18, irritable bowel syndrome19,20, asthma21,22, and many others.

Collecting, registering, and reflecting on the data are the most fundamental steps of self-tracking, and the frequency with which data is captured and recorded can vary. Approximately half of U.S. adults who self-track record their data

†The first two authors wish to be regarded as joint First Authors.
routinely (e.g., every day or every two weeks), and half only track when they experience changes in their health. Past work has found that routine tracking is important for reflection, which is a necessary step towards common goals of tracking (e.g., behavior change), and included recommendations about how often to self-track. Li et al. describe that reflection can happen in short and long term. The former happens in a short time after collection, and helps to make users aware of their current status. The latter happens days or weeks after collection and involves deeper self-reflection, allowing users to identify patterns by comparing information from different periods. Through short term reflection, routine tracking can improve self-knowledge by providing more opportunities to reflect after collecting data. Routinely tracking is also important to long-term reflection, since missing data can limit users’ ability to interpret and reflect on their data. Although routinely registering personal data could influence self-tracking practices, little is known about what leads users to self-track routinely, and what are the outcomes of doing so.

In this study, we examined predictors of routine self-tracking behavior and its effects on health management based on a reanalysis of the data collected by the 2012 Pew Survey on Tracking for Health. We conducted a quantitative analysis of the data using binomial logistic regressions to understand what leads to routine self-tracking behavior and how this behavior subsequently influences health management practices. The two main research questions that we aimed to address through this study are:

**Research Question 1:** What factors predict routine self-tracking behavior?

**Research Question 2:** Is routine self-tracking associated with improved health management practices?

The rest of the paper is organized as follows. First, we describe the methods used in the research. Then, we present the findings followed by the discussion. Finally we present the limitations of the study and conclusions.

**Methods**

The analyses reported in this paper were based on the Pew Survey on Tracking for Health conducted between August 7 and September 6, 2012. The survey was administered through phone interviews in both English and Spanish, and involved a total of 3,014 adults living in the United States. Among them, 55.6% were female. The median age was 53.

**Data**

Each survey began with questions about demographic information (e.g., age, income) and health status (e.g., health rating, chronic conditions, recent changes). Then, the survey asked questions specifically about self-tracking, such as whether the respondent tracked fitness-related information (weight, diet, or exercise) or other health indicators (e.g., blood pressure, glucose). The survey then asked the method(s) that the respondent used to do self-tracking, as well as its effects, such as whether self-tracking affected their approach to self-management, or whether it led them to ask questions to physicians. The full instrument of the survey is provided on the Pew Research Center website.

In this study, we included all survey questions pertinent to our research objectives. They are listed in Table 1. Because we are interested in the characteristics of the respondents who used self-tracking to manage their own health, in our analysis, we included only those who answered “yes” to one of the following two questions: 1) whether they tracked their own weight, diet, or exercise routine (Q24); and 2) whether they tracked other health indicators such as blood pressure, blood sugar, sleep patterns, and headaches (Q25). Participants who tracked health data for others (e.g., family members) but not for themselves were excluded from our sample. Those participants who tracked health data both for themselves and for others were included.

Our two research questions focus on Q27, which asks the respondents, with respect to “the health indicator you pay the MOST attention to,” whether they self-track this indicator routinely, or if they collect data only when something comes up or changes (see Table 1). In this study, routine self-tracking is characterized by measuring and recording data continuously at regular intervals, such as daily or weekly. Those who do not track routinely might only record events when they experience a change in their health, or adopt a new habit. We refer to the first type of behavior as “routine self-tracking,” and the second type as “occasional, event-triggered self-tracking.”

**Data analysis**

To answer the first research question, what factors predict routine self-tracking behavior, we analyzed the responses to questions Q3 (health problems or conditions the respondent lives with), Q4 (major health events in the past 12 months,
Table 1: Survey questions used in the analysis.

<table>
<thead>
<tr>
<th>Question</th>
<th>Options</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q2 In general, how would you rate your own health - excellent, good, only fair, or poor?</td>
<td>1) Excellent 2) Good 3) Fair 4) Poor</td>
</tr>
<tr>
<td>Q3 Are you now living with any of the following health problems or conditions?</td>
<td>1) Diabetes or sugar diabetes 2) High blood pressure 3) Asthma, bronchitis, emphysema, or other lung conditions 4) Heart disease, heart failure or heart attack 5) Cancer 6) Any other chronic health problem or condition I haven’t already mentioned</td>
</tr>
<tr>
<td>Q4 In the last 12 months, have you personally...</td>
<td>1) Faced a serious medical emergency or crisis? 2) Gone to the emergency room or been hospitalized unexpectedly? 3) Experienced any significant change in your physical health, such as gaining or losing a lot of weight, becoming pregnant, or quitting smoking?</td>
</tr>
<tr>
<td>Q24 Now thinking about your health overall... Do you currently keep track of your own weight, diet, or exercise routine, or is this not something you currently do?</td>
<td>1) Yes, keep track 2) No, not something R currently does</td>
</tr>
<tr>
<td>Q25 How about any other health indicators or symptoms? Do you happen to track your own blood pressure, blood sugar, sleep patterns, headaches, or any other indicator?</td>
<td>1) Yes 2) No</td>
</tr>
<tr>
<td>Q26 Thinking about the health indicator you pay the MOST attention to, either for yourself or someone else, how do you keep track of changes? Do you use...</td>
<td>1) Paper, like a notebook or journal 2) A computer program, like a spreadsheet 3) A website or other online tool 4) An app or other tool on your phone or mobile device 5) A medical device, like a glucose meter 6) Or do you keep track just in your head? 7) Other</td>
</tr>
<tr>
<td>Q27 How often do you update your records or notes about this health indicator? Do you do this on a regular basis, or only when something comes up or changes?</td>
<td>1) Regular basis 2) Only when something comes up or changes</td>
</tr>
<tr>
<td>Q28 Do you share these health tracking records or notes with anyone, either online or offline?</td>
<td>1) Yes 2) No</td>
</tr>
<tr>
<td>Q30 In which of the following ways, if any, has tracking this health indicator affected your own health care routine or the way you care for someone else?</td>
<td>1) Has it affected a decision about how to treat an illness or condition? 2) Has it changed your overall approach to maintaining your health or the health of someone you help take care of? 3) Has it led you to ask a doctor new questions, or to get a second opinion from another doctor?</td>
</tr>
</tbody>
</table>
such as an ER visit), Q26 (tools used to track the health indicator that the respondent pays the MOST attention to, such as a notebook or a computer program), and Q27 (routine tracking or occasional, event-triggered tracking). We hypothesize that whether an individual routinely tracks the health indicator that she or he pays the most attention to (Q27) is a function of the person’s health conditions (Q3); major recent health events (Q4); and the tool(s) used (Q26).

To answer the second research question, is routine self-tracking associated with improved health management practices, we analyzed whether the tracking behavior—routine or occasional, event-triggered (Q27)—may predict responses to Q30. Possible responses to Q30 include: (1) whether self-tracking affected the respondent’s decision(s) about how to treat an illness or condition; (2) whether it changed the respondent’s overall approach to health management; and (3) whether it led to asking clinicians new questions, or to looking for second opinions. We hypothesize that routine tracking, as opposed to occasional, event-triggered tracking, will more likely lead to a positive response to one or more of the Q30 options. We were not able to analyze other aspects of health management, such as those that involve family members and caregivers, as the survey did not cover them.

In both models, we controlled for age, sex, socioeconomic indicators (education, ethnicity, and income), and overall perception of one’s health (Q2: excellent, good, fair, or poor). In the second model, we also included the independent variables used in the first model (e.g., Q3 health conditions, and Q4 major recent health events) as control variables.

Findings

Descriptive statistics

Among the 3,014 respondents surveyed, 1,941 (64%) reported that they performed some form of health-related self-tracking activities. These 1,941 respondents are hereafter referred to as “study population.” Table 2 describes the study population (N=1,941) dichotomized into two groups based on their self-tracking style (routine vs. event-triggered). Out of the 1,941 respondents in the study population, 992 were routine self-trackers; and 949 only tracked occasionally when triggered by events.

The mean age of the respondents in the routine group was 56 years; and for those in the event-triggered tracking group it was 50.7 years. Reported health rating was not significantly different between the two groups, but a higher proportion of the respondents who tracked their health routinely reported that their health was fair (16.4%) or poor (5.1%), and a smaller proportion as excellent (27.1%), compared to the event-triggered group (14.2%, 3.3%, and 31%, respectively). This result suggests that those respondents who perceived their health to be worse were more likely to engage in routine self-tracking activities, but the differences are not statistically significant.

Table 2: Descriptive statistics for routine and event-based tracking.

<table>
<thead>
<tr>
<th></th>
<th>Routine</th>
<th>Event-based</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>992</td>
<td>949</td>
<td>1941</td>
</tr>
<tr>
<td>Age*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18-29</td>
<td>17.9%</td>
<td>24.7%</td>
<td>21.4%</td>
</tr>
<tr>
<td>30-49</td>
<td>30.2%</td>
<td>36.8%</td>
<td>33.5%</td>
</tr>
<tr>
<td>50-64</td>
<td>29.1%</td>
<td>21.6%</td>
<td>25.3%</td>
</tr>
<tr>
<td>65+</td>
<td>22.9%</td>
<td>17.0%</td>
<td>19.8%</td>
</tr>
<tr>
<td>Health rating</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>excellent</td>
<td>27.1%</td>
<td>31.0%</td>
<td>29.1%</td>
</tr>
<tr>
<td>good</td>
<td>51.3%</td>
<td>51.2%</td>
<td>51.3%</td>
</tr>
<tr>
<td>fair</td>
<td>16.4%</td>
<td>14.2%</td>
<td>15.3%</td>
</tr>
<tr>
<td>poor</td>
<td>5.1%</td>
<td>3.3%</td>
<td>4.2%</td>
</tr>
<tr>
<td>Chronic illness</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>diabetes*</td>
<td>18.5%</td>
<td>9.3%</td>
<td>13.8%</td>
</tr>
<tr>
<td>high blood pressure*</td>
<td>33.5%</td>
<td>25.1%</td>
<td>29.2%</td>
</tr>
<tr>
<td>lung diseases</td>
<td>14.1%</td>
<td>14.4%</td>
<td>14.3%</td>
</tr>
<tr>
<td>heart diseases</td>
<td>9.9%</td>
<td>8.0%</td>
<td>9.9%</td>
</tr>
<tr>
<td>cancer</td>
<td>3.7%</td>
<td>3.4%</td>
<td>3.6%</td>
</tr>
<tr>
<td>Events (12 months)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>emergency or crisis*</td>
<td>14.2%</td>
<td>10.5%</td>
<td>12.3%</td>
</tr>
<tr>
<td>ER visit*</td>
<td>22.0%</td>
<td>16.6%</td>
<td>19.2%</td>
</tr>
<tr>
<td>health change</td>
<td>21.2%</td>
<td>20.2%</td>
<td>20.7%</td>
</tr>
<tr>
<td>Type of tracking</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fitness</td>
<td>91.3%</td>
<td>89.3%</td>
<td>90.3%</td>
</tr>
<tr>
<td>Other indicators*</td>
<td>57.0%</td>
<td>41.8%</td>
<td>49.2%</td>
</tr>
<tr>
<td>Tool used</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>paper*</td>
<td>42.0%</td>
<td>30.3%</td>
<td>36.0%</td>
</tr>
<tr>
<td>program*</td>
<td>7.2%</td>
<td>3.5%</td>
<td>5.3%</td>
</tr>
<tr>
<td>website*</td>
<td>2.1%</td>
<td>0.9%</td>
<td>1.5%</td>
</tr>
<tr>
<td>app*</td>
<td>11.9%</td>
<td>5.0%</td>
<td>8.4%</td>
</tr>
<tr>
<td>medical device*</td>
<td>12.3%</td>
<td>4.6%</td>
<td>8.4%</td>
</tr>
<tr>
<td>memory*</td>
<td>37.6%</td>
<td>63.7%</td>
<td>51.0%</td>
</tr>
<tr>
<td>sharing*</td>
<td>37.7%</td>
<td>32.0%</td>
<td>34.8%</td>
</tr>
<tr>
<td>Health management</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>changed decision*</td>
<td>38.4%</td>
<td>30.3%</td>
<td>34.2%</td>
</tr>
<tr>
<td>changed approach*</td>
<td>53.3%</td>
<td>41.3%</td>
<td>47.2%</td>
</tr>
<tr>
<td>asked new questions</td>
<td>41.0%</td>
<td>41.6%</td>
<td>41.3%</td>
</tr>
</tbody>
</table>

*p<0.05
Most respondents with diabetes or high blood pressure self-tracked routinely \((p < 0.001)\). The difference was not significant for those with lung diseases (e.g., asthma, bronchitis, emphysema), heart disease (including heart failure and heart attack), and cancer. This difference may arise from health providers’ recommendations for patients with these conditions to use self-tracking. Both the reduced effort required for tracking using home devices, such as glucose meters, and the recommendation given by health providers may encourage routine self-tracking practices.

Among those who experienced severe health events such as ER visits in the last 12 months, a larger proportion self-tracked routinely \((p < 0.05)\). For other health changes, such as becoming pregnant, losing or gaining a lot of weight, or quitting smoking, the difference was not significant. Experiencing a health crisis, or seeking health care from an emergency department, constitute more severe health events in comparison with other health changes, thus severe health events might be more impactful for self-tracking practices.

While there was no significant difference between the event-based tracking and the routine tracking groups regarding fitness-related indicators and activities (Q24), a significantly larger proportion of those who track other health indicators (e.g., blood pressure, blood sugar, sleep patterns, and headaches - Q25) do so routinely \((p < 0.001)\). Most respondents who shared data collected through self-tracking also tracked routinely \((p < 0.01)\). In addition, the two groups differ significantly in their use of self-tracking tools. Most respondents who used pen and paper, computer programs, mobile apps, medical devices (e.g., blood pressure cuff) \((p < 0.001)\) or websites \((p < 0.05)\) self-tracked routinely. On the other hand, a larger proportion of those who only kept data in their memory \(i.e.,\) measuring but not registering the data) only self-tracked after a health related event \((p < 0.001)\). It is possible that those who are more diligent about tracking are more likely to track routinely, register, and share their data. But registering and sharing data could also influence users to self-track routinely.

The number of respondents who claimed that tracking activities had affected any decision about a treatment or changed their approach concerning health management was significantly higher among the group who performs tracking routinely \((p < 0.001)\). The third option, which asked whether tracking activities led respondents to ask new questions to health providers or to look for a second opinion, resulted in no significant difference between the two groups.

In summary, routine tracking was more common amongst those who tracked indicators not related to fitness, those who shared their data, and recorded their data on paper or on a digital system. Most respondents who reported that tracking had changed a decision about health, or changed their self-management approach, self-tracked routinely. Because these variables may correlate amongst one another, we used binomial logistic regressions to learn which were significant predictors of routine tracking practices, and to find whether routine tracking influenced reported effects of self-tracking on health management.

**Model testing results**

In this section, we present the results from the logistic regression models, which are reported in Table 3 and Table 4.

**Research Question 1: What factors predict routine self-tracking behavior?**

As shown in Table 3, the model results indicate that age \((p < 0.01)\), recent visit to the ER \((p < 0.05)\), and use of a majority of self-tracking tools are associated with a higher likelihood of adopting the routine tracking behavior. When the respondent relied on memory to do self-tracking, the chance that she or he would collect data routinely is significantly decreased \((p < 0.05)\). Gender, chronic conditions, and recent health crises or changes do not appear to have a significant effect.

These results indicate that, regardless of other conditions, the probability of routine self-tracking is higher for those who had recent ER visits and those who use technological tools. It is possible that people become more vigilant with their health after having a severe health event warranting a visit to the ER, which leads them to become more diligent in their tracking behavior; and that the use of technological tools such as websites and tracking apps facilitates this practice.

**Research Question 2: Is routine self-tracking associated with improved health management practices?**

Through three binominal logistic models, we tested whether the routine self-tracking behavior leads to better health management practices. The results are reported in Table 4. Because none of the control variables have a significant
effect, they are omitted from the table.

As shown in Table 4, as opposed to occasional, event-triggered tracking, those who self-track their health routinely have a significantly higher likelihood of changing their overall approach to maintaining health ($p < .01$). Further, routine tracking is negatively associated with asking new questions from doctors, or seeking second opinions. This result may be because this population, possibly through routine self-tracking, has become more knowledgeable about coping with their illnesses or conditions. Thus, they might have become less inclined to ask new questions or seek second opinions. Finally, we found no correlation between routine self-tracking and the likelihood of altering decisions on how to treat an illness or a condition.

**Discussion**

In our analysis, we investigated if factors such as the person’s health conditions, major recent health events, and the tool(s) used could predict routine self-tracking. We found that having had a recent emergency visit was the only significant health related predictor of routine self-tracking. This increased use of routine tracking practices may be motivated by a need to recover from the crisis, or manage continuing health outcomes resulting from the crisis by learning how to handle it or tracking symptoms. More routine tracking suggests that these people take the practice more seriously, or have a stronger reason to engage in it. This finding suggests that it might be worthwhile to investigate the use of self-tracking by patients who are discharged after a visit to an emergency or urgent unit, both to understand their current practices and to better support their health management after being discharged. It is likely that this demographic represents an opportunity for new technology that aims to support their needs.

The results also suggest that using electronic tools might help to increase routine tracking behavior. Those who used computer programs, websites, mobile applications, and medical devices were significantly more likely to routinely register their data, while those who only kept their measurements by memory were much less likely to use tracking routinely. Concerning the non-electronic tools approached by the survey, using only memory to track health indicators actually presented a negative correlation with routine tracking, while the results for pen and paper were not significant. Electronic tools for registering the tracked data, such as computer programs, websites, and mobile applications, might encourage routine tracking activities through the use of reminders, or features that encourage higher engagement or participation (e.g., social features, and game-like features such as medals and achievements). This finding indicates that, in cases wherein routine tracking is desired or important, users should seek to utilize electronic tools such as computer software or mobile applications to measure or record their data, and avoid keeping their data only by memory. It is also important to investigate the reason for this effect, to better understand what specific aspects of self-tracking tools offer support for tracking routinely.

We also investigated whether routinely self-tracking health indicators could be associated with improved health man-

### Table 4: Model with health management effects as dependent variables.

<table>
<thead>
<tr>
<th></th>
<th>p-value</th>
<th>OR</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Affected a decision about how to treat an illness or condition</td>
<td>0.415</td>
<td>1.11</td>
<td>(0.86-1.44)</td>
</tr>
<tr>
<td>Changed your overall approach to maintaining your health</td>
<td>0.007</td>
<td>1.38</td>
<td>(1.1-1.74)</td>
</tr>
<tr>
<td>Led you to ask a doctor new questions, or to get a second opinion</td>
<td>0.007</td>
<td>0.71</td>
<td>(0.56-0.91)</td>
</tr>
</tbody>
</table>
agement practices. Our results suggest that people who track routinely might become more knowledgeable about their health, and, as a result, adapt their health management approach based on their experiences with tracking, asking less questions to their clinicians. This finding indicates that, in comparison with event-driven tracking, routine tracking might better support patients’ self-efficacy and better assist them to improve their self-management skills. Because both self-efficacy and self-management skills are associated with improved health outcomes26, 27, these results indicate that routinely self-tracking could have a significant beneficial impact on users’ health. While the data analyzed is self-reported and cannot lead to conclusions regarding causality, our results can provide enough evidence to motivate future studies on the effects of routine self-tracking that examine such outcomes in more depth. In particular, future studies should investigate different aspects of health management that are not covered in this analysis, such as the role of caregivers and health outcomes.

In summary, the study results indicate that routinely tracking health indicators can have a significant effect on people’s approaches to their health management, and that might be caused by an increase in self-efficacy. Further, we have found that electronic tools used in tracking are very strong predictors of routine self-tracking, suggesting that they may encourage measuring and recording data at regular intervals. Lastly, we found that people who have had emergency visits to a hospital in the last 12 months are approximately 50% more likely to routinely track health indicators, suggesting that the process of recovery, or learning how to handle new health developments, are situations wherein tracking might be particularly important for patients.

At the same time that routine tracking might improve the results of health management, and it should be supported and encouraged by self-tracking technologies, different users might value routine tracking more than others. For instance, our results indicate that younger people, and those who have not visited the ER recently, are less likely to routinely self-track. In these cases a system that demands routinely tracking to present valuable results may not be desired by the users. These systems should support users’ priorities and capabilities.

This study has several limitations. Because the survey collected self-reported data, it is not possible to attribute causal relationships based on its analysis. The data is likely subject to bias associated with this type of data collection (e.g., social desirability bias, selection bias, acquiescence, halo effect). Additionally, because respondents might track multiple variables using different methods, and the data does not differentiate between these cases, it is possible that these instances created noise in the results. Further, the data provide limited detail regarding health outcomes, the exact variables tracked by each respondent, and the different tools used. These characteristics limit the conclusions that can be drawn from analyses. Lastly, because the survey was conducted in 2012, the influence of technologies on self-tracking activities might have changed due to the increasing popularity of mobile devices, and to new technologies that might have become available since.

Conclusion

Through an analysis of data obtained from the Tracking for Health survey1, we have investigated what health and demographic factors influence routine self-tracking practices, and how continuously measuring and recording data at regular intervals affects health management practices. Our results indicate that older people and those who have recently experienced a health emergency are significantly more likely to self-track routinely. Utilizing electronic self-tracking tools, such as smartphone applications, also increased the probability of routine tracking. Lastly, we found that those who routinely self-track are significantly more likely to report that tracking has influenced their health management practices.

Acknowledgements

We thank the Pew Research Center for making their data available, and CAPES and the NSF (Grant No. 1219197) for providing funding for this work.

References


A Cross-Sectional Study of Prominent US Mobile Health Applications: Evaluating the Current Landscape

Pierre-Antoine Fougerouse, MD¹, Mobin Yasini MD, PhD², Guillaume Marchand, MD², Oliver O. Aalami, MD¹

¹Stanford University School of Medicine, Department of Surgery, Division of Vascular Surgery, Palo Alto, United States; ²DMD Santé, Research and development department, Paris, France

Abstract

Mobile health (mHealth) could offer unprecedented opportunity to provide medical support closer to the users. We have selected some relevant criteria to describe 100 apps from Google Play store and Apple’s App Store’s top suggestions in medical category. These characteristics were compared based on the paid or free nature of the apps, the target users: consumers or healthcare professionals, and the platform: Android or iOS. Seventeen provided functionalities and 27 medical subjects covered by these apps were also extracted. Our study shows that even in top rated mHealth apps, a high proportion lacks some basic criteria regarding the quality of the apps including the presence of a privacy policy, describing content sources, participation of the target users in the app development, etc. Paid apps did not ensure better quality compared to free apps. The current mHealth market is not mature enough to be used widely and recommended by healthcare professionals.

Introduction

Mobile health (m-health) is a subdivision of the eHealth phenomenon that is in a perpetual state of active refinement. There is a vast amount of potential for m-health to positively affect health and healthcare processes¹. Mobile apps offer an unprecedented opportunity to provide medical support at the time/location of the demand². This opportunity can be harnessed by the general public (e.g. patients) or healthcare professionals who are in need of various specialized tools (means of interdisciplinary communication, ready access to health and medical information, simple consultation of medical records, and clinical decision-making support systems)³.

Although m-health apps are a relatively recent development, their virtual presence is exploding with no indication of slowing any time soon. There are over 259,000 health-related applications (apps) available in app stores (e.g. Google Play and Apple’s iOS “app store”) for smartphone devices. Approximately 100,000 m-health apps have been added as of the beginning of 2015⁴. Each month, about 1,000 new applications are put on the market⁵. These apps proffer myriad functionalities, ranging from simple text message reminders to platforms designed for sophisticated disease management. Naturally, the growth of the m-health market correlates with the increase in usage these apps are experiencing. In the next several years, upwards of three million free downloads and circa 300,000 purchased downloads are expected to be made of mHealth applications in the United States alone⁶. These values exhibit a swift rise in the prevalence of mobile technology in the medical and wellness fields. Being readily accessible and simple to procure ensures that smart devices (smartphones and tablets) are made very attractive in the eyes of medical professionals. We are reaching a turning point in the use of mobile technology in healthcare. Therefore there is a pressing need to ensure that patient safety is not compromised post-maturation of the field⁷. The continually-rising number of mHealth apps and the variety of obtainable functionalities render it burdensome for any kind of user (health professional or patient) to ascertain which apps excel in terms of quality.

There exist a great deal of potential dangers, and the reliability of some mHealth apps was investigated in several works of literature⁸–¹⁰. The involvement of a medical professional in the development of a given application¹¹, the accuracy and reliability of the content as utilized in diagnosis and patient management¹², the potential danger of camera functionalities in mobile devices to judge whether skin lesions are suspicious¹³,¹⁴, and deficiencies in self-management applications (in both diabetes¹⁵ and asthma¹⁶) were called into question by various researchers.

While the growing popularity of m-health is well-documented, its impact is not. The reported implications of mHealth interventions are mixed, with studies exhibiting modest benefits for some clinical diagnosis and management support outcomes¹. Furthermore, the information provided in app stores does not permit users to ascertain the quality of apps. The existing five-star rating system currently in use is not a reliable assessment method¹⁷. One is faced with a veritable
jungle of health-related apps, including very good apps as well as those that endanger patient/physician safety. In order to address this deficiency in today’s mobile health market, a research study was proposed to assess the current standing of mHealth applications available for download in the United States. We were interested in determining whether or not the current market could be considered a reliable source of medical applications. The main aim of this study was therefore to isolate characteristics of health apps currently available and identify existing gaps.

Methods

In collaboration with medical doctors from France and Stanford University, we have conducted a cross-sectional descriptive study of the top 100 medical applications available in the Google Play store and Apple's App Store. We selected the top 25 free apps from the App Store followed by 25 of the same from the Google Play Store. When an app appeared in both stores, it was only selected once and passed over in favor of the following app. For the payed apps we employed the opposite strategy, beginning with the Google Play Store and only afterwards the Apple Store. If the reported purpose of the app had no relation to medicine, it was not selected. Other exclusion criteria were that no apps that encapsulated other apps (a mini store of apps) and paid apps more expensive than 70 dollars were not selected.

The order of apps on the top lists are very subject to change. In the beginning of January 2017, we selected our apps in compliance with the aforementioned criteria. We did not alter our choices over the course of the study. The working group was composed of four physicians invested in medical informatics. The apps were installed on a relevant mobile device and analyzed one by one by each group member.

In order to define the criteria used over the course of the study, we hypothesized that we could utilize key elements of the mHealth Quality (mHQ) process. We have isolated from said process the most relevant criteria to assess the American market. The extraction criteria included informative and assessment criteria.

Apart from the characteristics of the app, we have analyzed each app to evaluate the use cases (functionalities) imparted by the app according to a previously published model of use cases for mHealth apps. We have also determined the relevant medical subject/specialty related to each app.

Each member of the working group had an Excel table to fill for each app. We organized a weekly meeting to discuss differing opinions and reach a consensus.

Descriptive data were provided for each criterion. The characteristics of both free and paid apps were compared according to their target demographic (consumers or health professionals) as well as the platform (iOS and Android). Chi-Square test was used to determine whether there was a significant difference between the elements of the information in the categories and P < 0.05 was considered statistically significant.

Results

Half of the analyzed apps were installed on Android phones, and the other half were installed on iOS ones. From each platform 25 paid apps and 25 free apps were selected. Whilst performing the selections, three apps were excluded from the study due to their non-medical nature. One app was not selected due to its high price and two others were excluded because they were essentially a mini store of apps.

The selected criteria were divided into informative elements and assessment elements. The informative elements include type of owner (start-up, corporation, patient association, etc.), target user, the topic the app was attempting to address, whether or not it belonged to the appendix A, B or C of the FDA document on Mobile Medical Applications, etc. Assessment elements include the existence of general terms and conditions of use, existence of privacy policy, presence of a health professional during app development, citation of content sources, etc. Table 1 provides descriptive explanations of various criteria and distribution of apps.

Most app criteria were evaluated in all 100 of the analyzed apps. However, some criteria were not applicable to all apps. For example, "user consent to personal data collection" and “ability to link to a connected device in order to collect data” were rendered obsolete for apps whose purpose was not to collect personal health data. The aforementioned criteria were only applicable to 58 apps. Another example is "citation of the content sources or bibliographic references" and "medical content is kept up to date" because only 71 of selected apps include medical content.

The pecuniary nature of some apps caused a significant difference (p<0.001) between those addressed to consumers and those targeting healthcare professionals. Apps whose target demographic consisted mostly of healthcare professionals were less likely to be free to download. However, 28% of free apps had in-app integrated purchases.
Table 1. Characteristics description of the health apps available on the US market.

<table>
<thead>
<tr>
<th>Characteristics (n)</th>
<th>Number of apps (%)</th>
<th>Price</th>
<th>Target users</th>
<th>Platform</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Free apps (%)</td>
<td>Paid apps (%)</td>
<td>P-value</td>
<td>C. (%)</td>
</tr>
<tr>
<td><strong>App provider (100)</strong></td>
<td></td>
<td></td>
<td></td>
<td>0.517</td>
</tr>
<tr>
<td>Corporate</td>
<td>30 (30)</td>
<td>13 (43)</td>
<td>17 (57)</td>
<td>19 (63)</td>
</tr>
<tr>
<td>Healthcare institution/ enterprise</td>
<td>16 (16)</td>
<td>6 (38)</td>
<td>10 (63)</td>
<td>9 (56)</td>
</tr>
<tr>
<td>Start up</td>
<td>35 (35)</td>
<td>20 (57)</td>
<td>15 (43)</td>
<td>27 (77)</td>
</tr>
<tr>
<td>Patient association / Non-profit</td>
<td>10 (10)</td>
<td>5 (50)</td>
<td>5 (50)</td>
<td>5 (50)</td>
</tr>
<tr>
<td>Individual</td>
<td>9 (9)</td>
<td>6 (67)</td>
<td>3 (33)</td>
<td>6 (67)</td>
</tr>
<tr>
<td><strong>if corporate (30)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Device company</td>
<td>3 (10)</td>
<td>3 (100)</td>
<td>0 (0)</td>
<td>3 (100)</td>
</tr>
<tr>
<td>Electronic Health Record</td>
<td>2 (7)</td>
<td>2 (100)</td>
<td>0 (0)</td>
<td>1 (50)</td>
</tr>
<tr>
<td>Insurance</td>
<td>4 (13)</td>
<td>4 (100)</td>
<td>0 (0)</td>
<td>4 (100)</td>
</tr>
<tr>
<td>Drug</td>
<td>0 (0)</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>Services provider</td>
<td>16 (53)</td>
<td>1 (6)</td>
<td>15 (94)</td>
<td>9 (56)</td>
</tr>
<tr>
<td>Other</td>
<td>5 (17)</td>
<td>3 (60)</td>
<td>2 (40)</td>
<td>2 (40)</td>
</tr>
<tr>
<td><strong>Price (100)</strong></td>
<td></td>
<td></td>
<td></td>
<td>&lt;10⁻³</td>
</tr>
<tr>
<td>Free</td>
<td>50 (50)</td>
<td>.</td>
<td>.</td>
<td>43 (86)</td>
</tr>
<tr>
<td>Paid</td>
<td>50 (50)</td>
<td>.</td>
<td>.</td>
<td>23 (46)</td>
</tr>
<tr>
<td><strong>If free, in-app purchase (50)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>36 (72)</td>
<td>36 (100)</td>
<td>0</td>
<td>33 (92)</td>
</tr>
<tr>
<td>Yes</td>
<td>14 (28)</td>
<td>14 (100)</td>
<td>0</td>
<td>10 (71)</td>
</tr>
<tr>
<td><strong>Target user (100)</strong></td>
<td></td>
<td></td>
<td></td>
<td>&lt;10⁻³</td>
</tr>
<tr>
<td>Consumers</td>
<td>66 (66)</td>
<td>43 (65)</td>
<td>23 (35)</td>
<td>.</td>
</tr>
<tr>
<td>Health professionals</td>
<td>34 (34)</td>
<td>7 (21)</td>
<td>27 (79)</td>
<td>.</td>
</tr>
<tr>
<td><strong>Existence of a &quot;contact&quot; tab for the users to ask their questions (100)</strong></td>
<td></td>
<td></td>
<td></td>
<td>0.663</td>
</tr>
<tr>
<td>No</td>
<td>30 (30)</td>
<td>14 (47)</td>
<td>16 (53)</td>
<td>21 (70)</td>
</tr>
<tr>
<td>Yes</td>
<td>70 (70)</td>
<td>36 (51)</td>
<td>34 (49)</td>
<td>45 (64)</td>
</tr>
<tr>
<td><strong>If yes, is there a hotline? (70)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>52 (74)</td>
<td>22 (42)</td>
<td>30 (58)</td>
<td>32 (62)</td>
</tr>
<tr>
<td>Yes</td>
<td>18 (26)</td>
<td>14 (78)</td>
<td>4 (22)</td>
<td>13 (72)</td>
</tr>
<tr>
<td>Description of the general terms and conditions of use (100)</td>
<td>0.016</td>
<td>0.405</td>
<td>0.420</td>
<td></td>
</tr>
<tr>
<td>---------------------------------------------------------------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td></td>
</tr>
<tr>
<td><strong>No</strong></td>
<td>56 (56)</td>
<td>22 (39)</td>
<td>34 (61)</td>
<td>35 (63)</td>
</tr>
<tr>
<td><strong>Yes</strong></td>
<td>44 (44)</td>
<td>28 (64)</td>
<td>16 (36)</td>
<td>31 (70)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Description of personal data privacy policy (100)</th>
<th>&lt;10⁻³</th>
<th>0.003</th>
<th>0.316</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>No</strong></td>
<td>47 (47)</td>
<td>11 (23)</td>
<td>36 (77)</td>
</tr>
<tr>
<td><strong>Yes</strong></td>
<td>53 (53)</td>
<td>39 (74)</td>
<td>14 (26)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>User consent for the personal data collection (58)</th>
<th>&lt;10⁻³</th>
<th>-</th>
<th>0.346</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>No</strong></td>
<td>24 (41)</td>
<td>9 (38)</td>
<td>15 (63)</td>
</tr>
<tr>
<td><strong>Yes</strong></td>
<td>34 (59)</td>
<td>28 (82)</td>
<td>6 (18)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>The app can link to a connected device to collect data (58)</th>
<th>0.187</th>
<th>-</th>
<th>0.431</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>No</strong></td>
<td>44 (76)</td>
<td>26 (59)</td>
<td>18 (41)</td>
</tr>
<tr>
<td><strong>Yes</strong></td>
<td>14 (24)</td>
<td>11 (79)</td>
<td>3 (21)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>FDA category A, B, C (100)</th>
<th>0.109</th>
<th>&lt;10⁻³</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>52 (52)</td>
<td>22 (42)</td>
<td>30 (58)</td>
</tr>
<tr>
<td>B</td>
<td>48 (48)</td>
<td>28 (58)</td>
<td>20 (42)</td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>.</td>
<td>.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>FDA approval for group B (48)</th>
<th>-</th>
<th>-</th>
<th>-</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>No</strong></td>
<td>47 (98)</td>
<td>27 (57)</td>
<td>20 (43)</td>
</tr>
<tr>
<td><strong>Yes</strong></td>
<td>1 (2)</td>
<td>1 (100)</td>
<td>0 (0)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Other certification approved? (48)</th>
<th>-</th>
<th>-</th>
<th>-</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>No</strong></td>
<td>46 (96)</td>
<td>26 (57)</td>
<td>20 (43)</td>
</tr>
<tr>
<td><strong>Yes</strong></td>
<td>2 (4)</td>
<td>2 (100)</td>
<td>0 (0)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>The app has passed a usability test process (100)</th>
<th>-</th>
<th>-</th>
<th>-</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>No</strong></td>
<td>100 (100)</td>
<td>50 (50)</td>
<td>50 (50)</td>
</tr>
<tr>
<td><strong>Yes</strong></td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>0 (0)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Participation of the target users in the conception or development of the app (100)</th>
<th>0.825</th>
<th>&lt;10⁻³</th>
<th>0.825</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>No</strong></td>
<td>71 (71)</td>
<td>35 (49)</td>
<td>36 (51)</td>
</tr>
<tr>
<td><strong>Yes</strong></td>
<td>29 (29)</td>
<td>15 (52)</td>
<td>14 (48)</td>
</tr>
</tbody>
</table>
We did not find a significant difference between target users of Android vs. iOS apps, none of the two major markets is oriented more towards consumers or healthcare professionals.

Seventy apps presented a means of contacting the owner via email. There was no significant difference between free and paid apps, between those addressed to consumers vs. healthcare professionals, or between the two platforms. 26% of selected apps provide a hotline to address users’ potential queries. There was a significant difference ($p < 0.009$) between free and paid apps: free apps provide more hotline services than paid ones.
More than half of the apps did not provide general terms and conditions of use. We found no substantial difference in availability between Android/iOS apps or between consumer-targeted/healthcare-professional-targeted ones. However, a larger amount of free applications offer general terms and conditions in juxtaposition with paid apps ($p<0.016$). This phenomenon occurs once again with the availability of privacy policies. Free apps include privacy policies more frequently than paid apps ($p<0.001$).

Only 59% of the apps collecting user health data require the explicit consent of their users pre-collection, and free apps were once again superior ($p<0.001$). No difference was found when we compared the two major platforms and target users.

There was only one app that boasted an ethical chart. This app was free, meant for healthcare professionals, and available in the Top 25 of both the Google Play and App stores (in compliance with the aforementioned selection criteria we studied the iOS version).

Seventeen disparate use cases were discovered in the 100 analyzed apps according to the Yasini and Marchand Model$^3$. Six use cases were related to health professionals and eleven to consumers. We discovered 226 use cases in these apps, meaning that the average number of use cases provided by each app is 2.26. Table 2 illustrates the various uses cases and their frequency in the selected apps.

**Table 2. Frequency of various uses cases in the top apps of the two major US stores.**

<table>
<thead>
<tr>
<th>Use Cases</th>
<th>number</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>C. Calculate and/or interpret data</td>
<td>30</td>
<td>13,3</td>
</tr>
<tr>
<td>C. Communicate/share information, social network</td>
<td>7</td>
<td>3,1</td>
</tr>
<tr>
<td>C. Database (drug, image, nutrition...)</td>
<td>21</td>
<td>9,3</td>
</tr>
<tr>
<td>C. Diagnostic/measurement tool</td>
<td>12</td>
<td>5,3</td>
</tr>
<tr>
<td>C. Health news</td>
<td>4</td>
<td>1,8</td>
</tr>
<tr>
<td>C. Information/Scientific popularization/Therapeutic patient education</td>
<td>19</td>
<td>8,4</td>
</tr>
<tr>
<td>C. Interaction with a health institution (Scheduling an appointment, Drug ordering...)</td>
<td>17</td>
<td>7,5</td>
</tr>
<tr>
<td>C. Locating a health service</td>
<td>18</td>
<td>8,0</td>
</tr>
<tr>
<td>C. Looking for information on health professionals/institutions</td>
<td>10</td>
<td>4,4</td>
</tr>
<tr>
<td>C. Tracking a physiopathological state</td>
<td>29</td>
<td>12,8</td>
</tr>
<tr>
<td>C. Treatment reminder/Managing the drug stock</td>
<td>6</td>
<td>2,7</td>
</tr>
<tr>
<td>HP. Database (drug, image, bibliography...)</td>
<td>22</td>
<td>9,7</td>
</tr>
<tr>
<td>HP. Decision support system, calculate and/or interpret data</td>
<td>3</td>
<td>1,3</td>
</tr>
<tr>
<td>HP. Diagnostic/measurement tool</td>
<td>1</td>
<td>0,4</td>
</tr>
<tr>
<td>HP. Managing professional activities (searching for job offers, calculating the fees...)</td>
<td>1</td>
<td>0,4</td>
</tr>
<tr>
<td>HP. Text book, journal, guidelines and synthesis of medical knowledge</td>
<td>18</td>
<td>8,0</td>
</tr>
<tr>
<td>Total</td>
<td>226</td>
<td>100</td>
</tr>
</tbody>
</table>

C: Consumers, HP: Healthcare Professionals

We were able to classify the chosen apps in 27 subject groups, of which 15 were consumer-oriented and 12 targeted health professionals. For consumer-targeted applications the most well-covered subject was general health status (29%), most often exhibited by apps attempting to centralize all patient data in order to provide a comprehensive synthesis of such. We also noted that an integral portion of consumer apps (18%) focused upon women's health issues such as menstrual cycles and pregnancy concerns. The subject of eight apps (12%) was managing the use and cost optimization of medications.

For apps directed towards healthcare professionals, atlases of anatomy dominated (26%). Second place was held by applications that could be harnessed by professionals of all medical specialties (20%) (e.g. drug databases). 17% of apps specifically for healthcare workers dealt with emergency medicine, and 11% of apps were created to aid students’ exam preparation. Figure 1 portrays the various subjects and their relative frequency.
Discussion

In this study, we analyzed 100 US mHealth applications available in the medical category of the Google Play store and Apple's iOS App Store. We described some of the characteristics of these apps including the app's target user, the app provider, presence of a physician in the core team of the company owning the app, providing the content sources when relevant, and providing terms and conditions of use and privacy policy. We also extracted the subjects they dealt with and their various use cases.

Although we have only analyzed the apps in the medical category, we have excluded three apps because of non-relevant topic to medicine. One of these apps was a music app and two others were games (that were not considered as serious medical related games). This shows that even in the top-rated apps in the market stores, some apps are not well categorized.

Contrary to our impression, paid apps did not provide a more qualified profile compared to free apps based on the evaluated criteria. This may be explained by the fact that mobile apps are often used by providers as another marketing tool to create loyal users. mHealth market is still in the phase that the app providers want to build their brand image and get their name out. Therefore, most of the providers select free apps to increase the download rate and potentially provide more qualified apps to keep their users continually.

Various mHealth apps collect and offer critical and private patient data. The data may be entered directly by the user, collected with the built-in features of the smart device (e.g., camera, microphone), an external sensor, or a connected device (e.g., connected thermometer). These data need a special focus on information security and privacy of mHealth
Potential damage to users through information security and privacy infringements have been discussed in various studies\textsuperscript{20–22}. Our study shows that still about half of the analyzed apps do not provide terms and conditions of use or privacy policy. The situation becomes more dramatic when we consider that the analyzed apps come from the top of the App Stores.

The involvement of healthcare professionals in mHealth apps development processes is another important criterion that ensures the reliability of health information\textsuperscript{17,23,24}. This study showed that 61\% of the studied apps were produced with the participation of a health professional in the core team of the app owner. Other studies in the literature reported the rate of this involvement between 12.8 \% to 48\% of analyzed apps\textsuperscript{20,24}. Although in our sample of apps we found a higher rate of professional involvement in the development of the apps, this is not yet satisfying particularly for the top listed apps of the App Stores.

It is critical that medical applications provide content that are accurate and reliable\textsuperscript{25} because patients and healthcare professionals can make critical decisions based on information provided by an app. We did not find the content source for 44\% of apps that provide medical information to their target users. When mHealth apps disseminate medical information, which was the case for 71\% of the analyzed apps, the information must be developed on the basis of reliable information resources and references. This study showed that 44\% of these apps did not mention their sources. This lack of transparency of the apps about the sources is in line with other studies in the literature who call in to question the adherence of the apps to established guidelines\textsuperscript{20,24,25}.

None of the studied apps could be considered as a medical device according to FDA classification\textsuperscript{19}. However, half of these apps were considered as borderline apps (group B). The need for regulating the apps is a real challenge of today’s mHealth market. The percentage of apps that resort to FDA authorization or other existing certifications (like European CE for medical devices) is too low. This shows that the market is not yet enough mature to the assessment and certification processes.

Various apps were multifunctional and provided various use cases in different medical fields. This indicates that the iOS and Android App Stores offer a wide selection of mHealth apps with a real diversity of usages that makes mobile devices ready to do almost everything in every medical specialty. However, this diversity prevents a “one size fits all” approach to ensuring the quality of these apps and information security\textsuperscript{18,21}.

One of the limitations of our study was the number of studied apps. The top-rated apps in the App Stores are based on the five-star rating that the users provide, number of downloads and some other criteria. However, the number of downloads may differ according to the target users of the apps (an app for diabetic patients may be downloaded more than an app that is addressed to the patients affected with a rare disease) and the five-star rating is not a reliable assessment system\textsuperscript{17,26}. In this study, we have selected all the studied apps from US App Stores. Further research with a multi country design and a more important number of apps to analyze is the real solution to validate these results.

Conclusion

mHealth apps could provide seamless access to adapted and context oriented health information and have the potential to decrease global health burdens. However, despite these promising potentials, there are some barriers and risks that prevent the users to trust mHealth apps. Recognizing these elements by all the stakeholders including app providers, app stores and users may lead to removal of these risks step by step. We are faced with a plethora and diversity of available mHealth apps. Therefore, implications for assuring the quality, security and privacy seem to be unclear and complex. Our study shows that the current situation of mHealth apps in general is not yet enough mature to be used widely and recommended by health professionals in the US market. Further research and development efforts are required to facilitate the wide range integration of mHealth apps in the real clinical pathways. These efforts could be designed in various axes including disseminating related guidelines for the users or app developers, implementing regulatory certification programs, designing adapted studies to reveal clinical evidence of mHealth apps considering the fast-paced nature of technology, and educating and encouraging app providers to protect their apps from information security and privacy and to test the usability and ergonomic aspects of their apps. Taking the right strategies in the future will help to benefit from the potential of mHealth apps to transform and improve the health care ecosystem and limit the barriers and risks.

References

4. research2guidance - The mHealth apps market is getting crowded. [Internet]. [cited 2017 Feb 28]. Available from: https://research2guidance.com/mhealth-app-market-getting-crowded-259000-mhealth-apps-now/
Achieving Logical Equivalence between SNOMED CT and ICD-10-PCS
Surgical Procedures

Kin Wah Fung¹, MD, MS, MA; Julia Xu¹, MD, PhD; Filip Ameye² MD, FRCS(Eng), FACS; Arturo Romero Gutiérrez³, MD; Arabella D’Havé⁴, MSc

¹National Library of Medicine, Bethesda, MD; ²National Institute for Health and Disability Insurance, Belgium; ³Ministry of Health, Social Services and Equality, Spain; ⁴Federal Public Service of Health, Food Chain Safety and Environment, Belgium

kwfung@nlm.nih.gov|xujc@mail.nlm.nih.gov|filip.ameye@riziv.fgov.be|aromerog@msssi.es|arabella.dhave@health.belgium.be

Abstract
Surgical procedures are coded in SNOMED CT in the electronic health record and in ICD-10-PCS in administrative systems. We compared the logical definitions of SNOMED CT concepts to the ICD-10-PCS axial components to identify overlap and gaps. The biggest discrepancy was in the surgical approach which was specified in all ICD-10-PCS codes but only in 8.7% of SNOMED CT surgical procedures. Among the top 100 commonly used ICD-10-PCS codes, 25% could be matched fully in meaning and logical definition to pre-coordinated SNOMED CT concepts. Using post-coordination, it was possible to represent the full meaning of 86% of ICD-10-PCS codes. Logical mapping between SNOMED CT and ICD-10-PCS is feasible but will be more productive if more SNOMED CT concepts can become fully-defined. Short of full logical matching, partial logical matches can also be useful in suggesting candidate maps for expert review and to support interactive post-coordination.

Introduction
The Standardized Nomenclature of Medicine, Clinical Terms (SNOMED CT) is steadily gaining momentum as the emerging international clinical terminology standard. The number of member countries has more than tripled (increased from 9 in 2007 to 30 in 2017) since the establishment of the International Health Terminology Standards Development Organisation (IHTSDO) in 2007. In early 2017, the IHTSDO has acquired the new name SNOMED International. SNOMED International now covers most primarily English speaking countries.¹ In the U.S., the Meaningful Use of Electronic Health Record (EHR) incentive of Centers for Medicare and Medicaid Services (CMS)², ³ specifies SNOMED CT as the terminology for the encoding of problem lists and procedures for the EHR, among other data elements. A similar effort is underway in Belgium to encourage the meaningful use of EHR and the use of standard terminologies like SNOMED CT.⁴ The adoption of SNOMED CT for clinical documentation is anticipated to increase steadily as the number of member countries continues to grow.⁵, ⁶

In the U.S. and in Belgium, for administrative and reimbursement purposes, hospital-based medical procedures are being reported in ICD procedure codes. Since 2015, ICD-9-CM procedure codes have been replaced by ICD-10-PCS (International Classification of Diseases, 10th Revision, Procedure Coding System) in both countries. ICD-10-PCS is a brand-new procedure classification system created by the U.S. Centers for Medicare and Medicaid Services (CMS) through a contract with 3M Health Information Systems.⁷ Other countries that have been using ICD-9-CM are expected to make similar transition to ICD-10-CM and ICD-10-PCS. For example, Spain has made the switch in 2016 and Portugal in March 2017.

While SNOMED CT and ICD-10-PCS are both used to encode procedures, they are designed with different principles to serve their specific purposes. It is likely that both coding systems will continue to be used. One way to mitigate the problem of having clinical and administrative data coded in disparate coding systems is to develop a map between the two systems. Under SNOMED International, there is a SNOMED CT to ICD-10-PCS Mapping Project Group with participation from U.S., Spain and Belgium to explore the creation of such a map.⁸ The group has studied various ways of automatically mapping between the two terminologies. These methods include lexical matching of the ICD-
10-PCS index, indirect mapping through the General Equivalence Map (GEMs, published by CMS) and ontological alignment between the SNOMED CT attributes and ICD-10-PCS axes. Among these approaches, lexical mapping has higher precision (89% of maps are useful) but low coverage (only 3% of SNOMED CT surgical procedure concepts can be mapped). Indirect mapping through GEM has lower accuracy (76% of maps are useful) but can find map for 10% of SNOMED CT surgical procedures. Ontological mapping has the lowest accuracy (45% of maps are useful) but can find map for 46% of the SNOMED CT procedure concepts studied.

In this study, we are expanding on our effort in ontological mapping. There are two prerequisites in order for ontological mapping to work. First, there has to be a high degree of similarity or compatibility between the underlying concept models of both systems. Second, the existing contents must be modeled adequately to support automatic mapping between both systems. Based on a list of frequently used ICD-10-PCS surgical procedure codes from a hospital information system, we assessed the extent to which the SNOMED CT concept model can cover the meaning of ICD-10-PCS codes. We also examined the degree of modeling of existing (pre-coordinated) SNOMED CT concepts and the feasibility of using post-coordination to achieve logical equivalence between SNOMED CT and ICD-10-PCS surgical procedures. We report our findings here and discuss the implications of our findings with respect to the quality of SNOMED CT modeling and the practicability of ontological mapping.

Methods

ICD-10-PCS is a radical departure from ICD-9-CM in terms of structure and design. Each procedure is defined by seven components (axes) namely section, body system, root operation, body part, approach, device and a 7th character for a qualifier if applicable. On the SNOMED CT side, procedure concepts are similarly defined by components (attributes) such as procedure site, method, access, device etc. The basis of ontological mapping is to align the ICD-10-PCS axes and SNOMED CT attributes in order to identify equivalence in meaning by comparing the logical definitions of concepts. In this study, we did a quantitative analysis of the attributes used in existing SNOMED CT surgical procedures and compared them to the corresponding components in ICD-10-PCS to identify overlaps and potential gaps.

We acquired a list of ICD-10-PCS codes with usage frequencies extracted from the data warehouse of Nebraska Medicine - Medical Center, a 600 acute-care bed facility with more than 1,000 physicians in all major specialties and subspecialties. In our study, we focused on medical and surgical procedures from section 0. The various sections can be distinguished by the first digit of the ICD-10-PCS code (e.g., 0 = medical and surgical, 1 = obstetrics, 2 = placement, etc.). Medical and surgical procedures represent the majority (86%) of all ICD-10-PCS codes. We identified the 100 most frequently used ICD-10-PCS medical and surgical procedures (codes beginning with 0). For each ICD-10-PCS code, two authors (JX and FA) independently looked for the SNOMED CT concept that most closely matched the meaning of the ICD-10-PCS procedure (meaning match). The meaning of the SNOMED CT concept was determined by its fully-specified name. The meaning match was characterized as exact or partial. For exact meaning matches, the logical definition of the SNOMED CT concept was examined to see whether it captured the full meaning of the ICD-10-PCS procedure (logical match). If the logical definition did not completely match the ICD-10-PCS procedure, it was extended according to the SNOMED CT concept model to see if full logical equivalence could be achieved. If full logical equivalence could not be achieved, the reason was recorded. The reasons might include: limitation of SNOMED CT concept model, missing necessary attribute, missing necessary value. For cases that the closest SNOMED CT concept was a partial meaning match, that SNOMED CT concept was used as the template to construct a post-coordinated SNOMED CT expression to capture the meaning of the ICD-10-PCS code as fully as possible. If full logical representation was not possible with post-coordination, the reason of failure was recorded. Results from the two reviewers were compared and differences discussed with a goal to reach consensus. A third reviewer (KWF) cast a third vote if consensus could not be reached. All three reviewers (all physicians) together discussed cases in which there was exact meaning match but incomplete logical match to see if the missing defining element was clinically significant. We found this to be necessary because some of the ICD-10-PCS component values were clinically trivial and missing them would not cause ambiguity in the clinical context. For example, circumcision was defined in ICD-10-PCS as having the external approach. Since there was no other approach to perform circumcision,
missing the external approach in the logical definition was considered not clinically significant. The review procedure is summarized in figure 1.

Figure 1. Finding meaning and logical definition match for ICD-10-PCS codes (FSN: fully-specified name)

**Results**

**Concept model analysis**

There are 19,796 surgical procedures in the January 2017 release of SNOMED CT identified as descendants of the concept *Surgical procedure (387713003)*. In the SNOMED CT concept model, 24 attributes (excluding the attribute *Is a*) are allowed for the modeling of procedures. (Table 1) The frequency of usage of these attributes is highly variable. For example, while the *Method* attribute is used in all concepts, 9 attributes (e.g., *Procedure morphology, Indirect morphology, Procedure device*) are used in less than 1% of concepts.

<table>
<thead>
<tr>
<th>Attribute</th>
<th>% concepts with attribute</th>
<th>Attribute</th>
<th>% concepts with attribute</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method</td>
<td>100.00%</td>
<td>Surgical approach</td>
<td>3.27%</td>
</tr>
<tr>
<td>Procedure site - direct</td>
<td>85.23%</td>
<td>Has intent</td>
<td>3.18%</td>
</tr>
<tr>
<td>Direct morphology</td>
<td>28.60%</td>
<td>Has focus</td>
<td>1.90%</td>
</tr>
<tr>
<td>Procedure site - indirect</td>
<td>13.27%</td>
<td>Procedure morphology</td>
<td>0.69%</td>
</tr>
<tr>
<td>Using device</td>
<td>13.14%</td>
<td>Indirect morphology</td>
<td>0.68%</td>
</tr>
<tr>
<td>Procedure site</td>
<td>8.80%</td>
<td>Priority</td>
<td>0.36%</td>
</tr>
<tr>
<td>Revision status</td>
<td>4.89%</td>
<td>Using energy</td>
<td>0.30%</td>
</tr>
<tr>
<td>Using access device</td>
<td>4.48%</td>
<td>Procedure device</td>
<td>0.27%</td>
</tr>
<tr>
<td>Access</td>
<td>4.30%</td>
<td>Indirect device</td>
<td>0.08%</td>
</tr>
<tr>
<td>Direct device</td>
<td>4.19%</td>
<td>Recipient category</td>
<td>0.03%</td>
</tr>
<tr>
<td>Direct substance</td>
<td>3.76%</td>
<td>Has specimen</td>
<td>0.01%</td>
</tr>
<tr>
<td>Using substance</td>
<td>3.45%</td>
<td>Route of administration</td>
<td>0.01%</td>
</tr>
</tbody>
</table>

Table 1. Frequency of use of attributes in SNOMED CT surgical procedure concepts

In ICD-10-PCS, there are seven axes each corresponding to a specific aspect of a procedure. (Table 2) The ICD-10-PCS code is made up of seven alphanumeric characters, one for each axis. The first five axes are always populated...
with clinically meaningful values (e.g., stomach, excision). In axis 6 (device) and axis 7 (qualifier), a placeholder value (Z) is allowed for null values when a component is absent (i.e., no device, no qualifier). Table 2 shows the proportion of ICD-10-PCS codes (2016 version) with non-null values for each of the axes, compared to the proportion of SNOMED CT concepts with attributes corresponding to a particular axis. For example, the root operation is always specified in ICD-10-PCS and a Method attribute is always present in SNOMED CT. Body system and body part are specified in all ICD-10-PCS codes. In SNOMED CT, body site is modeled by three attributes: Procedure site and its two children Procedure site – direct and Procedure site – indirect. Procedure site – direct denotes the body structure that is the direct aim of the procedure. For example, Amputation of the foot has Procedure site - direct = Foot structure. Procedure site – indirect is used when the body structure is not the direct target of the procedure. For example, Removal of calculus of urinary bladder has Procedure site – indirect = Urinary bladder structure. The more general attribute Procedure site is used to model high-level grouper type concepts such as Procedure on colon. However, sometimes the use of Procedure site and Procedure site – direct attributes is conflated. So in our analysis, body system and body part in ICD-10-PCS are considered to be equivalent to any of the three procedure site attributes in SNOMED CT. Overall, 97.1% of SNOMED CT concepts have at least one of the procedure site attributes. In ICD-10-PCS, body sites can also be used in axis 7 to denote the second body part of a bypass procedure (which typically connects two body parts). For example, for Bypass stomach to jejunum, axis 4 has the value stomach and axis 7 has value jejunum. In SNOMED CT bypass procedures, both anatomic sites being connected are represented as Procedure site – direct attributes. So in the case of bypass procedures, Axis 7 corresponds to the second (if present) Procedure site – direct attribute in SNOMED CT. There are other uses of axis 7 (e.g., diagnostic) which correspond to other SNOMED CT attributes (e.g., Has intent). Overall, while there are 17.2% of ICD-10-PCS codes with non-null axis 7 values, while 8.9% of SNOMED CT concepts have some corresponding attributes. The biggest discrepancy between SNOMED CT and ICD-10-PCS is in the surgical approach value, which is present in all ICD-10-PCS codes but only in 8.7% of SNOMED CT concepts.

<table>
<thead>
<tr>
<th>ICD-10-PCS axis</th>
<th>% of ICD-10-PCS codes with non-null values</th>
<th>Corresponding SNOMED CT attributes</th>
<th>% of SNOMED CT concepts with any of the attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Section</td>
<td>100%</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>2. Body system</td>
<td>100%</td>
<td>Procedure site Procedure site – direct Procedure site – indirect</td>
<td>97.1%</td>
</tr>
<tr>
<td>3. Root operation</td>
<td>100%</td>
<td>Method</td>
<td>100%</td>
</tr>
<tr>
<td>4. Body part</td>
<td>100%</td>
<td>Procedure site Procedure site – direct Procedure site – indirect</td>
<td>97.1%</td>
</tr>
<tr>
<td>5. Approach</td>
<td>100%</td>
<td>Access Using access device</td>
<td>8.7%</td>
</tr>
<tr>
<td>6. Device</td>
<td>54.5%</td>
<td>Using device Direct device Procedure device Direct substance Using substance</td>
<td>23.5%</td>
</tr>
<tr>
<td>7. Qualifier</td>
<td>17.2%</td>
<td>Procedure site – direct Has intent Direct substance Using substance</td>
<td>8.9%</td>
</tr>
</tbody>
</table>

Table 2. Distribution of non-null components in ICD-10-PCS and their corresponding attributes in SNOMED CT concepts (NA: not applicable)

*counting only bypass procedures with two or more Procedure site – direct attributes
**Concept matching**

<table>
<thead>
<tr>
<th>Match category</th>
<th>Number of concepts</th>
<th>Missing attribute</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Body site</td>
</tr>
<tr>
<td>Exact meaning match</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Full logical equivalence</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2. Close logical equivalence</td>
<td>24</td>
<td>2</td>
</tr>
<tr>
<td>3. Partial logical equivalence</td>
<td>7</td>
<td>3</td>
</tr>
<tr>
<td>Partial meaning match</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Full post-coordination</td>
<td>56</td>
<td>33</td>
</tr>
<tr>
<td>2. Partial post-coordination</td>
<td>13</td>
<td>9</td>
</tr>
<tr>
<td>Total</td>
<td>100</td>
<td>45</td>
</tr>
</tbody>
</table>

Table 3. Meaning and logical definition match between SNOMED CT and ICD-10-PCS and the distribution of missing attributes

The following is a detailed explanation of each match category with examples.

1. **Exact meaning match, full logical equivalence** (1 case)

   The ICD-10-PCS procedure *Occlusion of Bilateral Fallopian Tubes, Open Approach (0UL70ZZ)* matched exactly the meaning of the SNOMED CT concept *Open bilateral occlusion of fallopian tubes (176936004)*. The SNOMED CT logical definition fully captured the ICD-10-PCS components as shown below (drawn according to SNOMED CT Diagramming Guide\(^{16}\), omitting *Is a* relationships, same below):

   ![Diagram](image)

2. **Exact meaning match, close logical equivalence** (24 cases)

   The meaning of the SNOMED CT concept and ICD-10-PCS procedure matched exactly but the SNOMED CT logical definition did not capture all the ICD-10-PCS components. However, the missing component was considered clinically trivial. For example, ICD-10-PCS procedure *Resection of Appendix, Percutaneous Endoscopic Approach (0DTJ4ZZ)* exactly matched the meaning of *Laparoscopic appendectomy (6025007)*, which had these attributes:

   ![Diagram](image)
Strictly speaking, this was not an exact logical match for two reasons. First, the percutaneous approach was not represented. Moreover, ICD-10-PCS made a distinction between removal of the whole (resection) and part (excision) of a body part. In order to be logically equivalent, Entire appendix would have to be used instead of Appendix structure for the Procedure site – direct attribute. However, we did not consider these differences to be clinically significant because typically laparoscopic surgery was done percutaneously, and appendectomy implied removal of the whole appendix. Omitting these attributes did not significantly change the meaning of the concept. The logical match was close enough in the clinical context. Other examples of close logical matches included omission of the approach for endoscopic procedures (specified as ‘via natural or artificial opening’ in ICD-10-PCS) and the intent in biopsies (specified as ‘diagnostic’ in ICD-10-PCS). To facilitate automatic logical mapping, it would be possible to add back the missing attributes algorithmically (see Implications for logical mapping in Discussion).

3. Exact meaning match, partial logical equivalence (7 cases)

In these cases, the meaning match was exact, but the omission in the logical definition was clinically significant. For example, ICD-10-PCS procedure Bypass Stomach to Jejunum, Percutaneous Endoscopic Approach (0D164ZA) matched exactly the meaning of 307195003 Laparoscopic gastroenterostomy (307195003) with the following attributes:

The components of the ICD-10-PCS code could be captured completely by replacing the attribute Procedure site – indirect = Digestive tract structure by two attributes Procedure site – direct = Stomach structure and Procedure site – direct = Jejunal structure. Of the 7 cases we found, 6 cases could have completely matching logical definitions using the existing SNOMED CT concept model. In the remaining case of the ICD-10-PCS procedure Excision of Stomach, Percutaneous Endoscopic Approach, Vertical (0DB64Z3), complete logical representation in SNOMED CT was not possible because the surgical method (sleeve resection or vertical resection) was not available in SNOMED CT to fully represent the procedure in Laparoscopic sleeve gastrectomy (427074001).

4. Partial meaning match, full representation possible by post-coordination (56 cases)

The SNOMED CT concept matched part of the ICD-10-PCS meaning and a post-coordinated expression could be constructed based on that SNOMED CT concept to represent all the ICD-10-PCS components. For example, the ICD-10-PCS procedure Dilation of Left Ureter with Intraluminal Device, Via Natural or Artificial Opening Endoscopic (0T778DZ) was partially matched in meaning to Insertion of stent into ureter (428817001). If we changed the Procedure site - indirect attribute value to Structure of left ureter and added the attribute Using access device = Endoscope, device we could capture all the components of the ICD-10-PCS procedure.

Partial meaning match found in SNOMED CT concept Insertion of stent into ureter (428817001):
Post-coordinated expression could be created to express full meaning of ICD-10-PCS code *Dilation of Left Ureter with Intraluminal Device, Via Natural or Artificial Opening Endoscopic (0T778DZ)*:

5. Partial meaning match, full representation not possible by post-coordination with current SNOMED CT concept model (13 cases)

In these cases, full representation of the ICD-10-PCS meaning was not possible because of one of the following reasons:

- Cardinality – the current SNOMED CT description logic profile was not able to express the exact number of occurrence of an attribute. So it was not possible to capture the number of bypass grafts in this ICD-10-PCS procedure *Bypass Coronary Artery, Two Arteries from Aorta with Autologous Venous Tissue, Open Approach (021109W)*.
- Absence of body part concept – there was no SNOMED CT body part concept corresponding to the ICD-10-PCS procedure. For example, in the ICD-10-PCS procedure *Insertion of Monitoring Device into Upper Artery, Percutaneous Approach (03HY32Z)*, upper artery is defined as any artery above the diaphragm in ICD-10-PCS. In the anatomy model of SNOMED CT, there was no such corresponding body part concept.
- Absence of surgical action concept – the required action concept was not available in SNOMED CT. For example, in the ICD-10-PCS procedure *Extraction of Abdomen Skin, External Approach (0HD7XZZ)*, extraction referred to removal by any means other than surgical excision, e.g. debridement. In SNOMED CT, there was no action concept corresponding to ‘non-excisional removal’.

**Discussion**

SNOMED CT distinguishes itself from most other terminologies by its description logic underpinning. The logical definition of concepts makes it possible to compute concept equivalence and subsumption, among other things. ICD-10-PCS is a compositional terminology which makes it inherently computable. This makes it possible, at least in theory, to develop a computational way to link the two terminologies.\(^{17,18}\) For this to work, we have to assume that the SNOMED CT logical concept model is able to represent the meaning of ICD-10-PCS procedures and that existing concepts are modeled sufficiently for this purpose. This study examines these assumptions and identifies some hurdles which need to be overcome.
Improving SNOMED CT logical definitions and concept model

Currently, only 42% of all SNOMED CT procedure concepts are designated as fully-defined. This means that in about 60% of procedures, some essential defining attributes are missing and that the meaning of the concept is not fully reflected by the logical definition. Increasing the proportion of fully-defined concepts will improve the capability of SNOMED CT to support computation. The proportion of fully-defined concepts has been recognized by SNOMED International as an important quality metric for SNOMED CT. It is a daunting task to completely model tens of thousands of concepts and some prioritization is necessary. In this study, we have identified some commonly performed procedures in clinical practice which would be a good place to start. In our opinion, the highest priority should be given to those concepts that match an ICD-10-PCS procedure completely in meaning but the logical definition is incomplete because of the omission of some clinically significant attributes (about 7% of cases in our study). These concepts already exist in SNOMED CT and additional work is needed to make their definitions more complete. We are proposing that only attributes which are clinically meaningful and significant to be added, because omitting them can lead to ambiguity in the clinical context. Improving the logical definition of these concepts will not only facilitate mapping to ICD-10-PCS, but will avoid potentially erroneous results in description logic computation e.g., for data aggregation and inferencing. Next-in-line would be the cases with partial meaning matches which can be fully represented by the existing concept model. This represents a larger number of cases (56% of cases in our study) and will require significantly more effort. To match the full meaning of these ICD-10-PCS procedures, new concepts that are sufficiently modeled will need to be added to SNOMED CT. Can we justify the effort? The answer depends on whether adding the concepts will provide high enough utility to SNOMED CT users. We would argue that for procedures that are indeed commonly performed but missing in SNOMED CT, the effort will be worthwhile. Our study data set is based on a single hospital and may not be representative enough. To identify the list of priority concepts to work on, SNOMED International will need to gather more data sets from representative sources.

Cases that cannot be fully represented with the existing concept model present another challenge. We have identified three kinds of reasons for failure. In a small number of cases in our study, full logical representation cannot be achieved because of the problem of cardinality. The current SNOMED CT description logic profile cannot specify the number of occurrence of an attribute. Adding that capability to the SNOMED CT logic profile would be a significant change. We do not think that the ICD-10-PCS mapping use case alone is sufficient to justify this. The other cases in our study involve missing body structure or method concepts. Adding new concepts (provided that they are compatible with SNOMED CT editorial principles) generally requires less effort and causes less disruption. For example, adding the body structure concept Part of greater omentum will enable the full representation of the ICD-10-PCS procedure Excision of Greater Omentum, Open Approach (0DBS0ZZ). Adding the method Sleeve resection will allow the concept Laparoscopic sleeve gastrectomy (427074001) to become fully defined.

Implications for logical mapping

Our study can shed light on the feasibility and expected yield of using ontological alignment to map between SNOMED CT and ICD-10-PCS. In general, the SNOMED CT procedure model aligns reasonably well with the ICD-10-PCS axes. With the existing SNOMED CT concept model and existing attributes and values, 86% of ICD-10-PCS procedures can be fully represented as a SNOMED CT expression. By adding some new body part and method concepts, the coverage can go even higher. This speaks to the overall feasibility of logical mapping.

However, based on the existing modeled content in SNOMED CT, only a small proportion of cases can be mapped completely using the logical definition. This includes the cases we identified as full and close logical definition matches (25% of cases in our study). The ideal scenario is a full meaning and definition match, but it only occurs in one case. The cases with exact meaning match and close definition match are also amenable to logical matching. Since the missing attributes in these cases are clinically trivial, they can be implied and added automatically based on other characteristics of the procedure. For example, Access = Percutaneous approach can be added to all laparoscopic procedures, and Has intent = Diagnostic intent can be added to all biopsies. This will then allow fully-automated logical mapping to ICD-10-PCS procedures.
To increase the yield of logical mapping for the other cases will be more difficult. One solution would be for SNOMED International to improve the coverage and enrich the logical content of commonly performed procedures as discussed above. Whether this will occur depends on the availability of resources among various competing priorities.

So far, our discussion has focused on full logical equivalence between a SNOMED CT concept and an ICD-10-PCS procedure. However, even with the existing SNOMED CT content, partial logical matches can be found for most cases. Usually, a broader SNOMED CT concept can be found matching some (but not all) of the ICD-10-PCS components. For example, the ICD-10-PCS procedure *Resection of Spleen, Open Approach (07TP0ZZ)* can be matched through partial logical matching to *Total splenectomy (174776001)* in SNOMED CT through the attributes *Procedure site – direct = Entire spleen* and *Method = Excision – action*. The only missing attribute is the open approach. From a mapping perspective, these partial matches can be useful. One potential use case will be to suggest candidate mappings to human reviewers to expedite the creation of a manually-validated map. Another possible use case is to support real-time interactive post-coordination by clinical users. Starting with a partial match, the system can prompt the user to provide additional information until a full match is achieved. For example, when the user types in ‘splenectomy’, a list of SNOMED CT concepts will appear in a pick list. When the user picks ‘total splenectomy’, the system will further prompt the user to pick between open and laparoscopic approach. This is because based on the partial logical matching, *Total splenectomy (174776001)* is mapped to two ICD-10-PCS procedures: *Resection of Spleen, Open Approach (07TP0ZZ)* and *Resection of Spleen, Percutaneous Endoscopic Approach (07TP4ZZ)* and the user can click on the appropriate choice to hone in on the optimal ICD-10-PCS code. Furthermore, the system can also add either the attribute *Access = Open approach - access* or *Using access device = Laparoscope, device* to the logical definition of the concept *Total splenectomy*, thus creating a post-coordinated expression that fully captures the meaning of the chosen ICD-10-PCS procedure. In addition to cases with missing attributes, partial logical matching can also handle cases in which an exact match in the value of an attribute cannot be found. For example, an ICD-10-PCS procedure on the left jugular vein can be partially matched to a SNOMED CT procedure on the jugular vein. At the time of data entry, the user picking this SNOMED CT concept will be prompted to specify the laterality, which can also be captured in as a post-coordinated expression. One important consideration when using real-time post-coordination for data capture is to minimize the number of clicks needed to find the optimal code. We are building an experimental tool to study the different ways of presenting and prompting for additional information.

We recognize the following limitations in our study. We only focused on surgical operations in ICD-10-PCS and SNOMED CT but the scope of the two systems includes other types of medical procedures (e.g., obstetrical procedures, imaging studies). Our study sample was based on data from one hospital and the results may not be generalizable.

**Conclusion**

Logical mapping between SNOMED CT and ICD-10-PCS by aligning the SNOMED CT defining attributes and ICD-10-PCS axes is feasible. The biggest gap is in the surgical approach which is specified in all ICD-10-PCS codes but only in 8.7% of SNOMED CT concepts. However, in some cases the missing attributes are not clinically ambiguous and can be implied. In 25% of ICD-10-PCS codes, there is exact meaning and close logical match to a pre-coordinated SNOMED CT concept. Overall, 86% of ICD-10-PCS codes can be represented completely by the existing SNOMED CT concept model.

**Acknowledgements**

The authors thank Dr. James Campbell for providing the ICD-10-PCS data set from the Nebraska Medical Center, Omaha, Nebraska. The source of the data set was supported by the Research Information Technology Office and grant PCORI CDRN-1501-26643 CDRN 04: The Greater Plains Collaborative This work was supported in part by the Intramural Research Program of the National Institutes of Health and the National Library of Medicine.
References
1. SNOMED International Member Countries. Available from: http://www.snomed.org/members/.
8. SNOMED CT to ICD-10-PCS Mapping Project Group. Available from: https://confluence.ihtsdotools.org/display/SCTICD10PCS.
15. SNOMED CT Procedure Concept Model. Available from: https://confluence.ihtsdotools.org/display/DOCEG/6.2.2+Attributes+used+to+define+Procedure+concepts.
Toward Automated Pre-Biopsy Thyroid Cancer Risk Estimation in Ultrasound

Alfiia Galimzianova, PhD, Sean M. Siebert, BS, Aya Kamaya, MD, Terry S. Desser, MD*, Daniel L. Rubin, MD, MS*
Stanford University School of Medicine, Stanford, CA, USA

Abstract

We propose a computational framework for automated cancer risk estimation of thyroid nodules visualized in ultrasound (US) images. Our framework estimates the probability of nodule malignancy using random forests on a rich set of computational features. An expert radiologist annotated thyroid nodules in 93 biopsy-confirmed patients using semantic image descriptors derived from standardized lexicon. On our dataset, the AUC of the proposed method was 0.70, which was comparable to five baseline expert annotation-based classifiers with AUC values from 0.72 to 0.81. Moreover, the use of the framework for decision making on nodule biopsy could have spared five out of 46 benign nodule biopsies at no cost to the health of patients with malignancies. Our results confirm the feasibility of computer-aided tools for noninvasive malignancy risk estimation in patients with thyroid nodules that could help to decrease the number of unnecessary biopsies and surgeries.

Introduction

The incidence of thyroid nodules, both benign and malignant, has been consistently increasing in the United States in the recent decades. Much of this increase, it is believed, is due to increased utilization of imaging, with concomitant increased detection of asymptomatic thyroid nodules. In fact, autopsy studies report that up to 50-67% of adults nationwide are expected to have thyroid nodules, while only 0.2% of the population is reported to have thyroid cancer. Current definitive diagnosis of thyroid nodules requires tissue biopsy or even surgery, while only 5-7% of these nodules are found to be malignant. This inevitably exposes the majority of the patients to unnecessary health risks associated with these invasive tests and increases societal healthcare costs substantially. Therefore, there is a critical need for methods to reliably estimate the malignancy risks of thyroid nodules to decrease the number of invasive interventions being performed in low-risk benign nodules. Intensive research has been underway in the radiology, endocrinology, and surgery communities to attempt to identify those patients at high risk and who merit invasive diagnostic intervention, yet despite this, accurate diagnosis based on imaging findings remains very challenging.

Ultrasound imaging (US) is the standard-of-care imaging modality used to visually assess the risks of malignancy, define the necessity of biopsy-based definitive diagnosis and guide the fine-needle aspiration biopsy. In its recent guidelines for 2015, the American Thyroid Association recommends biopsy for thyroid nodules at size thresholds specified for five nodule appearance patterns associated with different risks of malignancy. However, such pattern-oriented approaches are not collectively exhaustive descriptors of nodule appearance in US and thus are not able to provide any scores to some nodules. Therefore, there is a need for methods that provide comprehensive evaluation of thyroid nodules based on collectively exhaustive sets of US features. Several Thyroid Imaging Reporting and Data System (TIRADS) classification systems have been proposed over the last years with the aim of providing a systematic approach evaluating the cancer risks of thyroid nodules based on multiple US features. The newly developed TIRADS lexicon from the American College of Radiology (ACR) provides a set of standardized US imaging descriptors (visual semantic features) used in expert annotation-based scoring systems (TIRADS classifiers) to estimate the risk that a nodule is a cancer before biopsy. However, extraction of these qualitative features imposes additional burden on radiologists and is subject to intra- and interrater variability. Moreover, the scoring systems for nodule malignancy are often oversimplified due to their design for human use. Automated analysis of nodule properties by extraction of computational features from US images and estimation of the risks using machine learning based classifiers could help reduce the expert labor and eliminate the associated variability.

*These authors contributed equally to this work
In this work, we present a computational framework for automated cancer risk estimation from US images, which is based on analysis of a rich set of computational features using a random forest classifier (RF). Unlike existing TIRADS classifiers, the proposed framework does not require qualitative assessments from the expert radiologist. We validated the framework on a dataset of US images of cases with biopsy-proven diagnosis. Our proposed framework compares favorably to other methods and has potential value for computer-aided diagnosis to help decrease the high number of biopsies in patients ultimately found to have benign disease, while identifying those patients at risk of thyroid cancer.

Materials and Methods

Datasets

Ultrasound imaging data of patients that underwent thyroid nodule examination at Stanford Hospital from year 2010 to 2015 were collected retrospectively with the approval by the institutional review board (IRB). Our final dataset consisted of 47 malignant and 46 benign biopsy-confirmed nodules from 93 patients (74 females, 19 males; mean age 55.9±15.4 years). For each nodule, its principal transverse and longitudinal projection US images were included. Blinded to the diagnosis, an expert radiologist reviewed the images, outlined each nodule, and recorded descriptors of visual characteristics of nodules using the recently established ACR TIRADS consensus lexicon8. Viewing, outlining and annotation of the images was performed in the electronic Physician Annotation Device (ePAD)14 (Figure 1). The images were preprocessed to reduce the US-specific artifacts such as speckle using nonlocal means based speckle filtering15. The advantage of this filtering method over traditional smoothing is its ability to preserve the edges, which is crucial for reliable extraction of computational features.

Semantic features for radiologist-based malignancy risk estimation

We derive a set of ten semantic features directly from the radiologist’s annotations (Table 1). As reference malignancy risk estimators, we use existing expert annotation-based TIRADS classifiers. Using the ACR TIRADS lexicon, we reformulate and implement five such TIRADS classifiers as described in the works of Park et al.9, Kwak et al. in 201110, Kwak et al. in 201311, Zayadeen et al.12, and Russ et al.13. In their work, Park et al.9 used 12 US features to assess the risk of malignancy using linear regression, and define six TIRADS categories based on the estimated risk. Kwak et al. in 201110 distinguished eight US features suspicious of malignancy, which were incorporated into a logistic regression-based malignancy risk quantification and the count-based classification system. Kwak et al. in 201311 used five US features of malignancy to build their weighted sum based TIRADS scoring system. Zayadeen et al.12 proposed a scoring system based on the defined major and minor malignant features, and also benign features. Russ et al.13 describes French TIRADS, a five-tier scoring system based on five malignant and six benign features. ACR TIRADS lexicon provides terms sufficient to describe these major US features of thyroid nodules and thus all five TIRADS classifiers were implemented using our expert annotations.

<table>
<thead>
<tr>
<th>Semantic feature</th>
<th>Possible values of radiologist annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Composition</td>
<td>Solid</td>
</tr>
<tr>
<td></td>
<td>Predominantly solid</td>
</tr>
<tr>
<td></td>
<td>Predominantly cystic</td>
</tr>
<tr>
<td></td>
<td>Cystic</td>
</tr>
<tr>
<td></td>
<td>Spongiform</td>
</tr>
<tr>
<td>Echogenicity</td>
<td>Hyperechoic</td>
</tr>
<tr>
<td></td>
<td>Isoechoic</td>
</tr>
<tr>
<td></td>
<td>Hypoechoic</td>
</tr>
<tr>
<td></td>
<td>Very hypoechoic</td>
</tr>
<tr>
<td>Shape</td>
<td>Taller-than-wide</td>
</tr>
<tr>
<td></td>
<td>Wider-than-tall</td>
</tr>
<tr>
<td>Border</td>
<td>Smooth</td>
</tr>
<tr>
<td></td>
<td>Irregular</td>
</tr>
<tr>
<td></td>
<td>Lobulated</td>
</tr>
<tr>
<td></td>
<td>Ill-defined</td>
</tr>
<tr>
<td>Halo</td>
<td>Present</td>
</tr>
<tr>
<td></td>
<td>Absent</td>
</tr>
<tr>
<td>Extrathyroidal extension</td>
<td>Present</td>
</tr>
<tr>
<td></td>
<td>Absent</td>
</tr>
<tr>
<td>Punctate echogenic foci</td>
<td>Present</td>
</tr>
<tr>
<td></td>
<td>Absent</td>
</tr>
<tr>
<td>Macroc calcifications</td>
<td>Present</td>
</tr>
<tr>
<td></td>
<td>Absent</td>
</tr>
<tr>
<td>Peripheral calcifications</td>
<td>Present</td>
</tr>
<tr>
<td></td>
<td>Absent</td>
</tr>
<tr>
<td>Comet-tail artifacts</td>
<td>Present</td>
</tr>
<tr>
<td></td>
<td>Absent</td>
</tr>
</tbody>
</table>
Computational features for automated malignancy risk estimation

To describe appearance of nodules identified by the radiologist in a quantitative manner, we selected a rich set of computational features that encode the echogenicity, texture, margin and shape properties of thyroid nodules. Nodule echogenicity was expressed by nodule intensity features, i.e., intensity histogram and intensity statistics. Nodule composition and presence of echogenic foci was expressed by texture and intensity features: rotation-invariant local binary pattern, Gabor filter bank, gray level co-occurrence matrix based features, and intensity difference between the nodule and its surrounding tissue. Nodule margin was computationally characterized via edge sharpness and local area integral invariant descriptor statistics at multiple spatial scales. Nodule shape had the following computational counterparts: ellipsoid fit, compactness, and roughness features. Nodule halo and peripheral calcifications were quantified by local intensity features, specifically by intensity histogram over the nodule margin. Although some of the semantic features, such as extrathyroidal extension, cannot be characterized easily from nodule outline alone, the edge sharpness features can be used as a surrogate, since the edges are expected to be blurred at the boundary portion where the thyroid capsule is invaded by the nodule\textsuperscript{16}. Overall, 480 computational features were collected.

Classification

We formulate the automated malignancy risk estimation as a posterior probability of malignancy learned by a machine learning classifier on an annotated training dataset. For both radiologist-annotated semantic features and computer-derived computational features, we trained an RF classifier to predict whether the nodule is benign or malignant. Classification by RFs consists of creating an ensemble of decision tree (DT) classifiers, each learned on a subsample of the original training dataset, and predicting the outcome by averaging responses across the DTs. The subsamples are drawn with replacement and also consist of randomly selected features. The RFs are characterized by the number of grown trees and the depth of the trees (equivalent to the number of leaves) in the DTs.

We chose RF because it generally has superior performance in medical image analysis\textsuperscript{17} and bioinformatics research\textsuperscript{18}, and it also has several additional desirable properties: 1) RFs can handle situations when the number of features are considerably higher than the number of data samples and do not require explicit dimensionality reduction as some other methods, 2) they can work with both categorical and continuous features and their combinations, and 3) they provide a probabilistic output. Therefore, in formulation of RF classification as discriminating benign vs malignant nodules, we will use the malignancy probability as the risk estimate.

\textbf{Figure 1} Example of annotation of a thyroid nodule on a transverse US image using the ePAD image viewing and annotation tool.
Evaluation

We used five expert scoring systems, or TIRADS classifiers, as reference malignancy risk estimators. The annotations provided by the radiologist were transformed into a mineable feature set consisting of binary and categorical predictors. Overall, 10 semantic and 480 computational features were used to train RF classifiers (RF-Semantic and RF-Computational). Leave-one-out hyperparameter tuning and model fitting was performed. Within each fold, the minimal number of the leaves in the decision trees were selected by searching among values \{2,5,10,20\}, while the number of trees in the forest was fixed to value 50 across all folds. The RFs were trained on the biopsy-confirmed diagnoses of the nodules in two-label formulation, i.e., benign vs malignant.

We measured the predictive performance of the classifiers using receiver operating characteristic (ROC) curves built on the malignancy scores as predictions and nodule malignancy indicator as reference labels. As a quantitative evaluation metric, we used the corresponding values of area under the ROC curve (AUC), which is an estimate of the probability that, for the randomly selected malignant nodule and randomly selected benign nodule, their scores will be in the correct order, i.e., the malignant nodule will receive a score higher than the benign nodule.

Results

The performance summary of the five TIRADS and the two RF classifiers is demonstrated in Figure 2 using ROC curves. For TIRADS classifiers, performance in terms of AUC was AUC=0.73 for Park et al.\(^9\), AUC=0.72 for Kwak et al.\(^{2011}\), AUC=0.81 for Kwak et al.\(^{2013}\), AUC=0.76 for Zayadeen et al.\(^{12}\), and AUC=0.77 for Russ et al.\(^{13}\). The proposed computational framework (RF-Computational) achieved AUC=0.70 and its semantic counterpart (RF-Semantic) provided AUC=0.75 (Figure 2). As expected, the classifiers that rely on radiologist annotation performed better than RF based on computational features alone, however, the performance was still comparable, which can also be noticed in Figure 2 as the ROC curves are at a close proximity to each other.

To study the clinical relevance of the methods, we analyzed their behavior at the highest true positive rate, or sensitivity. We recorded the number of correctly identified benign nodules, i.e., true negative count, in the absence of misidentified malignant nodules, i.e. false negative count close to zero. An ideal classifier would correctly identify all benign nodules without missing any malignant nodules, thus resulting in the number of true negatives equal to the number of benign nodules while having false negative count as zero. The results on our dataset for all the implemented classifiers are summarized in Table 2. Although at non-zero missed malignancies (FN=1,2 or above) the TIRADS classifiers achieved higher number for spared benignities, implementation of such rules in practice
would result in a decrease of benign nodule biopsies at the cost of misdiagnosis of patients with cancer (FN>0). Only the RF classifiers were able to provide a positive number of true negatives with no false negatives, which is the number of biopsies that could be avoided at no cost to patients with malignancies.

**Discussion**

In this work, we presented a computational framework for evaluating thyroid nodules (benign vs. malignant) on US images, with performance comparable to that of expert annotation-based classification systems. The malignancy estimation performance with respect to the biopsy-confirmed diagnosis for our framework was 0.70 in terms of AUC, which was lower but comparable to performance of the classifier based on expert-annotated semantic features (AUC 0.75) and to values obtained by the expert annotation-based TIRADS classifiers (AUC ranged from 0.72 to 0.81). Comparability in terms of classification performance shows potential of such computational features to enhance or perhaps even replace annotation by radiologists, which is laborious, expensive and prone to intra- and interrater variability.

The clinical relevance of the classifiers as a potential pre-biopsy malignancy risk estimator was further analyzed in terms of the number of unnecessary biopsies that could have been avoided, i.e., nodules correctly identified as benign, at the minimal levels of misidentified malignancies. In our experiments, the proposed computational framework achieved the highest number of correctly identified benign nodules (5 out of 46) in the absence of missed malignancies (values in italics in Table 2). Although the TIRADS systems we compared our system against could identify more benign cases, they did so with more misidentified malignancies. Another disadvantage of TIRADS classifiers in this respect is the limited control over the threshold for benign nodule selection. This is due to the oversimplified rules that lay foundation to such qualitative classification systems and resulted in small number of possible thresholds, e.g. five nodule malignancy scores in the work by Park et al. As can be seen in Table 2, only the RF classifiers provided thresholds for all the analyzed missed malignancy (false negative) counts.

Although some prior work has been done in thyroid cancer diagnosis using computational features derived from ultrasound images, the studies were limited to considerably smaller datasets (20 patients). More recently, Wu et al. have shown that their framework of radiologist-annotated semantic features and machine learning classifiers can provide results comparable to that of expert malignancy scoring. However, to the best of our knowledge, our work is the first to demonstrate a framework of computational features and a machine learning classifier having a performance comparable to expert classification. Our computational framework could serve as a basis to develop computer-aided diagnosis tools that ultimately could help reduce the high number of unnecessary biopsies in patients with benign nodules.

Thyroid cancer diagnosis currently relies on cytopathological analysis, which is invasive and carries risks of bleeding, infection and potential damage to adjacent structures. Identifying patients who have low risk for cancer is important to avoid such unnecessary health risks and decrease societal healthcare costs. Recent attempts to address the issue by introducing more reliable pre-biopsy cancer risk estimation using standard-of-care US images are based on radiologist-
annotated standard descriptors, which are laborious to collect and subject to inter- and interrater variability\textsuperscript{21}. Ultrasound is invaluable to noninvasive characterization of thyroid nodules; however, current guidelines are limited to recording a limited number of US feature of the nodules\textsuperscript{8}. By computing a rich set of features from these images, our proposed framework maximizes the use of the information available in the images. The use of the computer-aided diagnosis tools that are based on frameworks similar to ours could improve the management of thyroid nodules and result in decreasing the number of unnecessary biopsy or surgical risks, while more appropriately directing care in patients who actually need more invasive management.

A limitation of this study was the use of only two principal projections, transverse and longitudinal, for extraction and analysis of semantic and computational features. Although the selection of these projections is standard in practice to evaluate nodule properties such as shape and size, the use of multiple image frames from the ultrasound exams could better other US features and will be analyzed in future work. Obtaining consistent US features is also challenged by the variability in positioning the transducer by the operator; an extended study with multiple operators to study inter- and intra-rater variability can give more insight on robustness of certain features and limitations of others. An additional limitation of our work was the use of a single radiologist for determining the semantic features, though that radiologist is an academic expert. Future studies assessing the impact of inter-reader variation in assessments among experts on our results could be helpful.

**Conclusion**

In this paper, we presented a computational framework for estimating the risk of cancer in thyroid nodules by analysis of US images. The framework computes a rich set of computation features in the US images and estimates the malignancy probability using random forest classifier. Performance of the framework in terms of AUC was lower but comparable to that of five baseline expert radiologist annotation-based TIRADS classifiers. Given that our framework does not require qualitative assessments from the expert radiologist, it could reduce the effort required to evaluate these images and reduce variation in practice. In addition, the number of biopsies that a classifier could help to avoid without missing any malignancies was the highest for our framework. Our results confirm the feasibility of computer-aided diagnosis systems for thyroid cancer risk estimation. Such systems could provide second-opinion malignancy risk estimation for clinicians and ultimately help decrease the number of unnecessary biopsies and surgeries.

**Acknowledgements**

This research was supported in part by grants from GE Medical Systems (Blue Sky Initiative at Stanford University) and the National Cancer Institute, National Institutes of Health, U01CA142555, 1U01CA190214, and 1U01CA187947.

**References**


6. Haugen BR, Alexander EK, Bible KC, Doherty GM, Mandel SJ, Nikiforov YE, et al. 2015 American Thyroid Association Management Guidelines for Adult Patients with Thyroid Nodules and Differentiated Thyroid


**Effectiveness of a Cloud-Based EHR Clinical Decision Support Program for Body Mass Index (BMI) Screening and Follow-up**

Shruti Gangadhar, MS¹, Nam Nguyen, MS¹, James W. Pesuit, BA¹, Alina N. Bogdanov, MA¹, Lee Kallenbach, PhD¹, Jessica Ken, MPH¹, Joe Vasey, PhD¹, Richard M. Loomis, MD¹

¹Practice Fusion, San Francisco, CA

**Abstract**

The effectiveness of a clinical decision support (CDS) program encouraging clinicians to record patient’s Body Mass Index (BMI) and document appropriate follow-up plans is evaluated. Test (4,987 practices, 33,445 clinicians) and control groups (881 practices, 6,316 clinicians) were selected using stratified random sampling. Three CDS alerts for BMI screening and follow-up based on evidence based clinical quality guidelines were displayed at the point of care in a cloud-based EHR. The effectiveness of the CDS program was measured over 4 months by tracking recorded BMI and documented follow-up plans. Over the program, BMI recording increased minimally and documentation of follow-up plans increased 5-fold (p=0.05) compared to the control group. The overweight test group patients (18-64yo) gained less weight (p=0.06) than the control group and underweight patients gained more weight (p<0.01) during the program period. Outcome studies with longer follow-up periods are needed to further confirm positive outcomes.

**Introduction**

According to the Centers for Disease Control and Prevention (CDC), more than one-third (36.5%) of U.S. adults are obese¹. Obesity prevalence in the United States (U.S.) is on the rise and is expected to reach 42% of the population by 2030². Obesity is a risk factor for conditions including heart disease, stroke, type 2 diabetes, and certain types of cancer, some of the leading causes of preventable death²,³. Various institutions including the National Institutes of Health (NIH) and CDC have provided guidelines to help classify weight based on Body Mass Index (BMI)⁴,⁵. Anyone with a BMI between 25 and 29.9 would be classified as overweight and anyone with a BMI over 30 would be classified as obese⁴. In 2003, the U.S. Preventive Services Task Force recommended that primary care practitioners (PCPs) screen all adults for obesity and offer behavioral interventions and intensive counseling to affected individuals⁶. However, practice guidelines alone have failed to elicit appropriate clinician behavior⁷. Studies have shown that CDS systems improve clinical practice in preventive care and computer-assisted drug ordering and dosing systems⁸. While various studies have been done to test the effectiveness of CDS in management of high blood pressure, diabetes care, and asthma care⁸,⁹, there have been a limited number of studies that evaluate the effectiveness of a weight management CDS on clinician behavior and patient outcomes¹⁰.

A manufacturer of a chronic weight management treatment sponsored this CDS program for BMI screening and follow-up on the Practice Fusion EHR platform (PF-EHR). Our goals were to first measure the effectiveness of the CDS programs and second to find any evidence of improvements in patient outcomes associated with the CDS programs. The PF-EHR is cloud-based with 30,000 active practices in all 50 states facilitating over 5M office visits per month representing 6.7% of US ambulatory care. This study evaluates the impact of the weight management CDS program on clinician behavior and patient outcomes.

**BMI Clinical Decision Support Program Development**

For this study, three computer-based CDS alerts were implemented based on evidence-based clinical guidelines and quality measures¹¹. The CDS logic implemented is represented below (Figure 1).
Figure 1. CDS Alert logic

The first alert ("BMI not recorded") prompted clinicians to record BMI for patients age 18 years or older who did not have a recorded BMI in the prior 6 months (Figure 2).

Figure 2. Screenshot of alert: BMI not recorded.

For patients who did have a BMI recorded during the prior 6 months, the second alert ("Overweight, document follow-up plan") identified patients aged 18-64 whose most recent BMI recording during the prior 6 months was >= 25 and patients of age>65 years with BMI >=30 and prompted clinicians to document a weight loss plan (Figure 3).

Figure 3. Screenshot of alert: Overweight, document follow-up plan

The third alert ("Underweight, document follow-up plan") identified patients aged 18-64 whose most recent BMI recording during the prior 6 months was <18.5 and patients of age>65 years with BMI <23 and prompted clinicians to document a weight gain plan (Figure 4).
In the PF-EHR, BMI is auto-calculated when height and weight are documented in the vitals area of the patient chart, so for this study, BMI documentation equates to documentation of height and weight. In addition, documentation of a BMI related follow-up plan was counted when a clinician indicated in the assessment area of the patient chart that they performed one of two follow-up plans: “Calculated BMI above normal parameters and a follow-up plan was documented” or “Calculated BMI below normal parameters and a follow-up plan was documented”. The list of follow-up plans that clinicians could choose from were BMI follow-up plans available in the PF-EHR as a structured data associated to SNOMED concepts: weight controlled education (procedure), nutrition counseling, nutrition supplement (dietary) therapy, peripherally acting anti-obesity drug, counseling about physical activity (exercise), referral to weight maintenance regimen service, bariatric operative procedure (surgery), and next appointment.

The CDS logic for all three alerts excluded pregnant women, patients who received palliative care, refused BMI measurement, were in an urgent or emergent medical situation where time was of the essence and to delay treatment would jeopardize the patient’s health status, or any other reason documented in the medical record by the clinician explaining why BMI measurement was not appropriate during the timeframe.

**Measuring Effectiveness**

Test (4,987 practices, 33,445 clinicians) and control groups (881 practices, 6,316 clinicians) were selected prior to the program using a stratified random sampling method at the practice level. Eligible practices (5,871) were chosen based on history of patients meeting the CDS alert criteria (previous patients with BMI recorded and previous diagnosis of obesity). Practices were then stratified on all combinations of the following: tertiles of number of active patients per practice, tertiles of total visits by patients, tertiles of BMI values, tertiles of counts of BMI Screening/Intervention/Assessments, tertiles of visits by patients diagnosed with obesity, tertiles of number of MDs and NPs per practice and specialties (PCPs (Family Medicine, Internal Medicine, and General Practice), Obstetrics and Gynecology, Endocrinology), geographic region (West, Southwest, Midwest, Southeast, and Northeast). 15% of practices from each stratum were randomly selected for inclusion into the control group (strata combinations with less than 5 practices were combined and randomly sampled). The remaining practices comprised the test group.

Non-parametric Mann-Whitney tests were performed for continuous parameters and Chi-squared tests for categorical parameters to evaluate test and control group similarity. High p-values indicated that the control group adequately represented the test group (total visits by patients with a BMI captured (p=0.99), count of BMI follow-up plans documented (p=0.95), patients aged 65+ with low BMI (p=0.94), patients aged 65+ with normal BMI (p=0.88), visits by patients diagnosed with obesity (p=0.58), specialty (p=0.72), number of MD/NP clinicians per practice (p=0.99), number of active patients per practice (p=0.86)).

Test and control practices were monitored through a baseline period from 8/1/2014 to 11/30/2014 and through the CDS alerts program period from 12/1/2014 to 3/31/2015. CDS alerts were not shown to the control group, however visits where the CDS criteria were met were recorded for comparison. Percentage of visits for BMI recorded for visits where alerts shown (test group)/alerts should have shown (control group) were calculated and compared for the program period. The study did not continue past 4 months due to changes of the BMI documentation user-interface on the PF-EHR platform which changed the conditions of the study.

**Results**

During the 4 month program period, of the 39,761 clinicians in our program (33,445 test, 6,316 control), 11,123 clinicians were alerted. The alerted clinicians saw 1,154,304 distinct patients (981,155 test, 173,149 control) over 3,226,031 visits (2,766,603 test, 459,428 control). "BMI not recorded" was displayed to 8,856 clinicians and 285,947 patients, over 779,508 visits. Overall, percentage of visits with recorded BMI were equivalent between the test and control group (53.0% vs 53.6%, p=0.36, Figure 5).
However, for the subset of visits that were shown (test group) or were eligible for (control group) the "BMI not recorded" alert, there was a significantly higher percentage of visits with recorded BMI in the test group than the control group (27.9% vs 12.1%, p < 0.01, Figure 6).

"Overweight, document follow-up plan" was displayed to 10,018 clinicians, attributed to 837,721 patients over 2,713,986 visits. 9.8% of alerted patients (82,170 patients) had a follow-up plan documented in the program period. Documentation of follow-up plan for overweight patients increased by a factor of 5 for most of the program period compared to the baseline period in the test group (Figure 7). A smaller factor increase was observed in the test group between 12/11/14 and 1/11/15 attributed with a decrease of office visits during the winter holiday. The control group had no change BMI documentation of follow-up plan for BMI between the baseline and program period.
“Underweight, document follow-up plan” was displayed to 7,036 clinicians, attributed to 92,309 patients over 311,382 visits. 7.7% of alerted patients (7160 patients) had a follow-up plan documented in the program period. Documentation of follow-up plan for underweight patients increased by a factor of 5 for most of the program period compared to the baseline period in the test group (Figure 8). A smaller factor increase was observed in the test group between 12/11/14 and 1/11/15 attributed with a decrease of office visits during the winter holiday. The control group had no change BMI documentation of follow-up plan for BMI between the baseline and program period for the control group.

Figure 8. Percent of visits with follow-up plan documented by week alerted with "Underweight, document follow-up plan"
Overall percent of visits with documented follow-up plans for abnormal BMI increased in the test group in the program period compared to the baseline period (1.24% vs 0.29%, p<0.01). The control group showed no difference in percent of visits with documented follow-up plans (0.26% vs 0.22%, p=0.80, Figure 9).

**Figure 9.** Percent of visits with documented BMI follow-up plan by week.

Weight control education (procedure), nutrition counseling, and nutrition supplement (dietary) therapy were the top three Follow-up plans documented (Table 1).

**Table 1.** Type of Follow-Up Plans documented for Abnormal BMI.

<table>
<thead>
<tr>
<th>Follow-Up Plans</th>
<th>Control Group % (4 months)</th>
<th>Test Group % (4 months)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weight control education (procedure)</td>
<td>52.0%</td>
<td>61.6%</td>
</tr>
<tr>
<td>Nutrition counseling</td>
<td>40.7%</td>
<td>27.6%</td>
</tr>
<tr>
<td>Nutrition supplement (dietary) therapy</td>
<td>5.4%</td>
<td>7.2%</td>
</tr>
<tr>
<td>Peripherally acting anti-obesity drug</td>
<td>1.1%</td>
<td>2.2%</td>
</tr>
<tr>
<td>Counseling about physical activity (exercise)</td>
<td>0.1%</td>
<td>0.6%</td>
</tr>
<tr>
<td>Referral to weight maintenance regimen service</td>
<td>0.4%</td>
<td>0.5%</td>
</tr>
<tr>
<td>Bariatric operative procedure (surgery)</td>
<td>0.2%</td>
<td>0.4%</td>
</tr>
<tr>
<td>Next Appointment</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
</tbody>
</table>
Patient Outcomes:

Differences in patient BMI recorded at the start of the program and the end of the program were examined for patients with visits that were alerted and that had at least one BMI recorded between 12/1/2014 and 12/31/2014 and at least one BMI recorded between 2/1/2015 and 3/31/2015. The first BMI recorded between 12/1/2014 and the last BMI recorded between 2/1/2014 and 3/31/2015 were compared.

Table 2 compares the impact of follow-up plans during the four-month follow-up period between the test and control groups. While the patients (18-64 years) in the test group did not lose weight, there is weak evidence that they gained less weight during the follow-up period than the control group (p=0.06). There was no significant difference in weight change for patients 65+ (p=0.40). While underweight patients (18-64 years and 65+) gained significantly more weight than those in the control group (p<0.01, p=0.02).

Table 2. Percent changes in BMI for patients that had a recorded BMI at the beginning and end of the program:

<table>
<thead>
<tr>
<th>Patient Category</th>
<th>% Weight Change in Test Group</th>
<th>% Weight Change in Control Group</th>
<th>p-values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overweight patients aged 18-64 years</td>
<td>+0.02% SD=4.93</td>
<td>+0.32% SD=4.11</td>
<td>P=0.06</td>
</tr>
<tr>
<td>Underweight patients aged 18-64 years</td>
<td>+3.98% SD=17.85</td>
<td>-0.82% SD=3.12</td>
<td>P&lt;0.01</td>
</tr>
<tr>
<td>Overweight patients aged 65 years and older</td>
<td>-0.41% SD=4.43</td>
<td>-0.49 SD=4.39</td>
<td>P=0.40</td>
</tr>
<tr>
<td>Underweight patients aged 65 years and older</td>
<td>+0.79% SD=8.7</td>
<td>-2.79%SD=8.56</td>
<td>P=0.02</td>
</tr>
</tbody>
</table>

Discussion

Though we observed a small increase in BMI recorded for visits that received the “BMI not recorded” alert, the overall percentage of visits with a documented BMI did not change significantly. This suggests that the “BMI not recorded” alert is not as effective at a population level. However, the significant increase in percentage of documentation of follow-up plans indicate that follow-up plan alerts were effective in altering clinician behavior. Our study had a few limitations. Firstly, the follow-up period was relatively short. Although the follow-up period was short, we did see improved patient outcomes. Further studies with a longer follow-up periods are needed to confirm the positive outcomes. Secondly, test and control groups were chosen based on practice criteria. Future studies controlling better for patient demographics characteristics (such as more granular geography, age, socioeconomic characteristics, genomic traits, and co-morbidities) would provide additional insight on patient outcomes.

Conclusion

CDS systems are increasingly being adopted by the medical community to help clinician compliance with evidence based guidelines and best practices with the intent of improving patient outcomes. This study has shown that CDS systems can be an effective tool to help clinicians follow weight management clinical guidelines and improve patient outcomes.

References

1. [https://www.cdc.gov/obesity/data/adult.html](https://www.cdc.gov/obesity/data/adult.html)
The “Safety Net” of Community Care: Leveraging GIS to Identify Geographic Access Barriers to Texas Family Planning Clinics for Homeless Women Veterans

Lori Gawron, MD, MPH\textsuperscript{1,2}, Warren B. P. Pettey, MPH, CPH\textsuperscript{1,3}, Andrew Redd, PhD\textsuperscript{1,3}, Ying Suo, MStat\textsuperscript{1,3}, David K. Turok, MD, MPH\textsuperscript{2}, Adi V. Gundlapalli, MD, PhD, MS\textsuperscript{1,3,4,5}

\textsuperscript{1}VA Salt Lake City Health Care System, Informatics, Decision-Enhancement and Analytic Sciences Center (IDEAS 2.0), Salt Lake City, UT; Departments of \textsuperscript{2}Obstetrics and Gynecology, \textsuperscript{3}Internal Medicine, and \textsuperscript{4}Biomedical Informatics, University of Utah School of Medicine, Salt Lake City, UT; \textsuperscript{5}US Department of Veterans Affairs National Center for Homelessness Among Veterans, Philadelphia, PA

Abstract
The Veterans Healthcare Administration (VHA) is developing a civilian referral system to address specialty access issues to VHA healthcare. Homeless women Veterans may not have the resources to navigate referral systems when travel to VHA Medical Centers (VAMCs) is limited, especially for family planning needs. Recent Texas legislation restricted funding to civilian, publically-funded family planning clinics, limiting comprehensive services. This study’s goal was to assess geographic availability of VAMCs and family planning clinics for homeless Texan women Veterans. We identified 3,246 Texan women Veterans, age 18-44y with administrative homelessness evidence anytime between 2002-2015. Significant clusters of homeless women Veterans were near VHA facilities, yet mean travel distance was 24.1 miles (range 0-239) to nearest family planning clinic compared to 82.6 miles (range 0.8-316.4) to nearest VAMC. Community clinics need ongoing civilian funding support if the VHA is to rely on their geographic availability as a safety net for vulnerable Veterans.

Introduction
Women Veterans are the fastest growing population of Veterans Health Administration (VHA) eligible users, nearly doubling in the past decade and numbering more than 2 million in 2015. \textsuperscript{1,2} Women Veterans who access care in the VHA are a vulnerable population with a high prevalence of medical and mental health comorbidities \textsuperscript{1,3}, military sexual trauma experiences \textsuperscript{4}, and have a 4-fold increased risk of homelessness compared to the civilian population \textsuperscript{5}. In order to meet the unique needs of this growing population, the VHA prioritized expansion of comprehensive primary care that includes reproductive health care by a single, women’s health provider.\textsuperscript{6} This comprehensive care should include family planning services but may require travel to an urban, Veterans Healthcare Administration Medical Center (VAMC) for more advanced services, such as insertion of an intrauterine device (IUD), due to provider availability or training.\textsuperscript{7} To avoid this travel barrier, the VHA is increasingly developing policies to partner with healthcare providers closer to Veterans’ homes.\textsuperscript{8} This community referral process (Veterans Choice Program) can be challenging for Veterans to utilize. They must first make contact with a VHA provider through a VAMC or Community-Based Outpatient Clinic (CBOC), obtain a referral for non-VHA or “Choice” care, wait for an authorization process, and then wait until an appointment is scheduled with an eligible civilian provider. Unfortunately, this practice only works if trained civilian providers are geographically accessible to meet the Veterans’ needs.

Homeless women are at increased need for family planning services, due to elevated risks for sexual trauma and infections (STIs), unintended pregnancy, and adverse outcomes, such as preterm birth.\textsuperscript{9,13} Women Veterans are at greater risk for homelessness than civilian women due to the prevalence of military sexual trauma (MST), unemployment, disability, poor overall health, and anxiety disorders or post-traumatic stress disorder.\textsuperscript{14} Homeless women Veterans are highly reliant upon the VHA and, unlike Veterans with more resources, may not be able to overcome travel barriers or the logistics of the “Choice” referral process to access family planning services-especially for more time-sensitive needs, such as access to emergency contraception or STI treatment. Additionally, they are less likely to have public or private insurance to access non-VHA care through civilian healthcare providers, thus publically-funded family planning clinics are their safety net for services. Publically-funded clinics typically rely upon Title X federal funding, payments from Medicaid insurance plans, and state-funded family planning waivers, with federal policies requiring all qualified providers remain eligible to participate in care funded
by these federal programs. Recent political challenges to this healthcare infrastructure have progressed in several states and at a national level and Texas became the first state to enforce a string of legislative initiatives beginning in 2011 that drastically impacted publically-funded family planning clinics. These initiatives excluded Planned Parenthood clinics from federally funded programs, diverted family planning funds to other government programs, and shifted funds from comprehensive family planning providers to those with limited service abilities. By 2015, 25% of publically-funded Texas family planning clinics had closed. Additionally, clinics offering comprehensive contraceptive services, including IUD placement, decreased from 71% to only 46% of those remaining, and Medicaid-funded childbirths had risen.

In order for the VHA to ensure the family planning needs of women Veterans are met, an assessment of safety net services, especially in rural settings where VAMCs are not accessible and many high-risk women Veterans reside, needs to be completed. Texas is an extreme example with several years of data on the impact of restrictive funding changes for these clinics. Some of these same changes, such as exclusion of Planned Parenthood clinics from Medicaid funding and limiting Medicaid and Title X coverage, are being considered at a national level as part of the debate over U.S. healthcare coverage. These changes would further limit Veterans from seeking care outside the VHA, even when VHA clinics are geographically distant for impoverished Veterans. Therefore, the objectives of this study were to assess the geographic availability of VAMCs and publically-funded family planning clinics for homeless Texan women Veterans through information visualization tools. This work is a necessary first step to develop geographic interventions to improve gender-specific care within the VHA or guide policies for non-VHA care, as each region will likely have unique needs.

Methods
The University of Utah Investigational Review Board and the Research and Development Committee at the Veterans Administration Salt Lake City Health Care System approved this study.

We identified women Veterans in the state of Texas with administrative evidence of homelessness at any time ("ever-homeless") who accessed the VHA between fiscal years (FY) 2002-2015 using a national VHA research database of administrative and clinical data managed by the Veteran’s Informatics and Computing Infrastructure (VINCI). We included only women Veterans age 18-44y in birth cohorts for each fiscal year. Homelessness was identified through International Classification of Diseases, Ninth Revision (ICD-9) codes, Homeless Care stop codes, or treatment specialty codes. In reality, the exact moment a Veteran becomes homeless is known only to the Veteran or those close to them. They come to the attention of the healthcare system when they have been identified as being homeless by a provider and coded in administrative data based on statements by the Veteran regarding housing instability or an interest in VHA homeless services that are available to those experiencing homelessness. This method of identifying homelessness among Veterans using administrative data has been validated and has been peer-reviewed and published by the Office of Inspector General of the US Department of Veterans Affairs and VHA researchers. Thus, in this study, we use administrative data to identify those women Veterans who are experiencing homelessness and have been noted as such in VHA electronic data.

Demographic variables included age at entry into VHA services, marital status, race/ethnicity, and religious affiliation. Military characteristics included branch of service, combat exposure, and results (positive or negative) of screening for a history of military sexual trauma (MST) in VHA. While MST status was not routinely screened for, or reliably captured, in structured data until 2004, Veterans who accessed care prior to 2004 were screened at follow-up encounters.

We defined the geographic “residence” of the homeless women Veteran’s last known zip code reported at the VHA site where the homelessness designation was made. For women with only post office box (PO Box) numbers, we assigned the associated zip code as their “residence”, as the PO boxes were assumed to be in close proximity to where the homeless women Veterans were staying. The population density of homeless women Veterans per zip code were plotted on a map with locations of VAMCs. The geographic distribution of all 18-44-year-old women Veterans using county-level counts from the National Center for Veterans Analysis and Statistics of women veterans were used as map background.

Travel distance was measured first as the number of drive miles between the Veteran’s “residence” zip code and the nearest Texas VAMC. Drive distance was calculated as miles using a 2010 network of roads and using the “Closest Facility” method in the Network Analyst extension of ESRI’s ArcGIS software suite. To visualize the change in availability of publically-funded family planning clinics before and after the exclusion of Planned Parenthood, we mapped all available Texas Planned Parenthood Clinics, as of 2010, using publically available county-level data.
from the Guttmacher Institute. To capture currently available (March 2017) publically-funded clinics that provide family planning services, we used publically-available data from the Texas Health and Human Services “Healthy Texas Woman” website and searched by zip code under “Find a Doctor”. We then geocoded these clinic addresses. Travel distance was calculated in the same manner as for distance to VAMCs, measured as the number of drive miles between the Veteran’s “residence” zip code and the nearest currently available clinic according to the Texas Health and Human Services website. Some clinics did not appear appropriate for homeless women Veterans, i.e. high school and teen health clinics, and were censored from analysis.

As the stage of Texas is geographically large with the bulk of the population clustered in several relatively urban areas, we used a local spatial statistic (Cluster and Outlier Analysis (Anselin Local Moran’s I), part of the spatial statistics extension for ESRI’s ArcGIS software suite) to identify zip codes that had non-random similarity and dissimilarity count values (both high and low values) of homeless women veteran as compared to neighboring zip codes. Results were expressed as Z scores, with high values indicating similarity with neighboring zip codes and low Z scores indicating dissimilar values, and as p-values. We also applied a false discovery rate correction to account for multiple testing dependency and spatial dependency (gives a more conservative answer). The method also classifies features that have statistically significant values according to how different their Z Score values.

Cluster: High are features with neighbors that have similarly high values, and High Outlier are features that have high values near neighbors with low values.

As the real-world travel burden (or the experienced conceptualization of space) is based on road travel distance rather than linear distance, we compared each zip code count value of homeless women veterans to its closest 30 neighbors where the neighborhood was defined as road distance. The reason for this approach is that while two points (an origin and a destination) can be linearly very close, actual travel may be much farther (and sometimes much farther) if a bona fide barrier (such as a mountain, valley, or river) prevents an approximate straight drive between the origin and destination. We reasoned, therefore, that a street network conceptualization of space was a more appropriate for comparing places.

We then mapped these analysis values to compare the results to the proximity of VHA facilities and to the countywide count of all women veterans of childbearing years. While CBOC VHA facilities do not routinely provide specialty family planning services, such as IUD placement, veterans who visit these facilities can be identified as being homeless or be referred for non-VHA/Choice care, and therefore we wanted to see if these facilities were also near significant clusters of homeless women veterans.

Results

We identified 41,747 women Veterans nationally with administrative evidence of homelessness who had accessed the VHA between FY2002-2015. We then limited the cohort to state of Texas zip codes only, resulting in 3,246 homeless Texan women Veterans for analysis. The majority (35%) was in the age group 18-34 years at time of entry into the VHA during the study period. Those included were most likely to have served in the army (59%) and had low combat exposure (10%). Most (48%) identified as African American and non-Hispanic (85%). The homeless women Veterans also had a high prevalence of being positive for military sexual trauma screening (39%).

Table 1. Sociodemographic and military characteristics of ever-homeless women Veterans in Texas who accessed the Veterans Healthcare Administration between fiscal years 2002-2015 (N=3,246)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Value</th>
<th>n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age group</td>
<td>18-34y</td>
<td>1127 (35)</td>
</tr>
<tr>
<td></td>
<td>35-44y</td>
<td>1054 (32)</td>
</tr>
<tr>
<td></td>
<td>&gt;44y</td>
<td>1065 (33)</td>
</tr>
<tr>
<td>Race</td>
<td>Black</td>
<td>1555 (48)</td>
</tr>
<tr>
<td></td>
<td>White</td>
<td>1315 (41)</td>
</tr>
<tr>
<td>Ethnicity</td>
<td>Not Hispanic</td>
<td>2773 (85)</td>
</tr>
<tr>
<td></td>
<td>Hispanic</td>
<td>300 (9)</td>
</tr>
<tr>
<td>Marital status</td>
<td>Married</td>
<td>636 (20)</td>
</tr>
</tbody>
</table>
Homeless women Veterans were frequently co-located in areas of high population density for all reproductive age women Veterans in Texas. High countywide counts of women veterans were located in counties with large cities, such as Dallas, Fort Worth, Houston, Austin, San Antonio, Corpus Christi, and El Paso. (Figure 1)

![Figure 1. Map: Geographic distribution of homeless women Veterans in relation to all reproductive age women Veterans and Veterans Healthcare Administration Medical Centers](image)

The mean distance from Texas Zip Codes to VAMCs was 82.6 miles (range 0.8-316.4). Counties with the greatest travel distance were in the far west near El Paso and far south near Brownsville, both along the Rio Grande River and sharing a national border with Mexico. (Figure 2)
We completed a cluster and outlier analysis for statistically significant clusters of homeless women veterans by zip code in relation to all reproductive age women veterans (represented in Figure 1 as shaded counties) and VHA Facilities. “Cluster: High” represents a statistically significant count near other features with high values, and “High Outlier” represents a statistically significant count near other features with low values. Homeless women veterans tended to be clustered around VA facilities (both VAMC and CBOC) in counties that had general high counts of women veterans. Of Texas’ 2,629 Zip codes, 90 (almost all in relatively urban areas) were classified as “Cluster: High” and three were “High Outlier” (two in El Paso and one in Houston), representing 1,286 (40%) of the 3,246 homeless women veterans. While there was no overall clustering of homeless women veterans in El Paso, there were two zip codes with in the city of El Paso with counts high enough to be considered statistical outliers. Notably, the headquarters for the active Army base, Fort Bliss, is in El Paso. (Figure 3)
The number of publically-funded family planning clinics decreased starting in 2011, following the exclusion of Planned Parenthood Clinics from funding sources, and other clinics were added as funding recipients, even though they do not provide comprehensive family planning services. To represent these changes, we explored the drive distances from zip codes to current publically-funded family planning clinics (as of March 2017) and also show the count by county of formerly funded 2010 Planned Parenthood Clinics. As of March 2017, homeless women Veterans have a mean drive distance of 24.1 miles (range 0-239) to the nearest publically-funded family planning clinic. In particular, Planned Parenthood Clinics formerly had a heavy concentration in the far south, which now require some of the longest driving distances for women to access care at currently funded clinics. (Figure 4)
Discussion

In order to overcome access barriers to specialty care, the VHA is increasing the focus on community civilian partners to meet the healthcare needs of the Veteran population. High-risk, homeless women Veterans may not be able to navigate the often complex process to access this option, yet they still need the same civilian healthcare safety net that non-Veterans use, especially for family planning services. Women with a history of military or lifetime sexual trauma and homeless women are vulnerable populations at increased risk for repeat sexual trauma. The high prevalence of MST in this study population highlights the needs for access to reproductive healthcare. In using Texas as an example of the impact of restrictive legislative family planning policies, we found publically-funded family planning clinics are more geographically accessible than VAMCs for homeless women Veterans. Unfortunately, these clinics do not routinely provide comprehensive family planning service, especially in more rural areas, as funding changes limit services, close clinics, and uproot affiliated healthcare providers. Ongoing assessments of comprehensive Texas family planning services found clinics were able to serve 54% fewer clients between 2011 and 2021-2013. In a similar comparison to our results, closure of Texas abortion clinics resulted in increased one-way travel distance for women from 22 miles to 85 miles with additional out of pocket expenses and reported difficulty in clinic access. Public or private insurance coverage and non-VHA/Choice benefits will not increase access, when trained providers are not geographically available to accept the coverage. Additionally, the most socio-economically disadvantaged women will not be able to access coverage options, as many of the same states with the most restrictive Medicaid eligibility requirements are also states without family planning waivers or national family planning Title X clinic funding to fill gaps for uninsured women.

National VHA priorities include improved access to healthcare for women and homeless Veterans, and family planning services are an essential component of preventive healthcare. Despite these commitments, many women Veterans currently bypass VHA services due to lack of availability of gender-specific care, confusion regarding VHA eligibility and travel distance to the nearest VHA facility. Longer drive time is also associated with attrition from VHA health care in women Veterans, identifying a continued need for improvement in service provision.
VHA policies and priorities attempt to standardize healthcare services and quality across the nation. Despite these goals, there are vast geographic differences for Veterans to access publically-funded, civilian services between “conservative” states like Texas and more “liberal” states such as California, where family planning services and providers are abundant due to MediCal coverage and Title X funded clinics. As the nation weighs conservative changes to federal family planning funding and restructuring of the entire VHA referral system, the adverse outcomes seen in vulnerable Texan women and the resultant impact on our women Veterans needs to be considered.

We acknowledge several limitations of this study. The use of structured data to identify homelessness among women Veterans may lead to misclassification of their true homelessness status. We also looked at any housing instability (ever homeless), instead of a limited timeframe around homeless designation. Even though homelessness is not typically a constant over many years, many socio-economic stressors pre-date and persist beyond the administrative capture of homelessness and would likely result in the need for healthcare safety net clinics. We made the assumption that homeless women Veterans who cannot access care at a VAMC would also have challenges in accessing a non-VHA/Choice referral allowing for clinical coverage outside of publically-funded clinics. The VHA will pay for transportation for many Veterans who are eligible for this benefit, dependent upon level of disability or service-connection, but the Veteran again needs to navigate the system to access this coverage option. The current clinic mapping does not include homeless clinics that may offer some free family planning services, but previous literature shows family planning options are highly variable and limited in these settings.

Conclusions

In summary, homeless women Veterans represent the most vulnerable subset of women Veterans at greatest risk for adverse reproductive outcomes. Current restrictive funding changes have limited the safety net services of publically-funded family planning clinics in the state of Texas, despite the improved geographic availability of these clinics over VAMCs. As national efforts are underway to establish civilian care as the “cure” to access issues within the VHA, the impact of civilian funding changes in Texas on the availability of family planning services and providers for women Veterans need to be considered. The current push to exclude Planned Parenthood clinics from federal funding will increase the unmet need of comprehensive family planning services nationwide, including for our vulnerable homeless women Veterans.
References


Quantifying the Impact of Trainee Providers on Outpatient Clinic Workflow using Secondary EHR Data

Isaac H. Goldstein2, Michelle R. Hribar, PhD1, Sarah Read-Brown2, Michael F. Chiang, MD, MA1,2

1Medical Informatics & Clinical Epidemiology, and 2Ophthalmology
Oregon Health & Science University, Portland, OR

Abstract

Providers today face productivity challenges including increased patient loads, increased clerical burdens from new government regulations and workflow impacts of electronic health records (EHR). Given these factors, methods to study and improve clinical workflow continue to grow in importance. Despite the ubiquitous presence of trainees in academic outpatient clinics, little is known about the impact of trainees on academic workflow. The purpose of this study is to demonstrate that secondary EHR data can be used to quantify that impact, with potentially important results for clinic efficiency and provider reimbursement models. Key findings from this study are that (1) Secondary EHR data can be used to reflect in clinic trainee activity, (2) presence of trainees, particularly in high-volume clinic sessions, is associated with longer session lengths, and (3) The timing of trainee appointments within clinic sessions impacts the session length.

Introduction

Outpatient healthcare clinics face ongoing pressures to see more patients in less time due to concerns about the accessibility and cost of healthcare.1,2 Significant challenges can impede outpatient clinic efficiency including increased patient demand with limited clinic resources, ad-hoc scheduling methods, increased clerical burdens and possible negative effects of the introduction of electronic health records (EHRs).3–6 Given these pressures and challenges, methods to study and improve clinical workflow are becoming ever more significant. In previous studies, we developed methods to improve clinic workflow through secondary use of EHR timestamps and discrete event simulation.7,8 These results showed how EHR timestamps adequately approximated clinic workflow timings for use in large scale simulations. In this study, we make use of EHR timestamps in order to address an important unanswered question: the impact of trainee providers on academic outpatient clinic workflow.

In its Common Program Requirements, The Accreditation Council for Graduate Medical Education (ACGME) states that “The specialty education of physicians to practice independently is experiential, and necessarily occurs within the context of the healthcare delivery system”.9 Despite this commitment to train within the health care system, there are few studies on graduate medical education’s impact on outpatient clinics, and many have been small in size and have had conflicting findings. For example, two studies found trainees improved workflow by shortening patient wait times,10,11 while a different study found trainees lengthened appointment times.12 More recently, we conducted a large scale investigation of the relationship between presence of trainees and outpatient appointment length and found that appointments with trainees were significantly longer than appointments without trainees (Goldstein, IH et al. IOVS 2017 58: ARVO E-Abstract 5060). Spurred on by the high demand for emergency physicians and the association between emergency department crowding and adverse health outcomes, a number of research groups have quantified and refined the impact trainee providers have on the workflow of emergency departments.13–15 Some groups found trainees are associated with both length of stay and the number of patients who leave without being seen, while others have found the association was negligible.16–18

This sparse and conflicting literature leaves many unanswered questions about the relationship between trainees and outpatient clinic workflow. This study utilizes EHR data to quantify the impact of residents and fellow trainees on half-day clinic sessions in an outpatient ophthalmology clinic. Our results show that trainees can impact the overall length of clinic sessions significantly, with implications for secondary use of EHR data, clinic efficiency, and provider reimbursement models.
Methods

This study was approved by the Institutional Review Board at Oregon Health & Science University (OHSU).

Study Environment

OHSU is a large academic medical center in Portland, Oregon. The department of ophthalmology includes over 50 faculty providers who perform more than 115,000 outpatient examinations annually. The department provides primary eye care, and serves as a major tertiary referral center in the Pacific Northwest and nationally. Typically, the department has 15 residents and 10 fellows per academic year.

EHR Dataset

For our study, we identified “stable faculty providers” who worked at OHSU for at least 6 months before and after the study period. This minimized bias from providers with growing or shrinking clinical practices. Providers who did not have a standard clinical practice, who did not use the EHR, who had a small number of appointments with trainees (less than 3), or who had only appointments with trainees were excluded. Demographics for study providers and trainees (gender, age, and ophthalmic sub-specialty) were gathered using publicly-available data. We queried check in and check out times, as well as the primary billing code, from OHSU’s clinical data warehouse (EpicCare; Epic Systems, Verona, WI).

Determining the involvement of a trainee in patient encounters is not straight forward given that there isn’t a standard method for recording when trainees participated in encounters. For the purposes of this study, we considered trainees to be involved in the appointment if there was a record in the EHR encounter. This record can appear in several areas; we used audit log entries as markers for trainee activity. We measured the time that trainee providers used the EHR during each patient appointment, while the patient was checked in. A trainee was considered present for an appointment if they used the EHR for more than two minutes during the patient’s appointment (sometimes called trainee appointments in this paper). Appointments were excluded if they were missing either a check-in or checkout time, if there was a non-physician trainee present, or if there were two of the same kind of trainee present. The final exclusion criteria eliminated 50 appointments and focused our data on common workflows with physician trainees.

From this data set of appointments, we aggregated the encounter data into half-day clinic sessions. Session length, the dependent variable in our models, was calculated as the first patient check-in time of the session subtracted from the last patient checkout time of the session. A trainee session was defined as a half-day clinic session with more than one patient appointment in which a trainee was present. Because volume is a big determinant of the length of a clinic session we included patient volume as a factor in our model. To simplify analysis, we categorized clinic sessions as low, medium, and high. A low patient volume session was defined as one with ≤6 patients, a medium patient volume session was defined as one with 7-14 patients, and a high patient volume session was defined as one with >14 patients.

Data Validation

To validate our method of using audit log entries to determine the presence of a trainee in an appointment, we (SRB) conducted a thorough manual chart review of 50 appointments, and determined whether or not a trainee was present for the appointment if one or more of four criteria were met:

1. Attestation: The provider specifically affirmed they reviewed the activity of a trainee provider.
2. Revision History: The chart’s revision history showed activity by a trainee provider during the exam.
3. Edited Exam Element: The chart showed a trainee provider edited exam elements.
4. Signature: A trainee provider signed the note following the exam.

We then compared the study’s method for determining the presence of trainees, as well as each of the four criteria individually, to the results of this chart review. We further tested the robustness of our method of using 2 minutes of audit log activity by creating new data sets with adjusted time cutoffs and comparing them to the original.

Exclusions

Because we are studying the effect of trainees on regularly scheduled clinic sessions, we excluded sessions whose patient volume was in the bottom quartile for that session’s provider. These low volume sessions typically represent limited clinic sessions (e.g. a few post-op appointments on a non-clinic day). Since some of our analysis involved
dividing clinic session into quartiles, we excluded all sessions with fewer than four appointments. Finally, we 
excluded one outlier session with a session length of 16 hours, which was due to an invalid checkout time.

Data Analysis
To analyze the impact of trainees on session length, we constructed multiple linear and linear mixed models with 
interacting terms. Models were used to analyze the impact on session length of: (1) the presence of trainees in a 
session (2) the percentage of trainee appointments in a session (3) the length of trainee appointments, and (4) the 
timing of trainee appointments. For the linear mixed models without interacting terms, p-values were obtained 
through type II Wald chisquare tests, for those with interacting terms, multiple comparisons and the Holm-
Bonferroni method were used. For all tests, significance was defined as p<.05. All data processing and statistical 
calculations were conducted in R
, models were constructed via lme4, p-values were calculated via the Anova 
function from the car package, and the glht function from the multcomp package.

Results

Overview of providers and sessions

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Providers (N = 33)</th>
<th>Residents (N=20)</th>
<th>Fellows (N=20)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex - no. (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>13 (39.4)</td>
<td>9 (45.0)</td>
<td>8 (40.0)</td>
</tr>
<tr>
<td>Age - yr †</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Median</td>
<td>44.0</td>
<td>30.0</td>
<td>32.0</td>
</tr>
<tr>
<td>Interquartile range</td>
<td>39.0 - 53.0</td>
<td>29.0-31.0</td>
<td>30.0-33.0</td>
</tr>
<tr>
<td>Provider type - no. (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M.D.</td>
<td>29 (87.9)</td>
<td>20 (100.0)</td>
<td>18 (90.0)</td>
</tr>
<tr>
<td>O.D.</td>
<td>4 (12.1)</td>
<td>0 (0.0)</td>
<td>2 (10.0)</td>
</tr>
<tr>
<td>Sub-specialty - no. (%)‡</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Glaucoma</td>
<td>4 (12.1)</td>
<td>NA</td>
<td>3 (15.0)</td>
</tr>
<tr>
<td>Pediatrics</td>
<td>4 (9.0)</td>
<td>NA</td>
<td>2 (10.0)</td>
</tr>
<tr>
<td>Retina</td>
<td>4 (12.1)</td>
<td>NA</td>
<td>4 (20.0)</td>
</tr>
<tr>
<td>Oculoplastic</td>
<td>3 (9.0)</td>
<td>NA</td>
<td>1 (5.0)</td>
</tr>
<tr>
<td>Optometry</td>
<td>3 (9.0)</td>
<td>NA</td>
<td>2(10.0)</td>
</tr>
<tr>
<td>Uveitis</td>
<td>3 (9.0)</td>
<td>NA</td>
<td>2 (10.0)</td>
</tr>
<tr>
<td>Comprehensive</td>
<td>3 (9.0)</td>
<td>NA</td>
<td>0 (0.0)</td>
</tr>
<tr>
<td>Cornea</td>
<td>3 (9.0)</td>
<td>NA</td>
<td>4 (20.0)</td>
</tr>
<tr>
<td>Genetics</td>
<td>2 (6.0)</td>
<td>NA</td>
<td>2 (10.0)</td>
</tr>
<tr>
<td>Neuro-Ophthalmology</td>
<td>2 (6.0)</td>
<td>NA</td>
<td>0 (0.0)</td>
</tr>
<tr>
<td>Oncology</td>
<td>2 (6.0)</td>
<td>NA</td>
<td>0 (0.0)</td>
</tr>
<tr>
<td>Low Vision</td>
<td>1 (3.0)</td>
<td>NA</td>
<td>0 (0.0)</td>
</tr>
<tr>
<td>Appointments - no. (% Trainee Appointments)§</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Study Population</td>
<td>42690 (24.9)</td>
<td>4042</td>
<td>6422</td>
</tr>
<tr>
<td>Sessions - no. (%Trainee Sessions)ǁ</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Study Population</td>
<td>3764 (41.4)</td>
<td>836</td>
<td>907</td>
</tr>
<tr>
<td>Low Patient Volume</td>
<td>1064 (25.8)</td>
<td>183</td>
<td>88</td>
</tr>
<tr>
<td>Medium Patient Volume</td>
<td>1632 (39.4)</td>
<td>375</td>
<td>322</td>
</tr>
<tr>
<td>High Patient Volume</td>
<td>1068 (59.8)</td>
<td>278</td>
<td>497</td>
</tr>
</tbody>
</table>

Table 1. Characteristics of faculty ophthalmology providers, residents and fellows.* Demographic information for the providers and trainee 
providers, as well as summary statistics for the study appointments and sessions.

* Providers were identified based on having worked at the study institution for 6 months before and after the study period.
† Age and length of practice are calculated based on the beginning of the study period Jan 1, 2014.
‡ Because of rounding, the percents do not add to 100
§ Trainee appointments have either residents, fellows, or both. Note the number of trainee appointments is the not the sum of the resident appointments and the 
fellow appointments, as some appointments have both. The total number of trainee appointments was 10635
ǁ Trainee Sessions have two or more resident or fellow appointments per session. Note the number of trainee sessions is the not the sum of the resident 
and the fellow sessions, as some sessions have both. The total number of trainee sessions was 1557

The time period of this study was from January 1, 2014 to December 31, 2014. Data regarding attending faculty providers are shown in Table 1. There were 33 faculty attending providers who met study inclusion criteria. These 
faculty providers had 3,764 half-day clinic sessions which met study criteria, of which 1,557 were trainee sessions.
In total 12,247 appointments and 2,197 sessions were excluded.

Low volume sessions had on average 5 appointments, medium volume sessions had on average 10 appointments, high volume sessions had on average 19.8 appointments. Low volume trainee sessions had on average 3.4 trainee appointments, medium volume trainee sessions had 5.2 trainee appointments, and high volume trainee sessions 9.8 trainee appointments. In this study, 39.4% of providers were female, and the median age was 44.0 years. Twelve different ophthalmic sub-specialties were represented, with the largest numbers in glaucoma, pediatrics, and retina (N=4 each).

During the calendar year, there were 10 Post Graduate Year (PGY) 2 and PGY-3 residents, 9 PGY-4 residents, and 23 fellows. As our data are from the calendar year 2014, these numbers represent two different classes of trainee providers: the 2013-2014 year, and the 2014-2015 year.

Validation and robustness of trainee identification method

Table 2 shows the validation results of comparing our method of using audit log entries to manual chart review. Of the methods tested, only this study’s method based on audit log entries had a sensitivity score of 1, it also was the only method with a specificity score less than 1, with a score of .97. In addition to validating our method, we also wished to test its robustness. To this end we created two additional data sets, one where a trainee was present if for more than one minute a trainee used the EHR during the exam (>1), and another where a trainee was consider present if for more than three minutes a trainee used the EHR during the exam (>3), and compared the number of appointments with trainees, number of sessions with trainees, and model impact of the presence of trainees on session length to those of the study’s data set (>2), shown in Table 3. As the definition was loosened or restricted, the appointments and sessions with trainees increased and diminished predictably, with >3 having the largest difference in trainee appointments with an 8.6% decrease compared to >2. Notably, for all three data sets, the model impact of the presence of trainees on session length (discussed below) was significant, and the size of the impact was similar.

Association of presence of trainees and session length

Figure 1 summarizes the session data regarding the association of presence of trainees with session length, and shows the spread of session lengths for trainee vs. non-trainee sessions, grouped by patient volume to help eliminate differences due to the number of patients seen at the clinic. For each category of patient volume, the median session length for trainee sessions is larger than the median session length of non-trainee sessions. To see when the presence...
of trainees produced a significant difference in session length, we must also consider the effect of providers on this difference. We constructed a linear mixed model where half-day clinic sessions were the input and session length the output, the fixed effect was the presence of trainees, the random effect was providers. This model showed that trainee sessions were associated with an increase of 10.3±2.4 minutes compared to non-trainee sessions (p<.001). To further analyze this effect, we added patient volume as an interacting fixed effect along with the presence of trainees. This more detailed model showed that high volume sessions with trainees were associated with an increase in session length of 21.0±3.9 minutes (p<.001), but the association was not significant for medium and low volume sessions.

**Association of percentage of trainee appointments and session length**

---

**Figure 1: Distribution of trainee and non-trainee session lengths.** Boxplots of session length, grouped by volume and presence of trainees. In all Categories, session length varies greatly, though the median session length of sessions with trainees is larger than the median session length of sessions with no trainees.

**Figure 2: Percent trainees versus session length.** We compared the percent of a session’s appointments which were trainee appointments to session length. For all volume groups, as the percentage of trainee appointments increases, so too does average session length. Trend lines are significant with p<.001. Thin lines are 95% confidence intervals.
We next analyzed the association between the amount of trainee activity and clinic session length. Figure 2 displays the session data regarding this association with sessions grouped by their volume label. All three trend lines were significant, with low volume sessions seeing an increase of 3.5±.7 minutes per 10% increase, medium volume sessions seeing an increase of 2.8±.5 minutes per 10% increase, and high volume sessions seeing an increase of 2.8±.8 minutes per 10% increase.

To test this further, we developed another linear mixed model with the percentage of trainee appointments as the fixed effect, and the individual providers as the random effect. This model showed that the percentage of trainee appointments was associated with an additional .86±.4 minutes per 10% increase (p<.03). When this model was expanded to take patient volume per half-day clinic session into account, the impact of the percentage of trainee appointments was significant for high patient volumes, with an impact of an additional 2.1±.6 minutes per 10% increase (p<.004).

**Trainee appointment length**

Next, we investigated how the length of trainee appointments affected clinic length. To assess this, we used a ratio of a session’s average trainee appointment length compared to the session’s provider’s average appointment length for appointments from sessions with no trainees. Table 4 shows the results of our linear mixed models: the ratio and patient volume are interacting fixed effects and the providers are the random effect. Overall, an increase in the ratio was associated with a significant increase in session length of 7.1±.5 minutes per 10% increase, with the greatest increase for high volume sessions. This association was significant for all patient volumes (p<0.001).

**Trainee appointment timing**

Finally, we analyzed the association of the timing of trainee appointments and session length. For these tests, we divided each session into quartiles (roughly representing each hour of the session) and labeled each appointment with the quartile when it occurred.

As described above, low volume trainee sessions had on average 3.4 trainee appointments, medium volume trainee sessions had 5.2 trainee appointments, and high volume trainee sessions 9.8 trainee appointments. Keeping this in mind, the first conditions we developed were based on patient volume, and tested whether it mattered when a corresponding number of minimum trainee appointments occurred. For example, there were on average 5.2 trainee appointments in each medium volume session, meaning at least one quartile had 2 trainee appointments. Thus one condition tested was “two or more trainee appointments in the 1st quartile.”

Our only significant result from these tests was from testing the conditions “3 or more trainees in the nth quartile” among high volume sessions (N=1068). To test these conditions we created a model where the fixed, independent effects were the true/false conditions “3 or more trainees in the 1st, 2nd, 3rd or 4th quartile” and the random effect was patient volume. The only condition which proved to be significant was “3 or more trainees in the 1st quartile”, which was associated with an increase in session length of 27.8±6.0 minutes (p<.001).

Moving from absolute numbers of trainees, we examined a variety of conditions concerning the percentage of overall trainee appointments that occurred in a particular quartile of the clinic. For these conditions we used the full data set and a model with a true/false rule interacting with patient volume as the fixed effects, providers as the random effect. Our analysis showed that for high volume sessions, meeting the condition that 60% of their trainee appointments were in the first half of the session was associated with an increase in session length of 10.4±3.2 minutes (p<.002).

<table>
<thead>
<tr>
<th>Volume</th>
<th>Number of Sessions</th>
<th>Model Effect*</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>275</td>
<td>6.7±1.2</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Medium</td>
<td>643</td>
<td>5.3±.8</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>High</td>
<td>639</td>
<td>8.3±.8</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Overall</td>
<td>1557</td>
<td>7.1±.5</td>
<td>&lt;.001</td>
</tr>
</tbody>
</table>

*Reported as minutes per 10% increase in trainee appts.

Table 4: Association of the ratio of average trainee appointment time to average provider non-trainee appointment time with session length. For trainee sessions we looked at the ratio between the session's average trainee appointment length and the session's provider's average appointment time from sessions with no trainees. As the ratio increases in size, session length also increases, both overall and for all patient volumes. P-values were calculated via Wald type II chi squared tests and multiple comparisons and the Holm-Bonferroni method.
Lastly, we tested the condition “One or more trainees in the last quartile.” The model was constructed similarly to the previous one, using the full data set and patient volume as an interactive term. For high volume sessions, meeting this criteria was associated with an increase of 20.1±3.7 minutes (p<.001), for medium and low volume clinic sessions there was no significant difference. When volume was not included in the model, meeting this criteria was associated with an increase in session length of 12.1±2.4 minutes (p<.001).

Discussion

There were three key findings from this study: (1) Secondary EHR data can be used to reflect in-clinic trainee activity, (2) presence of trainees, particularly in high-volume clinic sessions, can be associated with longer session lengths, and (3) the timing of trainee appointments within clinic sessions is associated with session length.

The first key finding is that EHR data can be used to reflect in-clinic trainee activity. For this paper we determined that a trainee was present for a patient appointment if for more than two minutes a trainee provider used the EHR during the appointment. When we tested this method for sensitivity and specificity, both scores were extremely high (Table 2). While manually reviewing a patient’s revision history proved to be nearly as sensitive and more specific, our method is considerably faster. Though some form of automation could be developed to review revision history, this would be somewhat cumbersome, and given the success of our method, we believe it to be a good choice for this kind of analysis.

While numerous papers have been written using EHR data to answer clinical questions, and a robust literature of time motion studies analyze clinic workflow, few studies have used big data repositories of EHR data to address issues of clinic workflow. It is possible that one of the reasons so few studies considering the relationship between trainees and workflow in outpatient clinics exist because determining when trainees see patients by any other means is a prohibitively resource intensive endeavor. Our study provides one way to address this issue.

The second key finding is that presence of trainees, particularly in high volume clinic sessions, can be associated with longer session lengths. Overall, the presence of trainees in a half-day clinic session was associated with a ten minute increase in session length. However, once volume was included into the model, this association was only significant for high volume clinic sessions (sessions with more than 14 patients), which were about 20 minutes longer. Additionally, our study found that, for high volume sessions, as the percentage of appointments seen by trainees increases, so too does session length, suggesting that not only the involvement of trainees, but the degree of the involvement can affect session length.

While we found on average that appointments with trainees are longer than appointments without trainees, this is certainly not always the case. We were interested in how much the length of a trainee appointment was related to session length, but wanted to account for the fact that trainee appointments are not always longer than non-trainee appointments. Thus, we developed the ratio of a session’s average trainee appointment length to the session’s provider’s average non-trainee appointment length, to focus in on what happened as the length of trainee appointments became significantly longer than non-trainee appointments. Our results found that, for all patient volumes, as this ratio increased so too did session length (Table 4).

To our knowledge, there has been little research on the relationship between trainees and session length. One paper used simulations to predict that trainees would shorten both appointment time and session length, however the simulation was never corroborated with collected data. We believe our large scale-retrospective study gives a more accurate sense of the relationship between trainees and session length. Our results are also consistent with our previous study concerning trainees and outpatient clinic efficiency, which found that appointments with trainees are on average longer by 25 minutes, and that in some cases even appointments where no trainee was present from a trainee session may have longer appointment times (Goldstein, IH et al. IOVS 2017 58: ARVO E-Abstract 5060). We hypothesized this was because trainees were introduced as an additional step in the workflow of an appointment, leading to lengthened times and sometimes resource bottlenecks, which would also contribute to longer session lengths.

The third key finding is that timing of trainee appointments within clinic sessions is associated with session length. One goal of this study was to investigate not only if it mattered whether trainees were present during a clinic session,
but also if it mattered when they saw patients. In total we found three conditions concerning the timing of trainee appointments that, when met in high volume clinic sessions, led to increases in session length of 20-30 minutes. Our study found that concentrating trainee appointments in the first half of the session (in particular the first quarter) led to longer session lengths, probably caused by workflow bottlenecks created due to longer appointment times (Goldstein, IH et al. IOVS 2017 58: ARVO E-Abstract 5060). Our study also found session length increased when trainee appointments were concentrated in the last quarter of the session, this is probably because longer appointment times at the end of the day push back the final checkout time, and thus lengthen the session.

Together, these key findings tell us that trainees can be associated with a significant negative impact on clinic efficiency, and that this impact is affected by a variety of factors concerning trainee involvement, activity, and timing. While it seems unlikely that clinics could change the way trainee providers are involved in appointments due to educational factors, these findings at least help academic providers better understand how trainees affect their clinic workflow.

With MACRA, the United States is continuing to transition toward “value-based” models of provider reimbursement, which are based on a combination of quality and cost of care. Key findings 2 and 3 highlight a structural characteristic of academic outpatient clinics that should perhaps be considered as reimbursement models are refined, because presence of trainees is associated with increased time – and therefore lower efficiency. Specifically, our results suggest that, due to the presence of trainees academic outpatient providers see fewer patients in a given period of time than non-academic outpatient providers.

Limitations

There were several study limitations that should be highlighted. (1) There is a great deal of variance associated with clinic session length that is not associated with trainees. This most likely is why so many of the rules we tested failed, notable ones included: the percentage of trainee appointments in any given quartile of the clinic session, the number of trainee appointments in quarter four, and the lack of trainee appointments in quarter four. The fact that we could not establish significance for these factors does not speak to whether or not they are actually important factors in determining session length, simply that we were unable to do so with our data set and models. Further research seems warranted to help correct for this variance. (2) It is worth noting that we calculated session length using appointment checkout times, which may not be correct due to staff delays. This may also contribute to some of the variance in our models. (3) Finally, this study only considers appointments where trainees used the EHR as part of their involvement, there are training activities that do not involve the EHR which our study fails to take into account. Other research modalities would seem to be necessary to address limitations 2 and 3.

Conclusion

We have shown that secondary EHR timestamp data may be applied to arrive at important and timely conclusions about the impact of trainees on clinic workflow. Findings from this study demonstrate that presence of trainees is associated with longer clinic session lengths, and may have implications for clinical care, medical education, and policymaking regarding provider reimbursement.

Acknowledgements

Supported by grants T15LM007088, K99LM012238, and P30EY0105072 from the National Institutes of Health, (Bethesda, MD) and by unrestricted departmental support from Research to Prevent Blindness (New York, NY).

References


9. ACGME Common Program Requirements [Internet]. 2016;Available from: http://www.acgme.org/Portals/0/PFAssets/ProgramRequirements/CPRs_07012016.pdf


Inferring Clinical Correlations from EEG Reports with Deep Neural Learning

Travis R. Goodwin, MS, Sanda M. Harabagiu, PhD
The University of Texas at Dallas, Richardson, TX, USA

Abstract

Successful diagnosis and management of neurological dysfunction relies on proper communication between the neurologist and the primary physician (or other specialists). Because this communication is documented within medical records, the ability to automatically infer the clinical correlations for a patient from his or her medical records would provide an important step towards enabling health care systems to automatically identify patients requiring additional follow-up as well as flagging any unexpected clinical correlations for review. In this paper, we present a Deep Section Recovery Model (DSRM) which applies deep neural learning on a large body of EEG reports in order to infer the expected clinical correlations for a patient from the information in a given EEG report by (1) automatically extracting word- and report-level features from the report and (2) inferring the most likely clinical correlations and expressing those clinical correlations in natural language. We evaluated the performance of the DSRM by removing the clinical correlation sections from EEG reports and measuring how well the model could recover that information from the remainder of the report. The DSRM obtained a 17% improvement over the top-performing baseline, highlighting not only the power of the DSRM but also the promise of automatically recognizing unexpected clinical correlations in the future.

Introduction

Diagnosing and managing neurological dysfunction often hinges on successful communication between the neurologist performing a diagnostic test (such an Electroencephalogram or EEG), and the primary physician or other specialists. In 2005, Glick et al. studied malpractice claims against neurologists and found that 71% of the claims arose from “an evident failure of communication by the neurologist” and that the majority of the claims resulted from deficient communication between the neurologist and the primary physician or other specialists. In addition, Glick et al. found that 62.5% of claims included diagnostic errors and that 41.7% involved errors in “ordering, interpreting, and reporting of diagnostic imaging, follow-through and reporting mechanisms.” It is expected that these types of errors could be reduced, and communication could be improved by developing tools capable of automatically analyzing medical reports. Moreover, a recent Institute of Medicine Report advocated the need for decision-support tools operating on electronic health records for primary care and emergency room providers to manage referral steps for further evaluation and care of persons with epilepsy. Specifically, the ability to automatically extract and analyze the clinical correlations between any findings documented in a neurological report and the over-all clinical picture of the patient, could enable future automatic systems to identify patients requiring additional follow-up by the primary physician, neurologist, or specialist. Furthermore, systems capable of automatic analysis of the clinical correlations documented in a large number of reports could ultimately provide a foundation for automatically identifying reports with incorrect, unusual, or poorly-communicated clinical correlations mitigating misdiagnoses and improving patient care. It should be noted, however, that automatically identifying incorrect, unusual, or poorly-communicated clinical correlations has two critical requirements: (1) inferring what the expected clinical correlations would be for the patient and (2) quantifying the degree of disagreement or contradiction between the clinical correlations documented in a report and the expected clinical correlations for the patient. In this initial study, we focus on the first requirement by considering the clinical correlation sections documented in EEG reports.

The role of the clinical correlation section is not only to describe the relationships between findings in the EEG report and the patient’s clinical picture, but to also explain and justify the relationships so as to convince any interested health care professionals. Consequently, the clinical correlation section of an EEG report is expressed through natural language, meaning that the clinical correlations documented in the clinical correlation section are qualified and contextualized through all the subtlety and nuance enabled by natural language expression. For this reason, while it might appear sufficient to simply extract individual findings or medical concepts from the clinical correlation section, describing and justifying the clinical correlations requires producing coherent natural language. This requirement makes inferring the expected clinical correlation section from an EEG report a challenging problem because it requires not only identifying the correct clinical correlations, but also expressing those correlations through natural language which is by the content of the EEG report as well as the neurologist’s medical knowledge and accumulated experience.

In this paper, we present a novel Deep Section Recovery Model (DSRM) which applies deep neural learning on a large body of EEG reports in order to infer the expected clinical correlations for a patient based solely on the natural language content in his or her EEG report. The DSRM was trained and evaluated using the Temple University Hospital (TUH) EEG Corpus by (a) identifying and removing the clinical correlation section written by the neurologist and (b) training the DSRM to infer the entire clinical correlation section from the remainder of the report. At a high level, the DSRM can be viewed as operating through two general steps:
Step 1: word- and report-level features are automatically extracted from each EEG report to capture contextual, semantic, and background knowledge; and

Step 2: the most likely clinical correlation section is jointly (a) inferred and (b) expressed through automatically generated natural language.

Our experimental results against a number of competitive baseline models indicate the generative power of the DSRM, as well as the promise of automatically recognizing unusual, incorrect, or incomplete clinical correlations in the future. It should be noted that although we evaluated the DSRM by recovering the clinical correlation sections from EEG reports, the model automatically extracts its own features based on the words in a given report and (clinical correlation) section. Consequently, we believe the DSRM could be easily adapted to not only process addition types of medical reports, but to also to infer and generate medical language for other purposes, e.g., generating explanations for CDS systems, providing automated second opinions, and assessing and tracking documentation quality.

Background

The Deep Section Recovery Model (DSRM) presented in this paper was originally envisioned as part of a larger project to design an automatic patient cohort retrieval system (operating on natural language) for EEG reports. This system assigns different weights or importance to each section in an EEG report, with the clinical correlation section being the most important. Unfortunately, we found that as many as 1 in 10 EEG reports were missing a clinical correlation section. In previous work, we designed a binary classification model for automatically inferring the over-all impression (normal or abnormal) for an EEG report. This model was extended and adapted to produce natural language, forming the basis for the DSRM presented in this paper.

As a natural language generator, the DSRM incorporates advances from Natural Language Generation, Machine Translation, and Automatic Summarization. We briefly review each of these topics below.

Natural Language Generation. Natural language generation (NLG) is an area of study on how automatic systems can produce high-quality natural language text from an internal representation. Traditionally, NLG systems rely on a pipeline of sub-modules including content selection – determining which information the model should generate – and surface realization – determining how the model should express the information in natural language. These systems typically require supervision at each individual stage and cannot scale to large domains. In health care, NLG has traditionally focused on surface realization through a number of applications, including generating explanations, advice, or critiques in expert systems, as well as generating explanatory material for patients. These systems largely rely on templates and rule-based mechanisms for producing natural language content. By contrast, the DSRM jointly performs content selection (via latent feature extraction) and surface realization (using a deep neural language model) without requiring predefined rules or templates.

Machine Translation. Perhaps the most ubiquitous application of NLG, machine translation has been an active area of research for the last 50 years. While the earliest systems were largely rule-based, statistical machine translation (SMT) systems have become the focus of the field. Statistical machine translation systems typically rely on gold-standard word or sentence alignments between parallel texts in a source and target language and use machine learning to train models which can automatically translate between them. More recently, the advent of deep learning has enabled the design of systems which jointly learn to align and translate. The canonical work by Bahdanau et al. introduces the notion of neural attention, which allows the model to learn how words in the target language should be aligned to words in the source language without supervision. The DSRM extends this idea by incorporating an attention layer to learn the association between words in the clinical correlation section and those in the rest of the report.

Automatic Summarization. Automatic summarization systems can be typically divided into two categories: extractive summarization systems, which aim to select individual words or sentence from a document and “stitch” them together to form a summary, and abstractive summarization systems which consider structural and/or semantic information to produce a summary that can contain words not mentioned in the document. It has been shown that extractive summarization may not be sufficient for health care needs; rather, abstract summarization efforts should be preferred. Fortunately, as with SMT, advances in deep learning have allowed allowed summarization systems to learn an internal or embedded representation of a document which can be used as the basis for NLG using so-called Sequence-to-Sequence models. Consequently, the DSRM model adapts the notion of abstractive summarization and combines and extends Sequence-to-Sequence models with the attention mechanisms used by SMT systems.

Data

The experiments reported in this paper use the Temple University Hospital (TUH) EEG Corpus with a standard 3:1:1 split for training, validation, and testing sets. The TUH EEG Corpus is the largest publicly available collection of EEG reports and the first...
Table 1: Examples of EEG Report sections from the TUH EEG Corpus (each section was taken from a different EEG report).

<table>
<thead>
<tr>
<th>Section</th>
<th>Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clinical History</td>
<td>An elderly woman with change in mental status, waxing and waning mental status, COPD, morbid obesity, and markedly abnormal EEG. Digital EEG was done on XXXX XX, XXXX.</td>
</tr>
<tr>
<td>Introduction</td>
<td>The EEG was performed using the standard 10/20 electrode placement system with an EKG electrode and anterior temporal electrodes. The EEG was recorded during wakefulness and photic stimulation, as well as hyperventilation, activation procedures were performed.</td>
</tr>
<tr>
<td>Medications</td>
<td>Keppra, Aricept, Senna, Aricept, ASA, famotidine</td>
</tr>
<tr>
<td>Description</td>
<td>In wakefulness, the background EEG is very low voltage, relatively featureless with some 10 Hz activity in the background and a posterior dominant rhythm, which may be estimated at 7 Hz. The patient seems to have very brief lapses into sleep with diffuse 10 to 13 Hz activity and then spontaneous arousals. This pattern is a beta spindle and then an arousal can be identified throughout the record. Later portions of the record seem to demonstrate more sustained sleep, but with ongoing eye movements. HR: 66 BPM.</td>
</tr>
<tr>
<td>Impression</td>
<td>Abnormal EEG due to: 1. Slow and disorganized background. 2. Left occipital sharp waves, at times becoming somewhat periodic in sleep. 3. Some additional epileptiform discharges with more of a mid to posterior temporal localization.</td>
</tr>
<tr>
<td>Clinical Correlation</td>
<td>This tracing raises the possibility of a mechanism for seizures outside of the area of the abscess described above. The photoparoxysmal response is unusual and may be accentuated by the previous surgery in the posterior brain regions.</td>
</tr>
</tbody>
</table>

The publicly released collection which includes both the raw EEG signal data as well as the EEG report associated with each EEG session. The EEG reports were authored according to The American Clinical Neurophysiology Society (ACNS) guidelines for writing an EEG report which stipulate that all EEG reports should contain (a) an introduction, (b) a description of the EEG recording, and (c) an interpretation regarding the normality or abnormality of findings as well as a correlation to the patient’s overall clinical picture. In the TUH EEG Corpus, the introduction was typically divided into three sections: (1) the Clinical History section indicating the age and gender of the patient as well as a brief history of any medical conditions which may affect the EEG recording; (2) the Medications section which consists of a comma-separated list of any medications the patient is regularly taking that could influence the EEG recording; and (3) the Introduction section itself which describes the setting of the EEG, the configuration of electrodes, the patient’s state of consciousness, whether the patient had been fasting, and any other pertinent information about the EEG setting. The description of the EEG recording was represented by the Description section which provides a “complete and objective” list of any notable findings including details about all waveforms in the record as well as a description of the patient’s background electro-cerebral activity. The neurologist’s interpretation was documented in two sections in the TUH EEG Corpus: (1) the Impression section in which the neurologist documents whether the EEG recording indicates normal or abnormal brain activity as well as – particularly in the cast of abnormal brain activity – a list of the most important findings that lead to this conclusion; and (2) the Clinical Correlation section in which the neurologist ties the findings in the report to the over-all clinical picture of the patient. Table 1 provides an example of each section from the TUH EEG Corpus.

Figure 1: Simplified Architecture of the Deep Section Recovery Model (DSRM).

Inferring the Clinical Correlation Section

When writing the clinical correlation section of an EEG report, the neurologist considers the information described in the previous sections, such as relevant clinical history or notable epileptiform activities, as well as their accumulated medical knowledge and experience with interpreting EEGs. This type of background knowledge is difficult to capture with hand-crafted features because it is rarely explicitly stated; rather, it is implied through the subtlety, context, and nuance afforded the neurologist by natural language. Consequently, to approach this problem, we present a deep neural network architecture which we refer to as the Deep Section Recovery Model (DSRM). Illustrated in Figure 1, the DSRM consists of two major components:

- the **Extractor** which learns how to automatically extract (a) feature vectors representing contextual and background...
knowledge associated with each word in a given EEG report as well as (b) a feature vector encoding semantic, background, and domain knowledge about the entire report; and

- the Generator which learns how to use the feature vectors extracted by the Extractor to produce the most likely clinical correlation section for the given report while also considering the semantics of the natural language it is generating.

In order to train and evaluate the DSRM, we identified all EEG reports in the TUH EEG Corpus which contained a CLINICAL CORRELATION section and removed that section from the report. The model was trained to recover the missing clinical correlation section in the training set and evaluated based on the clinical correlation sections it inferred for reports in the test set. In the remainder of this section, we describe (1) the natural language pre-processing steps applied to the data, (2) the mathematical problem formulation, (3) the Extractor, (4) the Generator, (5) how the parameters of the model are learned from the training set, and (6) how the learned parameters are used to infer the most likely clinical correlation section for a (new) EEG report.

**Natural Language Pre-processing**

Before applying the Deep Section Recovery Model, we pre-processed each EEG report with three basic natural language processing steps: (1) sentence boundaries were identified using the OpenNLP* sentence splitter; (2) word boundaries were detected using the GENIA tokenzer, and (3) section boundaries were identified using a simple regular expression search for capitalized characters ending in a colon. These three pre-processing steps allowed us to represent each section of an EEG report as a sequence of words in which the symbols ⟨s⟩ and /⟨s⟩ were used to indicate the start and end of each sentence, ⟨p⟩ and /⟨p⟩ were used to indicate the start and end of each section, and ⟨d⟩ and /⟨d⟩ were used to indicate the start and end of each report.

**Problem Formulation**

In order to formally define the problem, it is necessary to first define the vocabulary as the set of all words observed at least once in any section (including the clinical correlation section) of any EEG report in the training set. Let V indicate the size or number of words in the vocabulary. This allows us to represent an EEG report as a sequence of V-length one-hot vectors corresponding to each word in the report, i.e., $R \in \{0,1\}^{N \times V}$ where $N$ is the length or number of words in the report. Likewise, we also represent a clinical correlation section as a sequence of V-length one-hot vectors; in this case, $S \in \{0,1\}^{M \times V}$ where $M$ is the number of words in the clinical correlation section. The goal of the Deep Section Recovery Model is to infer the most likely clinical correlation section for a given EEG report. Let $\theta$ be the learn-able parameters of the model. Training the model equates to finding the values of $\theta$ which assign the highest probabilities to the gold-standard (neurologist-written) clinical correlation sections for each EEG report in the training set; formally:

$$\theta = \arg\max_{\theta'} Pr(S|R; \theta')$$

(1)

We decompose the probability of a particular clinical correlation section being produced for a given EEG report (i.e., correctly identifying and describing the clinical correlations in the report) into two factors:

$$Pr(S|R; \theta) \approx Pr(e,h_1,\ldots,h_N|R; \theta) \cdot Pr(S|e,h_1,\ldots,h_N; \theta)$$

(2)

where the first factor is implemented by the Extractor and the second factor is implemented by the Generator.

**The Extractor**

The language in the clinical correlation section is intended to relate findings and observations described in the previous sections of the record to the over-all clinical picture of the patient. Consequently, in order to automatically produce the clinical correlation section, the goal of the Extractor is to automatically (1) identify important neurological findings and observations (e.g., “background slowing”), (2) identify descriptions of the patient’s clinical picture (e.g., “previous seizure”), and (3) determine the inferred relationship(s) between each finding and the clinical picture as described by the EEG report or implied by medical knowledge and experience (e.g., “observed epileptiform activity is consistent with head trauma”). It should be noted that the length and content of each EEG report varies significantly throughout the collection, both in terms of the sections included in each report as well as the content in each section. Moreover, when producing an EEG report, each neurologist writes in a different style, ranging between terse 12-word sections to 600-word sections organized into multiple paragraphs. Consequently, the role

*https://opennlp.apache.org/*
of the Extractor is to overcome these barriers and extract meaningful feature vectors which characterize semantic, contextual, and domain knowledge. To address these requirements, we implemented the Extractor using the deep neural architecture illustrated in Figure 2. The Encoder relies on five neural layers to produce feature vectors for each word in the report \((h_1, \cdots, h_N)\) as well as a feature vector characterizing the entire report \((e)\):

- **Layer 1: Embedding.** The role of the embedding layer is to embed each word in the EEG report \(R\) (represented as a \(V\)-length 1-hot vector) into a \(K\)-length continuous vector \(r_i^{(1)}\) (where \(K \ll V\)). This is accomplished by using a fully connected linear projection layer, \(r_i^{(1)} = R_i W_e + b_e\), where \(\{W_e \in \mathbb{R}^{V \times K}, b_e \in \mathbb{R}^{1 \times K}\} \in \theta\) correspond to the vocabulary projection matrix and bias vector learned by the Extractor.

- **Layer 2: Bidirectional Recurrent Neural Network.**
  Layer 2 implements a bidirectional recurrent neural network (RNN) using two parallel RNNs trained on the same inputs: (1) a forward RNN which processes words in the EEG report in left-to-right order and (2) a backward RNN which processes words in the EEG report in right-to-left order. This allows the forward RNN to extract features capturing any short- or long-range contextual information about each word in \(R\) provided by any preceding words in the EEG report (e.g. that “slowing” is negated in “no background slowing”). Likewise, the backward RNN extracts features capturing any short- or long-range contextual information provided by successive words in the EEG report (e.g. that “hyperventilation” described in the introduction section may influence the inclusion of “spike and wave discharges” in the EEG impression or description sections). Formally, the forward RNN maps the series word embeddings \(r_1^{(1)}, \cdots, r_N^{(1)}\) to a series of “forward” word-level feature vectors \(r_1^{(2)}, \cdots, r_N^{(2)}\), while the backward RNN maps \(r_1^{(1)}, \cdots, r_N^{(1)}\) to a series of “backward” word-level feature vectors \(r_1^{(2b)}, \cdots, r_N^{(2b)}\). In our model, the forward and backward RNNs were implemented as a series of shared Gated Recurrence Units\(^{22}\) (GRUs)*.

- **Layer 3: Concatenation.** The concatenation layer combines the forward and backward word-level feature vectors to produce a single feature vector for each word, namely, \(r_i^{(3)} = \left[ r_i^{(2)}, r_i^{(2b)} \right] \) where \([x; y]\) indicates the concatenation of vectors \(x\) and \(y\).

- **Layer 4: \(2^{nd}\) Bidirectional Recurrent Neural Network.** In order to allow the model to extract more expressive features, we use a second bidirectional RNN layer. This layer operates identically to the bidirectional RNN in Layer 2, except that the word-level feature vectors produced in Layer 3, i.e., \(r_1^{(3)}, \cdots, r_N^{(3)}\), are used as the input to the bidirectional RNN (instead of \(r_1^{(1)}, \cdots, r_N^{(1)}\) used in Layer 2). Likewise, the memory states produced in Layer 4 are denoted as \(f_2\) and \(b_2\), corresponding to the forward RNN and the backward RNN, respectively. Unlike the bidirectional RNN used in Layer 2, we use the final memory of the forward RNN (i.e. \(f_2\)) as the report-level feature vector \(e\) which will be used by the Generator.

- **Layer 5: \(2^{nd}\) Concatenation.** As in Layer 3, the second concatenation layer combines the forward and backward word-level features vectors produced in the previous layer. In the case of Layer 5, however, we used the resulting feature vectors \(h_1, \cdots, h_N\) as the word-level feature vectors which will be provided to the Generator.

* A GRU is a block of coordinated sub-layers in a neural network which learn to transform an input vector (e.g. \(r_i^{(3)}\)) into an output vector (e.g. \(r_i^{(2f)}\) or \(r_i^{(2b)}\)) by maintaining and updating an internal memory state. The memory state used in the forward RNN is denoted by \(f_1\) while the memory state used in the backward RNN is denoted by \(b_1\).
The Generator

The role of the Generator is to generate the most likely clinical correlation section for a given EEG report using the feature vectors extracted by the Extractor. It is important to note that because the clinical correlation sections vary both in terms of their length and content, the number of possible clinical correlations sections that could be produced is intractably high \( (V_{\text{MAX}} \) where \( V_{\text{MAX}} \) is the maximum length of a clinical correlation section). Consequently, we substantially reduce the complexity of the problem by modeling the assumption that each word in the clinical correlation section can be determined based solely on (1) the word-level feature vectors \( h_1, \ldots, h_N \) extracted by the Extractor, (2) the report-level feature vector \( e \) extracted by the Extractor, and (3) any preceding words produced by the Generator. This assumption allows us to define the probability of any clinical correlation section, \( S' \), having been produced by a neurologist for a given EEG report (i.e., the second factor in Equation 2) as:

\[
Pr(S'|R) = \prod_{j=1}^{M} Pr(S'_j|S'_{j-1}, \ldots, S'_1, e, h_1, \ldots, h_N; \theta)
\]  

To compute Equation 3, we designed the Generator to act as a type of Recurrent Neural Language Model (RNLM) which incorporates a Recurrent Neural Network (RNN) to produce one word in the clinical correlation section at-a-time while maintaining and updating an internal memory of which words have already been produced.

![Diagram of the Generator's architecture](image)

**Figure 3:** Detailed Architecture of the Generator under (a) Training and (b) Inference Configurations.

To improve training efficiency, the Generator has two similar but distinct configurations: one for training, and one for inference (e.g., testing). Figure 3 illustrates the architecture of the Generator under both configurations. The primary difference between each configuration is the input to the RNN: when training, the model embeds the previous word from the gold-standard clinical correlation section (e.g., \( S'_j \)) to predict \( S'_j \) while during inference the RNN operates on the embedding of the previously generated word (e.g., \( S'_{j-1} \)) to predict \( S'_j \). The Generator produces the natural language content of a clinical correlation section for a given EEG report using four layers (with the preliminary embedding layer in the training configuration acting as an extra “zero”-th layer):

- **Layer 0: Embedding.** The embedding layer, which is only used when the Generator is in training configuration, embeds each word in the gold-standard clinical correlation section \( S_j \) (represented by \( V \)-length 1-hot vectors) into an \( L \)-length continuous vector space, \( s_j^{(0)} \), where \( L \ll V \). This is accomplished by using a fully connected linear projection layer, \( s_j^{(0)} = S_j W^{(0)} + b^{(0)} \) where \( (W^{(0)} \in \mathbb{R}^{V \times L}, b^{(0)} \in \mathbb{R}^L) \in \theta \) correspond to the vocabulary projection matrix and vocabulary bias vector learned by the Generator.

- **Layer 1: Concatenation.** The first layer used in both configurations of the Generator is a concatenation layer which combines the embedded representation of the previous word with \( e \), the report-level feature vector extracted by the Extractor, \( s_j^{(1)} = [s_j^{(0)}, e] \) where \([x,y]\) indicates the concatenation of vectors \( x \) and \( y \) and \( s_0^{(0)} \) is defined as a zero vector.

- **Layer 2: Gated Recurrent Unit.** The second layer used by both configurations is a Gated Recurrent Unit (GRU). The GRU allows the model to accumulate memories encoding long-distance relationships between each produced word of the clinical correlation section,
We evaluated the performance of the Deep Section Recovery Model (DSRM) using the Temple University Hospital EEG Corpus\textsuperscript{5}.

In order to improve the quality and coherence of natural language produced by the Generator, an attention mechanism was introduced. The attention mechanism allows the Generator to consider all of the world-level feature vectors $h_1, \ldots, h_N$ produced by the Extractor for the given report, and learns the degree that each word in the EEG report influences the selection of (or aligns with) $S_j^\prime$; formally:

$$s_j^{(3)} = \sum_{i=1}^{N} \alpha_{ij} h_i$$

such that $\alpha_{ij}$ is an alignment vector used in the alignment model $\beta_{ij}$ which determines the degree that the $i^{th}$ word in the EEG report $R$ (represented by $h_i$) influences the $j^{th}$ word of the clinical correlation section $S_j^\prime$ (represented by $s_j^{(2)}$).

The role of the fourth layer is to combine the result of the previous attention layer with the result of the GRU in Layer 2, i.e., $s_j^{(4)} = s_j^{(3)} + s_j^{(2)}$.

In order to measure the probability of each word $S_j^\prime$ being produced for the given EEG report, we use a final softmax projection layer to produce a vocabulary-length vector $s_j^{(5)}$ in which the $v^{th}$ element indicates the probability that $S_j^\prime$ should be generated as the $v^{th}$ word in the vocabulary. $s_j^{(5)} = \text{softmax}(\langle V \rangle W_j + b_j)$ where softmax($x$) = \frac{\exp(x_v)}{\sum_{v=1}^{V} \exp(x_v)} and $v \in [1,V]$. This allows us to complete the definition of Equation 3:

$$Pr(S_j^\prime = v|S_{j-1}^\prime, \ldots, S_1^\prime, e, h_1, \ldots, h_N; \theta) = s_j^{(5)}$$

Training the Deep Section Recovery Model

Training the Deep Section Recovery Model (DSRM) is achieved by finding the parameters $\theta$ which are most likely to produce the gold-standard clinical correlation sections for each EEG report in the training set $T$. Formally, we model this by minimizing the cross-entropy loss between the vocabulary-length probability vectors produced by the model ($s_j^{(5)}$) and the one-hot vectors corresponding to each word in the gold-standard clinical correlation section ($S_j^\prime$).

$$L(\theta) \propto \sum_{(R,S) \in T} \left[ \sum_{j=1}^{M} \left[ s_j^{(5)} \log S_j + (1 - s_j^{(5)}) \log(1 - s_j^{(5)}) \right] \right]$$

The model was trained using Adaptive Moment Estimation (ADAM)\textsuperscript{24} (with an initial learning rate $\eta = 0.001$).

Inferring Clinical Correlations

Given $\theta$ learned from the training set, the clinical correlation section $S$ can be generated for a new EEG report $R$ using the inference configuration illustrated in Figure 3b. In contrast to the training configuration in which $S_j^\prime$ is selected using the previous word from the gold-standard clinical correlation section ($S_{j-1}^\prime$), during inference, the model predicts $S_j^\prime$ using the word previously produced by the model ($S_{j-1}^\prime$). It is important to note that, unlike training, we do not know the length of the clinical correlation section we will generate. Consequently, the model continually generates output until it produces the END-OF-SECTION symbol $\langle /p \rangle$. Thus, the length of the inferred clinical correlation section $M$ is determined dynamically by the model. When inferring the most likely clinical correlation section, it is necessary to convert the vocabulary probability vectors $s_1^{(5)}, \ldots, s_M^{(5)}$ to one-hot vocabulary vectors $S_j^\prime$ that can be directly mapped to natural language.*

Experiments

We evaluated the performance of the Deep Section Recovery Model (DSRM) using the Temple University Hospital EEG Corpus\textsuperscript{5} (described in the Data section) using a standard 3:1:1 split for training, validation, and testing sets. The performance of our model was compared against four baseline systems:

1. **NN:Cosine.** In this nearest-neighbor baseline, we represented each EEG report as a bag-of-words vector. This baseline infers the clinical correlation for a given EEG report by copying the clinical correlation associated with the EEG report in the training set whose bag-of-words vector had the least cosine distance to the bag-of-words vector representation of the given EEG report.

*Let $\delta_j = \text{argmax}(s_j^{(5)}); S_j^\prime$ is defined as the one-hot vector in which the $s_j^{(5)}$ value is 1 and all other values are zero.
2. NN:LDA. In the second nearest-neighbor baseline, we represented each EEG report as a latent topic vector which was computed by applying Latent Dirichlet Allocation\(^*\) to the EEG reports in the training set. This allowed us to infer the clinical correlation for a given EEG report by copying the clinical correlation associated with the EEG report in the training whose topic-vector representation has the least Euclidean distance to the topic-vector representation of the given EEG report.

3. DL:Attn-RNLM. The first deep-learning baseline considers a recurrent neural language model\(^2\) (RNLM) using the standard attention mechanism operating on the embedded word-representations of a given EEG report. This baseline closely resembles the DSRM if the Extractor component were removed.

4. DL:Basic-S2S. The second deep-learning baseline uses a standard Sequence-to-Sequence\(^\dagger\) model without attention. This baseline closely resembles the DSRM if word-level feature vectors (i.e., \(h_1, \ldots, h_N\)) were not extracted and only the report-level feature vector is considered by the Generator.

**Implementation Details**

Our model and the two deep learning baselines were implemented in Tensorflow* version 1.0. For all deep learning models, we used a mini-batch size of 10 EEG reports, a maximum EEG report length of 800 words, a maximum clinical correlation section length of 60 words, 200-dimensional vectors for word embeddings, and 256 hidden units in all RNNs based on a grid search over the validation set.

<table>
<thead>
<tr>
<th>System/Model</th>
<th>BLEU-1</th>
<th>BLEU-2</th>
<th>BLEU-3</th>
<th>ROUGE-1</th>
<th>ROUGE-2</th>
<th>ROUGE-3</th>
<th>WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>NN:Cosine</td>
<td>0.55334</td>
<td>0.40274</td>
<td>0.32137</td>
<td>0.54284</td>
<td>0.38516</td>
<td>0.31508</td>
<td>2.521</td>
</tr>
<tr>
<td>NN:LDA</td>
<td>0.51739</td>
<td>0.36316</td>
<td>0.28199</td>
<td>0.52389</td>
<td>0.36863</td>
<td>0.28686</td>
<td>2.891</td>
</tr>
<tr>
<td>DL:Attn-RNLM</td>
<td>0.57907</td>
<td>0.41619</td>
<td>0.32433</td>
<td>0.58196</td>
<td>0.41960</td>
<td>0.32755</td>
<td>2.315</td>
</tr>
<tr>
<td>DL:Basic-S2S</td>
<td>0.58992</td>
<td>0.36829</td>
<td>0.26806</td>
<td>0.47487</td>
<td>0.31170</td>
<td>0.23445</td>
<td>2.658</td>
</tr>
<tr>
<td>DSRM</td>
<td>0.68792</td>
<td>0.54686</td>
<td>0.46323</td>
<td>0.63523</td>
<td>0.50459</td>
<td>0.42894</td>
<td>1.631</td>
</tr>
</tbody>
</table>

*\(p<0.05\), **\(p<0.01\), ***\(p<0.001\); statistical significance against DSRM using the Wilcoxon signed-rank test.

**Experimental Setup and Results**

Evaluating the quality of automatically produced natural language (such as the inferred clinical correlation sections) is an open problem in the natural language processing community. Consequently, to quantify the quality of the clinical correlation sections inferred by all four baseline systems as well as the DSRM, we considered standard metrics used to evaluate machine translation, automatic summarization, and speech recognition.

We measured the surface-level accuracy of an automatically inferred clinical correlation section in two ways: (1) the Word Error Rate\(^2\) (WER) which measures how many “steps” it takes to transform the inferred clinical correlation section into the gold-standard clinical correlation section produced by the neurologist, where steps include (a) insertion, (b) deletion, or (c) replacement of individual words in the inferred clinical correlation; (2) the Bilingual Evaluation Understudy\(^2\) (BLEU) metric which is a commonly used analogue for Precision in language generation tasks. The surface-level completeness of each inferred clinical correlation section was measured using the Recall-Oriented Understudy for Gisting Evaluation\(^2\) (ROUGE), a commonly used analogue for Recall (i.e. Sensitivity) in language generation tasks. Finally, we measured the surface-level coherence between individual words or the semantics of multi-word expressions. For example, if the gold-standard clinical correlation contains “absence of epileptiform features”, then the excerpt “no epileptiform activity” would have BLEU-2 and ROUGE-2 scores of zero and a WER of 2 despite the fact that both excerpts express the same information. Consequently, these surface-level metrics should be interpreted as strict lower-bounds on the performance of each evaluated system.

Table 2 presents these results.

It can be seen that the DSRM achieved the best over-all performance. Moreover, it can be observed that the Attention Decoder (DL:Attn-Decoder) achieved the second-best performance. The Basic Sequence-to-Sequence model (Basic-S2S) as well as the Cosine and LDA nearest neighbor approaches achieved comparable, but only moderate performance. The high performance of the DSRM compared to the Basic S2S model indicates the importance of incorporating attention, allowing the model to discover latent relationships between words in the EEG report and each word in the clinical correlation section. Moreover, the improvement in performance shown by the DSRM compared to the Attention Decoder indicates that the clinical correlation cannot be generated

---

\(^*\)https://www.tensorflow.org/
Table 3: Comparisons of inferred and gold-standard clinical correlation sections for three EEG reports.

<table>
<thead>
<tr>
<th>Example 1</th>
<th>Report: 00005044_s03</th>
<th>Inferred: No epileptiform features are identified. If epilepsy is an important consideration, a repeat EEG capturing deeper stages or sleep deprivation prior to the EEG may be helpful to identify epileptiform activity.</th>
<th>Gold: There are no definitive epileptiform discharges, but there is an amplitude asymmetry and there is an asymmetry of wicket activity. Additional recording capturing more extensive sleep may be helpful to identify epileptiform activity.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Example 2</td>
<td>Report: 00010462_s01</td>
<td>Inferred: This EEG supports a severe underlying encephalopathy and diffuse disturbance of cerebral dysfunction involving both gray and white matter. Contributing factors can include some of the renal failure, acute or metabolic processes. The focal features described above should be correlated with imaging.</td>
<td>Gold: This abnormal EEG demonstrates a severe, diffuse disturbance of cerebral function involving both gray and subcortical white matter. This EEG pattern was communicated to the primary care team.</td>
</tr>
<tr>
<td>Example 3</td>
<td>Report: 000004928_s02</td>
<td>Inferred: This EEG is not suggestive of a metabolic or intermittent encephalopathy. The rare left with focal feature suggests conforms with underlying metabolic pattern.</td>
<td>Gold: As discussed with the team on the date of this recording, this EEG is most compatible with a metabolic encephalopathy.</td>
</tr>
</tbody>
</table>

Table 4: Likert scale used to assess over-all quality of inferred clinical correlation sections.

<table>
<thead>
<tr>
<th>Likert Score</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1:</td>
<td>(strongly disagree) clinical correlation section is incomprehensible</td>
</tr>
<tr>
<td>2:</td>
<td>(disagree) clinical correlation section is not correct</td>
</tr>
<tr>
<td>3:</td>
<td>(weakly agree) clinical correlation section is generally correct, but omits important information or contains additional false or inconsistent information</td>
</tr>
<tr>
<td>4:</td>
<td>(agree) clinical correlation section is correct but omits minor details</td>
</tr>
<tr>
<td>5:</td>
<td>(strongly agree) clinical correlation section is effectively equivalent to the gold-standard</td>
</tr>
</tbody>
</table>

Discussion

In order to analyze the automatically inferred clinical correlation sections produced by the DSRM, we manually reviewed 100 randomly selected EEG reports from the test set by comparing the inferred clinical correlation sections to the gold-standard clinical correlation sections written by the neurologists. The over-all quality of the inferred clinical correlation sections was assessed using the Likert scale illustrated in Table 4, with the DSRM obtaining an average score 3.491, indicating that the inferred clinical correlation sections are generally accurate, but may contain minor additional erroneous information or have minor omissions.

Table 3 illustrates the inferred clinical correlation as well as the gold-standard clinical correlation section for three EEG reports in the test set. Example 1 illustrates an example of a correct, but incomplete inferred clinical correlation section. Both the inferred and gold-standard clinical correlation sections agree that (1) no epileptiform discharges were observed, and (2) that a repeat EEG focusing on extensive sleep is needed. However, the gold-standard clinical standard includes additional details about asymmetry and asymmetry of wicket activity which the DSRM omitted.

Example 2 illustrates an inferred clinical correlation section which accurately expresses the diffuse disturbance of cerebral function. However, the inferred clinical correlation section additionally indicates a “severe underlying encephalopathy” which was not expressed in the gold-standard clinical correlation section. Moreover, the inferred clinical correlation section attempts to correlate the findings with the patients “renal failure, and acute, and/or metabolic processes” and indicates that these findings should be correlated with imaging. While these inclusions highlight the model’s ability to accumulate knowledge across the large corpus of EEGs in the training set in order to simulate experience, they also demonstrate that the model occasionally struggles to determine which information is (or is not) relevant.

The inferred clinical correlation illustrated in Example 3 illustrates a relatively rare (15% of reviewed EEG reports) but significant error: contradiction within the inferred clinical correlation sections. While the first sentence (incorrectly) states that the EEG does not suggest metabolic encephalopathy, the second sentence indicates that it does. This error strongly suggests that the performance of the model could be improved by developing and incorporating a more sophisticated loss function: the average cross-entropy loss (shown in Equation 5) considers each individual word in the inferred clinical correlation equally; thus, the incorrect inclusion of “not” in the first sentence has a very small impact on the loss despite it inverting the meaning of the entire sentence.
Conclusion

In this paper, we have presented a deep learning approach for automatically inferring the clinical correlation section for a given EEG report, which we call the Deep Section Recovery Model (DSRM). While traditional approaches for inferring clinical correlations would require hand-crafting a large number of sophisticated features, the DSRM learns to automatically extract word- and report-level features from each EEG report. Our evaluation on over 3,000 EEG reports revealed the promise of the DSRM: achieving an average of 17% improvement over the top-performing baseline. These promising results provide a foundation towards automatically identifying unusual, incorrect, or inconsistent clinical correlations from EEG reports in the future. Immediate avenues for future work include (1) considering more sophisticated loss functions which incorporate contextual and semantic information and (2) an in-depth study and evaluation of metrics for qualifying the degree of disagreement between a given clinical correlation section and the inferred or expected clinical correlation section.

Acknowledgements

Research reported in this publication was supported by the National Human Genome Research Institute of the National Institutes of Health under award number 1U01HG008468. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

A Configurational Analysis of Risk Patterns for Predicting the Outcome After Traumatic Brain Injury

Niku Gorji1, Zsolt Zador2,3, Simon Poon1,

1 School of Information Technologies, The University of Sydney, NSW, Australia; 2Department of Neurosurgery, Salford Royal Foundation Trust, Salford, Greater Manchester, United Kingdom; 3Institute of Cardiovascular Sciences, Centre for Vascular and Stroke Research, University of Manchester, Manchester, United Kingdom

Abstract

Exploring relationships between admission variables and outcome using regression models has been the focus of Traumatic Brain Injury (TBI) research. Although practical and well established, these approaches do not evaluate interactions between predictors. We therefore applied a set-theoretic logical analysis to the Corticosteroid Randomization after Significant Head Injury (CRASH) trial database. Complete data analysis of 6945 patients demonstrated 9 different configurations of admission variables were sufficient for favorable outcome in 87.5% of all cases and explained 57% of favorable outcomes (moderate disability or good outcome). We also evaluated the contrasting configurations for unfavorable versus favorable outcome. Results are largely in line with findings of previous studies however the influence of age fell behind GCS components, which is unexpected. Specifying a combination of admission parameters that are likely to translate into a given clinical outcome is appealing from a clinician’s perspective therefore our results have considerable translational value.

Introduction

Traumatic Brain Injury is a significant source of morbidity and mortality. TBI-related disability is quoted to be 5.3 million in the United States4 and 7.7 million in the European Union old member state5. Furthermore TBI affects younger population (<45 years), which contributes to the devastating impact on society. Prognostic models have given increasing insight into predictor importance highlighting patient age, motor response and imaging findings as the most influential predictors of outcome6. These findings helped tailor our assessment protocols and pointed out the variables that should be gathered for clinical trials7. In the past, TBI studies have investigated both multi-variable and single-variable models to assess the prognostic strength of variables on TBI outcome8-10. Multi-variable models8-10 focus on development and assessment of the combined effect of multiple variables on the outcome, while single-variable approaches focus on assessing the prognostic strength of one particular variable9. More recently machine-learning methods such as Bayesian Networks have been applied to TBI databases, which proved to be an appealing way to formalize intuitive as well as unexpected associations between variables7.

Model specification for multi-variate approach using regression, often rely on the inherent assumption that each variable has an independent effect on the outcome11. It is known that these techniques assess the net effect12 of a set of variables on an outcome and are not generally concerned with configurations and interaction of variables. For understanding complex biological conditions, the interactions between variables needs to be studied in a multi-dimensional manner.

In reviewing the methods that assess the effect of multiple variables on TBI outcome, we note that the mainstream techniques are marked by limitations in expressing the interactions between variables, and the role of these interactions in predicting the outcome. This means that in these techniques, the data is assumed to have just one ready answer for the magnitude of a variable’s effect on the outcome. When interaction terms are not modelled adequately, the accuracy of estimates in regression approaches can be affected by model misspecification11. If interaction terms are not modelled, the effect of individual independent variables are likely to be over-estimated.

Modeling interactions in a multi-variate analysis is not a straight forward task. Starting from single variables, all possible combinations of variables need to be investigated. Depending on the number of variables, multiple models can be generated and the validation of these models is non-trivial. Conventional statistical methods, cannot account for situations in which only specific combinations of variables reveal their impact on the outcome (conjunctural causation) or all paths that lead to an outcome need to be simultaneously uncovered (equifinality). These methods also
fall short in explaining situations in which a given combination of variables contributes to the presence of an outcome but at the same time is irrelevant for the absence of that outcome (causal asymmetry).

Despite the depth and breadth of recent investigations, there is limited generalized knowledge to model the complex interaction of variables and the prognostic value of these interactions in TBI. In this study our goal is to systematically investigate these interactions. While considering that the predictors of favorable outcome in TBI are not necessarily the negation or reversal of predictors of unfavorable outcome, we study the interaction of variables causative to this asymmetry, in a multi-dimensional, multi-variate manner.

Set-theoretic logical analysis methods can detect recurring causal patterns, and are well suited to help us explore a configurational model of TBI outcome. For this, we apply the method of Qualitative Comparative Analysis (QCA) which unlike statistical approaches, can address the three important phenomenon of conjunctural causation, equifinality and causal asymmetry inherent in modelling the concept of configurations. The general assumption behind the configurational approach applied here is that the interaction or combinations of different predictor variables can explain the difference in outcome classes. Hence, in comparison to statistical approach like regression that provides an estimate of impacts of the study variable on outcome in a specified model, rather QCA allows a study factor to participate in difference configurations affecting the outcome.

The paper proceeds as follows: First we briefly cover the background on TBI and the current state of research in this area and will introduce the explanatory variables included in our study. Next, we explain the analytical framework behind our study, followed by the research design. We then present the QCA results and offer a more substantive interpretation of risk patterns before concluding the paper with an assessment of the predictive power of the model compared to that of a simple logistic regression model followed by a discussion.

**Prognostic Models and Predictor Variables in TBI**

The International Mission for Prognosis and Analysis of Clinical Trials in TBI (IMPACT) set forth three prognostic models with different levels of complexity, using well-known predictors (age, Glasgow Coma Motor Score, and pupillary reactivity), computed tomographic characteristics (CT classification and traumatic subarachnoid hemorrhage), secondary insults (hypoxia or hypotension) and laboratory values on admission (Hb and glucose)\(^5,6,8\). These models can predict 6-month outcome in patients with severe or moderate TBI with good discriminative ability based on the Area Under Curve (AUC).\(^7\) Assessment and validation of these widely accepted prediction models on different cohorts has been the focus of many investigations. Externally, the IMACT models were validated against the Corticosteroid Randomization after Significant Head Injury (CRASH)\(^8\) trial findings. The CRASH trial included 10008 cases of patients with traumatic head injury within 8 hours of clinical assessment from 239 hospitals in 29 countries.

We’ve based our current study on the clinically relevant variables from previous studies by IMPACT and CRASH researchers who have identified age, motor score and imaging abnormalities as important predictors of clinical outcome in TBI\(^6,8,7\). Study variables include demographics, injury characteristics, computed tomography (CT) findings and Glasgow Outcome Scale (GCS, motor, verbal response and eye opening). Outcome measure were dichotomized as death or severe disability at 6 months.

From the 10008 cases in the CRASH dataset, about a third had one or more missing values and were omitted from our analysis. Our analysis is therefore based on the 6945 cases that had no missing values. Table 1 describes the characteristics of patient data in the CRASH dataset. The missing CT findings were responsible majority of the excluded values in the study (2191 of the 10008 patients, 21.9%). For the majority of these patients (2063) a CT brain was not performed at all whereas only 128 had one or more imaging findings not recorded in the dataset. We considered multiple imputations of missing data, which would technically be difficult to interface with the subsequent analysis. Furthermore previous studies with the CRASH trial dataset found no difference between imputed and complete datasets. We therefore choose to undertake a complete data analysis rather than imputing missing values. Another consideration regarding the dataset was the better early outcomes (14 days) for high-income countries, compared low-middle income regions. The 6 month outcomes (used in our study) were however similar between income regions.
<table>
<thead>
<tr>
<th>Variable category</th>
<th>Variable (abbreviation)</th>
<th>Category</th>
<th>Total cases with no missing values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Epidemiology</td>
<td>Sex (sex)</td>
<td>Male</td>
<td>5706</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Female</td>
<td>1239</td>
</tr>
<tr>
<td></td>
<td>Age (age)</td>
<td>&lt;20</td>
<td>892</td>
</tr>
<tr>
<td></td>
<td></td>
<td>20-24</td>
<td>1191</td>
</tr>
<tr>
<td></td>
<td></td>
<td>25-29</td>
<td>860</td>
</tr>
<tr>
<td></td>
<td></td>
<td>30-34</td>
<td>754</td>
</tr>
<tr>
<td></td>
<td></td>
<td>35-44</td>
<td>1199</td>
</tr>
<tr>
<td></td>
<td></td>
<td>45-54</td>
<td>899</td>
</tr>
<tr>
<td></td>
<td></td>
<td>&gt;55</td>
<td>1150</td>
</tr>
<tr>
<td>Injury Cause (cause)</td>
<td></td>
<td>Road traffic accident</td>
<td>4780</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Fall&gt;2 meters</td>
<td>920</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Other</td>
<td>1245</td>
</tr>
<tr>
<td>Major extracranial injury (ec)</td>
<td>Yes</td>
<td>1638</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>No</td>
<td>5307</td>
</tr>
<tr>
<td>Assessment</td>
<td>Eye opening (eye)</td>
<td>No response</td>
<td>2680</td>
</tr>
<tr>
<td></td>
<td></td>
<td>To pain</td>
<td>1261</td>
</tr>
<tr>
<td></td>
<td></td>
<td>To verbal stimulus</td>
<td>1764</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Spontaneous</td>
<td>1240</td>
</tr>
<tr>
<td></td>
<td>Motor response (motor)</td>
<td>No response</td>
<td>601</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Extension</td>
<td>407</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Abnormal flexion</td>
<td>515</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Withdrawal</td>
<td>933</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Localises</td>
<td>2723</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Follows commands</td>
<td>1766</td>
</tr>
<tr>
<td></td>
<td>Verbal response (verbal)</td>
<td>No response</td>
<td>2640</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Incomprehensible sounds</td>
<td>1124</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Single words</td>
<td>821</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Confused</td>
<td>2006</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Oriented</td>
<td>354</td>
</tr>
<tr>
<td></td>
<td>Pupillary response (pupils)</td>
<td>Both reactive</td>
<td>5791</td>
</tr>
<tr>
<td></td>
<td></td>
<td>No response unilateral</td>
<td>496</td>
</tr>
<tr>
<td></td>
<td></td>
<td>No response</td>
<td>658</td>
</tr>
<tr>
<td>Image findings</td>
<td>Petechial haemorrhage (phm)</td>
<td>Yes</td>
<td>1974</td>
</tr>
<tr>
<td></td>
<td></td>
<td>No</td>
<td>4971</td>
</tr>
<tr>
<td></td>
<td>Subarachnoid bleed (sah)</td>
<td>Yes</td>
<td>2206</td>
</tr>
<tr>
<td></td>
<td></td>
<td>No</td>
<td>4739</td>
</tr>
<tr>
<td></td>
<td>Obliterated 3rd ventricle or basal cisterns (obl)</td>
<td>Yes</td>
<td>1663</td>
</tr>
<tr>
<td></td>
<td></td>
<td>No</td>
<td>5282</td>
</tr>
<tr>
<td></td>
<td>Midline shift (mdls)</td>
<td>Yes</td>
<td>1021</td>
</tr>
<tr>
<td></td>
<td></td>
<td>No</td>
<td>5924</td>
</tr>
<tr>
<td></td>
<td>Hematoma (hmt)</td>
<td>Yes</td>
<td>2718</td>
</tr>
<tr>
<td></td>
<td></td>
<td>No</td>
<td>4227</td>
</tr>
<tr>
<td>Outcome</td>
<td>Outcome at 6 months</td>
<td>Death or severe disability</td>
<td>2763</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Moderate disability or good recovery</td>
<td>4182</td>
</tr>
</tbody>
</table>
Qualitative Comparative Analysis

The method of Qualitative Comparative Analysis (QCA) carries potential for analysis of complex dependencies in configurational data\(^\text{2}\). Ragin describes QCA as “an analytic technique designed specifically for the study of cases as configurations of aspects, conceived as combinations of set memberships”\(^\text{17}\). A configuration is a combination of variables that consistently produce (i.e. are sufficient for) the outcome\(^\text{9}\).

At its core, QCA is based on ideas from the field of logic synthesis\(^\text{30}\) to obtain the minimal Boolean sum-of-products (SOP) formulas that fully represents a given truth table of variables. The truth table lists all logically possible combinations of the variables based on the dataset included in the study.

The core algorithm in QCA, the Quine-McCluskey\(^\text{20-21}\) algorithm, was established in 1950s and used for minimization of Boolean logic formulas to find the smallest, logically valid combination of variables that have the largest coverage over all cases under investigation. The minimization process is based on repeatedly applying three laws of logic: 1) absorption (e.g. \(x_1.x_2 + x_1.x_2' = x_1\)), 2) idempotency or redundancy (e.g. \(x_1 + x_1 = x_1\)), and 3) the law of excluded middle (e.g. \(x_1 + x_1' = 1\)).

The Quine-McCluskey algorithm like any other logical analysis method is not concerned with the empirical validity of the formulas that are being discovered. It is the role of the analyst to design a valid foundation for analysis and then to assess the empirical validity of the findings. After listing all variables in a truth table, the analyst needs to select the threshold at which sufficient evidence for the outcome is defined. For example, if the analyst wants to uncover all combinations of variables that lead to a certain outcome 85% of the time, the sufficiency score needs to be set to 85%.

All combinations of conditions that meet this threshold are then included in further analysis. The analyst can also define the minimum number of occurrences of a certain combination for it to be included in the study. This gives the analyst the choice to include for example all combinations that appeared at least twice for favorable outcome.

The parameters of fit in QCA are consistency and coverage\(^\text{15}\). These parameters assess how consistently a combination of conditions appears in the data and the degree to which the findings cover or explain the dataset.

Steps of Analysis

For analysis, we used fsQCA\(^\text{22}\), a software developed by Ragin\(^\text{15}\) for configurational analysis. An implementation of QCA in R\(^\text{22}\) was also used for replication and comparison. The steps are schematically shown below (Figure 1).

![Figure 1. The 6 Steps in our analysis](image)

Variable Selection and Dimensionality Reduction

Since the computational cost of an exact multivariate logical analysis increases according to the number of variables included in the study, the algorithms used with these methods cannot process a large number of variables. The predictor variables in TBI dataset are nominal and multi-valued. When flattened, the total number of variables in the truth table sums up to 36 (including the outcome variable,). An exact analysis of 35 variables and one outcome requires 1.8 Petabytes of memory and could not be analyzed on conventional lab computers at the university. (6th Generation Intel® Core™ i7-6700T Processor (8M Cache, up to 3.60 GHz), 12GB Memory, 1TB hard drive). For this reason, we need to select the most informative variables and consider increasing the granularity of multi-value variables by merging multiple sub-categories.
Results
We employed the binary decision tree algorithm RPART\(^4\), which is an implementation of Classification and Regression Trees\(^5\) (CART) in R, to identify the most informative variables and the cut off point for each multi-level variables. We pruned the resulting decision tree using two different complexity parameters (0.001 and 0.01) and evaluated the predictive power of the resulting models based on the Area Under Curve (AUC). Table 2 compares the AUC of the two models with that of the original CRASH dataset. DeLong\’s test was used to formally compare the ROC curves for the different models.

The 11-var model showed no significant different (Delong p values >0.05) compared with the original model (non-binarized dataset). The alternative hypothesis was that the true difference in AUC is not equal to 0. Even though the dataset represents the same population, the paired ROC test is not applicable for comparing the ROC curves of the binarized models with that of the original model since the models are very different and are deemed to be unpaired by the built-in glm (general linear model) algorithm in R. At AUC 0.8175, the 9-variable model has a higher AUC than sensitivity based (AUC 0.8149) and specificity based (AUC 0.8132) models reported in earlier studies\(^6\). The Delong p-value for the 9-var model is less than 0.05 showing a more significant difference to the AUC of the original model compared to the 11-var model.

We test two models. The first model includes the 9 most informative variables based on the application of RPART, and the second model includes only 7 variables. Variable importance ranking for the two models is given in Table 3.

Table 3. Variable importance based on RPART for two models

<table>
<thead>
<tr>
<th>Variables</th>
<th>motor</th>
<th>verbal</th>
<th>eye</th>
<th>pupils</th>
<th>age</th>
<th>mids</th>
<th>oblt</th>
<th>hmt</th>
<th>sah</th>
<th>ec</th>
<th>phm</th>
<th>sex</th>
<th>cause</th>
</tr>
</thead>
<tbody>
<tr>
<td>9-var Model</td>
<td>30</td>
<td>18</td>
<td>15</td>
<td>13</td>
<td>11</td>
<td>5</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>7-var model</td>
<td>35</td>
<td>15</td>
<td>14</td>
<td>16</td>
<td>11</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Comparative Analysis Using QCA
The first step in an exact analysis of a dataset using the configurational approach is to construct a truth table of variables. Each case in the CRASH dataset will correspond to a row in truth table. A truth table represents a binary tree in which every input variable takes either a zero or one for value. The truth table for our dataset is constructed by replacing for each label \(i\) in variable \(x\) the \(i^{th}\) label of \(x\) with a new variable \(x_i\). This means that multi-level variables are flattened into binary variables by expanding column wise. Given that our dataset includes 6945 cases, and in QCA terms this represents a Large-N analysis, it is unlikely that we can find perfectly sufficient causal combinations. We tested multiple levels and decided to set the sufficiency threshold\(^7\) to 70%. The inclusion cut off point is kept at 1, meaning a single occurrence of a combination is enough to include it for further analysis.

Raw coverage (RC), unique coverage (UC) and consistency (CONS) are the parameters of fit and assess how consistently a combination of conditions appears in the data and the degree to which the findings cover or explain the dataset\(^8\). The dashes (-) in the result tables mean that presence or absence of the variable does not matter for the outcome of that configuration.

Analysis of the 9-Var Model
As shown in Table 4, 67.8% of the Configurations for favorable outcome with a consistency of 84.9% could be explained by 40 combinations. The top 6 configurations for favorable outcome based on this model are reported. The first four conditions cover more cases in the dataset based on their RC and UC. Table 5 shows top 6 configurations for unfavorable outcome.

Table 2. Model Evaluation

<table>
<thead>
<tr>
<th>Model</th>
<th>AUC</th>
<th>95% CI (Delong)</th>
<th>DeLong p*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original CRASH</td>
<td>0.8348</td>
<td>0.8252-0.8444</td>
<td>-</td>
</tr>
<tr>
<td>11-var Binarized</td>
<td>0.8235</td>
<td>0.8136-0.8334</td>
<td>0.1091</td>
</tr>
<tr>
<td>9-var Binarized</td>
<td>0.8175</td>
<td>0.8073-0.8276</td>
<td>0.01504</td>
</tr>
</tbody>
</table>

* Compared with Original CRASH

784
Table 4. Top 6 Configurations for Favorable Outcome Based on the 9-Var Model

<table>
<thead>
<tr>
<th>age</th>
<th>eye</th>
<th>motor</th>
<th>verbal</th>
<th>pupils</th>
<th>obtl</th>
<th>mds</th>
<th>hmt</th>
<th>sah</th>
<th>RC</th>
<th>UC</th>
<th>CONS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 &lt; 45</td>
<td>&lt;</td>
<td>localises or follows commands</td>
<td>&gt; single words</td>
<td>both reactive</td>
<td>no</td>
<td>-</td>
<td>no</td>
<td>0.335</td>
<td>0.028</td>
<td>0.905</td>
<td></td>
</tr>
<tr>
<td>2 &lt; 45</td>
<td>any response</td>
<td>-</td>
<td>&gt; single words</td>
<td>both reactive</td>
<td>no</td>
<td>no</td>
<td>-</td>
<td>0.370</td>
<td>0.014</td>
<td>0.899</td>
<td></td>
</tr>
<tr>
<td>3 &lt; 45</td>
<td>any response</td>
<td>localises or follows commands</td>
<td>-</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>-</td>
<td>0.318</td>
<td>0.000</td>
<td>0.896</td>
<td></td>
</tr>
<tr>
<td>4 &lt; 45</td>
<td>any response</td>
<td>localises or follows commands</td>
<td>-</td>
<td>both reactive</td>
<td>no</td>
<td>no</td>
<td>-</td>
<td>0.414</td>
<td>0.020</td>
<td>0.885</td>
<td></td>
</tr>
<tr>
<td>5 &lt; 45</td>
<td>withdrawal or less</td>
<td>-</td>
<td>both reactive</td>
<td>no</td>
<td>no</td>
<td>yes</td>
<td>0.086</td>
<td>0.015</td>
<td>0.806</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6 &lt; 45</td>
<td>any response</td>
<td>withdrawal or less</td>
<td>&gt; single words</td>
<td>-</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>0.315</td>
<td>0.079</td>
<td>0.872</td>
<td></td>
</tr>
</tbody>
</table>

As shown in Table 5, 42.9% of the Configurations for unfavorable outcome with a consistency of 87.2% could be explained by 63 combinations. The top configurations for unfavorable outcome are reported.

Table 5. Top 8 Configurations for Unfavorable Outcome Based on the 9-Var Model

<table>
<thead>
<tr>
<th>age</th>
<th>eye</th>
<th>motor</th>
<th>verbal</th>
<th>pupils</th>
<th>obtl</th>
<th>mds</th>
<th>hmt</th>
<th>sah</th>
<th>RC</th>
<th>UC</th>
<th>CONS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-</td>
<td>no response</td>
<td>withdrawal or less</td>
<td>Incomp. sounds or no response</td>
<td>no response/unilateral</td>
<td>-</td>
<td>-</td>
<td>yes</td>
<td>0.125</td>
<td>0.009</td>
<td>0.901</td>
</tr>
<tr>
<td>2</td>
<td>-</td>
<td>no response</td>
<td>withdrawal or less</td>
<td>Incomp. sounds or no response</td>
<td>-</td>
<td>yes</td>
<td>-</td>
<td>yes</td>
<td>0.087</td>
<td>0.018</td>
<td>0.889</td>
</tr>
<tr>
<td>3</td>
<td>-</td>
<td>-</td>
<td>withdrawal or less</td>
<td>Incomp. sounds or no response</td>
<td>-</td>
<td>-</td>
<td>yes</td>
<td>yes</td>
<td>0.078</td>
<td>0.024</td>
<td>0.857</td>
</tr>
<tr>
<td>4 &lt; 45</td>
<td>no response</td>
<td>withdrawal or less</td>
<td>-</td>
<td>-</td>
<td>no</td>
<td>no</td>
<td>-</td>
<td>-</td>
<td>0.059</td>
<td>0.001</td>
<td>0.921</td>
</tr>
<tr>
<td>5 &lt; 45</td>
<td>-</td>
<td>withdrawal or less</td>
<td>Incomp. sounds or no response</td>
<td>no response/unilateral</td>
<td>-</td>
<td>no</td>
<td>no</td>
<td>yes</td>
<td>0.030</td>
<td>0.004</td>
<td>0.848</td>
</tr>
<tr>
<td>6 &lt; 45</td>
<td>any response</td>
<td>-</td>
<td>Incomp. sounds or no response</td>
<td>no response/unilateral</td>
<td>no</td>
<td>-</td>
<td>yes</td>
<td>-</td>
<td>0.029</td>
<td>0.002</td>
<td>0.964</td>
</tr>
</tbody>
</table>

Analysis of the 7-Var Model

Since variable importance ranking of hmt and sah are the lowest in the rankings of our classification tree, we removed these two variables to evaluate the resulting configurations without them. The AUC of the 7-var model is 0.811 (95% CI: 0.8136-0.8334 (DeLong)). At DeLong’s p-value 9.474E-04 compared with the original model, the ROC curves of the two models were significantly different. It was found that 57.2% of the cases with favorable

---

1 Withdrawal or abnormal flexion or extension or no response for Motor Response assessment
outcome with a consistency of 85.7% could be explained by 9 combinations. From these 9 configurations in Table 6, we report on the top 6 that have the highest RC and UC.

Table 6. Top 6 Configurations for Favorable Outcome Based on the 7-Var Model

<table>
<thead>
<tr>
<th>age</th>
<th>eye</th>
<th>motor</th>
<th>verbal</th>
<th>pupils</th>
<th>obt</th>
<th>mds</th>
<th>RC</th>
<th>UC</th>
<th>CONS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>&lt; 45</td>
<td>-</td>
<td>localises OR follows commands</td>
<td>-</td>
<td>both reactive</td>
<td>no</td>
<td>no</td>
<td>0.483</td>
<td>0.058</td>
</tr>
<tr>
<td>2</td>
<td>&lt; 45</td>
<td>-</td>
<td>at least single words</td>
<td>both reactive</td>
<td>-</td>
<td>no</td>
<td>0.394</td>
<td>0.037</td>
<td>0.893</td>
</tr>
<tr>
<td>3</td>
<td>&lt; 45</td>
<td>any response</td>
<td>at least single words</td>
<td>both reactive</td>
<td>no</td>
<td>-</td>
<td>0.384</td>
<td>0.025</td>
<td>0.897</td>
</tr>
<tr>
<td>4</td>
<td>&lt; 45</td>
<td>any response</td>
<td>localises OR follows commands</td>
<td>-</td>
<td>-</td>
<td>no</td>
<td>no</td>
<td>0.423</td>
<td>0.003</td>
</tr>
<tr>
<td>5</td>
<td>&lt; 45</td>
<td>any response</td>
<td>localises OR follows commands</td>
<td>at least single words</td>
<td>-</td>
<td>no</td>
<td>0.364</td>
<td>2.3E-4</td>
<td>0.899</td>
</tr>
<tr>
<td>6</td>
<td>&lt; 45</td>
<td>no response</td>
<td>withdrawal or less</td>
<td>at least single words</td>
<td>-</td>
<td>no</td>
<td>no</td>
<td>0.003</td>
<td>0.003</td>
</tr>
</tbody>
</table>

With a raw coverage of 0.48, the configuration of row 1 in Table 6 explains the highest number of favorable outcomes covered by the total model (2740 cases), capturing the configuration “patients (below 45), with motor (localizes OR follows commands) AND pupils (both reactive) AND mds (no) AND obt (no).” This means that regardless of the value of eye and verbal, with 86% consistency, any configuration that matches row 1 results in favorable outcome. On the other hand, 44.5% of the cases of unfavorable outcome with a consistency of 83% could be explained by 20 combinations. Due to space limitation, we only report the top 8 configurations in Table 7 below.

Table 7. Top 8 configurations for Unfavorable Outcome Based on the 7-Var Model

<table>
<thead>
<tr>
<th>age</th>
<th>eye</th>
<th>motor</th>
<th>verbal</th>
<th>pupils</th>
<th>obt</th>
<th>mds</th>
<th>RC</th>
<th>UC</th>
<th>CONS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>45 and above</td>
<td>no response</td>
<td>-</td>
<td>incomp. sounds or no response</td>
<td>-</td>
<td>-</td>
<td>0.200</td>
<td>0.069</td>
<td>0.832</td>
</tr>
<tr>
<td>2</td>
<td>45 and above</td>
<td>-</td>
<td>-</td>
<td>incomp. sounds or no response</td>
<td>no response/unilateral</td>
<td>no</td>
<td>-</td>
<td>0.059</td>
<td>0.009</td>
</tr>
<tr>
<td>3</td>
<td>45 and above</td>
<td>-</td>
<td>-</td>
<td>incomp. sounds or no response</td>
<td>-</td>
<td>no</td>
<td>yes</td>
<td>0.034</td>
<td>0.006</td>
</tr>
<tr>
<td>4</td>
<td>-</td>
<td>no response</td>
<td>withdrawal or less</td>
<td>incomp. sounds or no response</td>
<td>no response/unilateral</td>
<td>-</td>
<td>-</td>
<td>0.239</td>
<td>0.043</td>
</tr>
<tr>
<td>5</td>
<td>-</td>
<td>no response</td>
<td>withdrawal or less</td>
<td>incomp. sounds or no response</td>
<td>-</td>
<td>-</td>
<td>yes</td>
<td>0.161</td>
<td>0.020</td>
</tr>
<tr>
<td>6</td>
<td>45 and above</td>
<td>no response</td>
<td>-</td>
<td>incomp. sounds or no response</td>
<td>-</td>
<td>-</td>
<td>no</td>
<td>0.035</td>
<td>0.001</td>
</tr>
<tr>
<td>7</td>
<td>45 and above</td>
<td>no response</td>
<td>-</td>
<td>incomp. sounds or no response</td>
<td>-</td>
<td>no</td>
<td>yes</td>
<td>0.018</td>
<td>0.001</td>
</tr>
<tr>
<td>8</td>
<td>45 and above</td>
<td>any response</td>
<td>withdrawal or less</td>
<td>both reactive</td>
<td>yes</td>
<td>-</td>
<td>yes</td>
<td>0.009</td>
<td>0.002</td>
</tr>
</tbody>
</table>

With a raw coverage of 0.48, the configuration of row 1 in Table 6 explains the highest number of favorable outcomes covered by the total model (2740 cases), capturing the configuration “patients (below 45), with motor (localizes OR follows commands) AND pupils (both reactive) AND mds (no) AND obt (no).” This means that regardless of the value of eye and verbal, with 86% consistency, any configuration that matches row 1 results in favorable outcome. On the other hand, 44.5% of the cases of unfavorable outcome with a consistency of 83% could be explained by 20 combinations. Due to space limitation, we only report the top 8 configurations in Table 7 below.

Note: Withdrawal or less means: withdrawal or abnormal flexion or extension or no response for Motor Response assessment.
With a raw coverage of 0.2, the configuration of row 1 in Table 7 explains the highest number of unfavorable outcomes covered by the total model; capturing the configuration "patients who are 45 and above with eye = (no response) AND verbal = (incomprehensible sounds or no response)."

Predicting Outcome with QCA

To evaluate the usefulness of the 7-variable QCA model, we compared its ability to predict the TBI outcome with that of a simple binary logistic regression (Logit) model:

\[
P(\text{TBI outcome}) = \text{Logit}(\beta_0 + \beta_1 \cdot \text{age} + \beta_2 \cdot \text{eye} + \beta_3 \cdot \text{motor} + \beta_4 \cdot \text{verbal} + \beta_5 \cdot \text{pupils} + \beta_6 \cdot \text{oblt} + \beta_7 \cdot \text{mul} )
\]

where \(P\) is the predicted probability of TBI outcome based on the assumption of linear relationship between the variables. The purpose of using this simple model for comparison is to show the difference between the results of a conventional additive model with that of QCA. The two models are based on very different assumptions. The linear logistic regression model assigns a weight to all independent variables and is additive in nature. The QCA model takes patterns of interactions between variables into account and outputs multiple combinations.

We compared the predictive power of the two models based on the number of true positives and false negatives they predict as well as their overall prediction accuracy. The results are shown in Table 10. Precision reports the percentage of correct predictions that the model makes. Recall reports the fraction of positive predictions that are truly positive. Accuracy of the model is the percentage of all true predictions from the number of predictions the model makes. One main difference between the two models is that the Logit model generates one model for the whole dataset, but the QCA only explains patterns in a fraction of the dataset.

### Table 10. Predicting favorable outcome in TBI in CRASH dataset

<table>
<thead>
<tr>
<th>Model</th>
<th>Precision</th>
<th>Recall</th>
<th>True Positive Rate</th>
<th>False Positive Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>7-var QCA favorable</td>
<td>2394/2790 = 0.857</td>
<td>2394/4182 (2790)</td>
<td>0.857</td>
<td>0.133</td>
</tr>
<tr>
<td>7-var QCA unfavorable</td>
<td>1232/1483 = 0.830</td>
<td>1232/2763 (1483)</td>
<td>0.830</td>
<td>0.169</td>
</tr>
<tr>
<td>7-var Logit favorable</td>
<td>0.755</td>
<td>3608/4182</td>
<td>0.881</td>
<td>0.447</td>
</tr>
<tr>
<td>7-var Logit unfavorable</td>
<td>0.735</td>
<td>1591/2763</td>
<td>0.553</td>
<td>0.317</td>
</tr>
</tbody>
</table>

* Favorable outcome = Moderate Disability or Good Recovery  
** Unfavorable outcome = Death or Sever Disability at 6 Months

If precision and recall for the two models are calculated based on the number of cases that they claim to explain, the QCA model benefits from higher accuracy:

\[
\text{Accuracy of the QCA model} = (2394+1232) / (2790+1483) = 0.848
\]

\[
\text{Accuracy of the Logit model} = (3684+1527) / (4182+2763) = 0.750
\]

However, when we evaluate the predictive power of each model on the full dataset, the Logit model demonstrates a better precision and recall than the QCA model in predicting outcome, but suffers from higher false positive rates for both cases of unfavorable and favorable outcome. These results highlights the advantages of using QCA particularly when variables that affect outcome positively do not necessarily have reverse effect when they are removed, hence enabling us to highlight the possible asymmetries in the way individual variables can influence the outcome through their participation in configurations.

For the 9-var model, the cases of favourable outcome that QCA did not cover totals 215 different combinations, and for cases of unfavourable outcome that number is 174. For the 7-var model, the numbers are 79 and 59 respectively. Some of these non-covered cases are single occurrences of the configuration of variables that could not be factored with other configurations.
Discussion

Our study demonstrated a different approach to evaluating predictors of clinical outcome in TBI. With methods of QCA we established multiple configurations for admission variables that are predictive of favorable versus unfavorable outcome. Most of the findings are intuitive, young age (<45), good neurological condition and lack of CT abnormalities are in keeping with favorable outcome. Whereas older age, poor neurological condition and CT findings such as mass effect or traumatic subarachnoid bleed are suggestive of an unfavorable outcome. These results are in line with previous studies, however an unexpected finding was that on formal variable importance ranking using RPART age fell behind the GCS components as well as pupillary response. This is further traceable in several of the configurations (1-3 Table 5) for unfavorable outcome where age does not appear. A further finding in our study is the dichotomization values for admission variables which we established using a binary decision tree algorithm (RPART). Binary adaptation of clinical features is appealing to clinicians because it simplifies patient assessment particularly in the emergency setting. We have demonstrated that collapsing multi-level variables into binary does not impact model performance when maintaining the full set or most of variables present in the original model (Table 2). Consequently, a binary model can potentially inform a simplified assessment protocol without substantial loss of clinical information.

The configurational asymmetric models uncovered through the application of set-theoretic methods such as QCA, make these methods appealing when there is a possibility of interactions between variables. Comparative analysis with QCA is receiving increasing attention among researchers from a variety of disciplines such as social science24, business and economics25-30, management and organization31, education32, and health policy research33, 34.

For analyzing the CRASH dataset with QCA, we made the models more parsimonious by removing some variables since the complexity of the dataset does not allow us to apply the exact procedure of QCA to the full dataset, or even the 11-var dataset. As a limitation, we forced the multi-level variables into dichotomies based on their first split in the classification tree. This first split is considered to be most informative for contrasting outcomes, however, with more granularity we might find different and possibly better results. These binary models however showed similar performance to models built with multi-level variables. A translational value of our findings is that the configurations of admission variables can be regarded as “typical” patient scenarios that are strongly predictive of a clinical outcome.

Conclusion

A configurational, asymmetric model of TBI outcome is investigated. We have demonstrated that the dichotomization of admission variables can provide basis of simplified assessment protocols that can usefully implemented in small centers for example without specialist capacity. From a clinician’s perspective is also useful to be presented with a set of “typical” scenarios that are suggestive of favorable versus unfavorable outcome. We evaluated the predictive power of a simple logistic regression model and a QCA model using the CRASH dataset. The Logit model demonstrates a better precision and recall than the QCA model in predicting outcome, but suffers from higher false positive rates for both classes of unfavorable and favorable outcomes. The QCA model only explains patterns in a fraction of the dataset while the Logit model attempts to cover the whole dataset. We are currently investigating a new heuristic based on logic synthesis and network analysis methods to overcome the inherent limitation in QCA in terms of the number of variables and the complexity of the dataset while automating the inclusion of interaction terms to develop minimal models that are maximally predictive.

References


30. Glaesser J, Cooper B. Selectivity and flexibility in the German secondary school system: A configurational analysis of recent data from the German socio-economic panel. European Sociological Review. 2010


789
Using Process Mining Techniques to Study Workflows in a Pre-operative Setting

Adela Grando, PhD1, Danielle Groat, MS1, Stephanie K. Furniss, MS1,2, Joshua Nowak, MS1,3, Regina Gaines MSc1, David R. Kaufman, PhD1, Karl A. Poterack, MD4, Tim Miksch, MBA2, Richard A. Helmers, MD5

1Department of Biomedical Informatics, Arizona State University, Phoenix, AZ
2 Mayo Clinic, Rochester, MN
3Department of Biomedical Engineering, Arizona State University, Phoenix, AZ
4 Department of Anesthesiology, Mayo Clinic, Phoenix, AZ
5Mayo Clinic, Phoenix, AZ

Abstract

Information technologies have transformed healthcare delivery and promise to improve efficiency and quality of care. However, in-depth analysis of EHR-mediated workflows is challenging. Our goal was to apply process mining, in combination with observational techniques, to understand EHR-based workflows.

We reviewed nearly 76,000 event logs from 15 providers and supporting staff, and 142 patients in a pre-operative setting and we inspected 3 weeks of interviews and video observations. We found that on average 44 minutes were spent per patient interacting with the EHR, 55% of the time of the patient visit was spent by personnel interacting with the EHR and for over 5% of the time personnel used or reviewed paper-based artifacts. We also discovered the handover-of-care network and compared frequency of interactions between personnel.

This study suggests that applying process mining in combination with observational techniques has vast potential for informing Mayo Clinic in the forthcoming EHR conversion.

Introduction

Current health information technologies (HIT), such as electronic healthcare records (EHRs), aim to improve healthcare delivery through increased access to patient information and better organization. However, recent HIT implementation and use have yielded mixed results in different clinical settings. Problems with EHR integration with existing clinical work processes impact workflow. This has resulted in extending the time required to complete tasks and has also resulted in the development of workarounds which can result in adverse events that compromise patient safety or quality of care delivered. With these new problems arising, we need a meaningful way to record and analyze clinicians’ interactions with HIT. This need coincides with Mayo Clinic’s interest to collect and analyze current workflow on clinician’s interactions in order to facilitate a large-scale EHR conversion.

In advance of an upcoming enterprise wide conversion to a converged EHR, the Mayo Clinic launched the ROOT (Registry Of Operations and Tasks) project, which aims to understand EHR-mediated workflows and clinician-HIT interaction. Surgical settings have been prioritized as the testbed for this project. Surgery is one of the most clinical and resource intensive activities in health care, constituted by multiple care transitions and a high incidence of adverse events. A main goal of the ROOT project is to create a repository of current high-definition snapshots of clinical and informational workflows at Phoenix, AZ, Rochester, MN, Jacksonville, FL, and in the Mayo Clinic Healthcare System before the transition to a new EHR.

Process mining techniques describe a family of a-posteriory automatic analysis techniques that exploit event logs, which record information about the sequences (traces) of events (activities) executed using electronic information systems. While process mining has been extensively used in business management, medical process mining is a relatively newer research field. In our prior research, which laid the methodological foundation for this work, we investigated variations in clinicians’ EHR workflow by integrating quantitative analysis of patterns of users’ EHR-interactions with in-depth qualitative analysis of user performance. This paper summarizes preliminary outcomes from the ROOT project resulting from observations, interviews, video ethnography and process mining analysis at the pre-operative (PreOp) setting at Mayo Clinic Hospital, Phoenix, Arizona.

Objectives

The objective of this study was to employ an integrative methodological approach leveraging process mining, observations, interviews and video ethnography, to understand patient workflows, clinical workflows, and interactions of personnel with electronic systems at the PreOp setting of one of the Mayo Clinic facilities. Four questions were posed to guide this research: Can we validate observed handover-of-care based on PreOP personnel’s sequential
interactions with patient cases?; How much time do personnel spend in electronic information systems?; How much time do personnel spend on documenting and/or reviewing data using paper based artifacts vs. electronic systems?; Is there a correlation between care paths followed and the time spent in the PreOp by different patient groups?.

Methods

We integrated outcomes of clinical observations, interviews or video ethnography with process mining techniques to increase our study sample and, in some cases, validate outcomes. The interviews, observations and videos guided the process mining and helped us to identify relevant questions to be answered by process mining techniques in order to better understand clinical and patient workflows. In turn, the process mining can be used to discern the distribution of patterns evident in our workflow.

For this study, two different teams collaborated to carry out different data gathering techniques and analysis.

Observations, Interviews and Video Ethnography

The first team conducted mixed-method data collection on-site over a 3-week period in 2016. Data collection methods included semi-structured interviews and multiple observational techniques to surface specifics about all roles and activities performed in PreOp. In this paper, we refer to the following set of methods collectively as “observational techniques”: shadowing, video ethnography using hand-held video camera and Morae™ video capture. Morae™ software is used for usability studies and it records user activity with no interruption to the user’s work. Using a webcam, it was also used to record audio of participants verbalizing their thoughts and video recording of the participant’s hands and desk space. Participants were asked to think-aloud as they performed HIT-based tasks to reveal activity, goals and cognitive processing, and allow for retrospective analysis.

The mixed-method approach was used by the team to identify, describe and validate the clinical workflow components and processes. The study presented here, draws on a subset of this analysis for an in-depth description of patient flow, personnel, and personnel’s coordination processes integral to efficient patient flow.

Process Mining Analysis

The second team was responsible for collecting and analyzing, using process mining techniques and tools, the event logs generated at the PreOp during two days that the first team was on-site. The team had access to nearly 76,000 event logs generated during the selected days. For those days, there was a daily average of 15 providers and supporting staff, and 71 patients.

The event logs mainly captured actions performed within the Cerner SurgiNet environment. SurgiNet is the EHR application used for tracking cases in the perioperative settings (i.e. pre-op, intra-op and post-op) and for most perioperative documentation. In addition, when applications outside the EHR system such as Synthesis, Smart Template Wizard, Schappt Book and Outlook email were launched, there were few if any event logs created.

![Figure 1: All of the attributes provided in a Cerner event log are split into two rows. The columns of interest in the first row provide the ending timestamp of the event (End Date, End Time), duration of the event (Elapsed Time), name of software function that triggered the event (Timer Name) within an information system (Exec Name), and the identifier of the provider (User Name) using the software. The second row contain the identifier of the patient (Metadata 1) that is related to the event and additional information regarding the software’s function (Subtimer).](image-url)

Each row in the log file was considered to be an event. Each event in the log file included 16 attributes: 1) End Date: date of the log file event; 2) End Time: accurate to the second when the log file event was completed; 3) Node Name: networking information; 4) Timer Name: high-level description of the log file event; 5) Elapsed Time: amount of time in seconds for the system to execute the log file events; 6) Pass: unknown; 7) Exec Name: name of the execution software; 8) Process ID: unknown; 9) Thread ID: unknown; 10) Username: provider identifier; 11) Client PC: workstation identifier; 12) Client IP: networking identifier; 13) Fail Code: unknown; 14) Subtimer: additional background regarding the log file event; 15) Result Count: unknown; 16) Metadata 1, 2 and 3: may contain a patient identifier; the patient identifier is not related to the patient medical record number or any other identifier that could be used to re-identify the patient. We provide in Figure 1 an example of an event in a log file. From all the information...
contained in the log files the most relevant for process mining analysis was end date and time, elapsed time, timer name in conjunction with the subtimer, username (provider identifier) and the metadata (patient identifier).

The Mayo Clinic also made available an event log dictionary that provided, for each potential event log that could be collected, a description of its interpretation. To better understand and associate users’ behavior and clinical workflow, the second team reviewed selected portions of the Morae™ video recordings and closely collaborated with the first team to generate the results reported here.

Morae™, Microsoft Excel, Java, and Disco™ were used for preprocessing and cleaning the EHR event logs. While the process mining analysis presented in this paper is based on two of the observed days, the full data set of Morae™ video recordings and researchers’ observations, were reviewed for log event analysis.

In Excel, event logs were sorted based on date and time. Files were converted from .csv to .xls format. The .xls files were then cleaning using a Java program. Once the files were cleaned, they were manually merged with Excel. Files were merged according to days so that multiple personnel were all in the same file. This organization allows for the process mining analysis to be done from either the personnel or patient perspective.

Once the .xls files were organized by days, they were imported into Disco™ and exported as XES ProM files. Disco™, a processing mining and automated process discovery software, was used for additional data cleaning and file conversion. The resulting XES files were imported to ProM for analysis. ProM is an open-source, free software, considered the “de-facto” process mining software. It contains hundreds of plug-ins (algorithms) for process mining analysis. Several ProM plug-ins were used to further refine the dataset for analysis and produce visualizations.

**Results**

First, we present outcomes from the application of observational techniques, in particular descriptions of PreOp personnel roles and responsibilities, patient-based workflows and personnel’s coordination process integral to patient flow in PreOp. Then, we build on those outcomes to apply process mining techniques to address three relevant questions posted by the observational team: Can we validate observed handover-of-care based on PreOP personnel’s sequential interactions with patient cases?; How much time do personnel spend in electronic information systems?; How much time do personnel spend on documenting and/or reviewing data using paper based artifacts vs. electronic systems?; Is there a correlation between care paths followed and the time spent in the PreOp by different patient groups?.

**Personnel and their interactions, captured through observational techniques**

PreOp relies on a number of personnel and roles to ensure safe patient care delivery and timely preparation for surgery. Roles specific to PreOp include the health unit coordinator (HUC), patient care associate (PCA), pre-operative nurse (RN), float nurse (Float RN), Desk nurse (Desk RN), and Team Lead (TeamLead RN).

HUCs are non-clinical staff who are stationed at the check-in desk in the Surgery Waiting Room. For patients arriving to the hospital on their day of surgery, the HUC station is their first stop. The HUC facilitates communication between patient’s family and clinicians during the patient’s perioperative care (i.e., pre-op, intra-op and post-op). For example, the HUC is responsible for notifying the Desk RN when the patient has arrived in the Surgery Waiting Room.

The PCA is the patient’s second point of contact in PreOp. PCAs guide the patient from the Surgery Waiting Room to the patient’s PreOp room, orient the patient to the room, and collect some of the patient’s vital signs measures (written on paper). For some patients, PCAs may help with surgery preparation activities. PCA’s do not use HIT in support of patient care.

Each day, RNs are assigned two PreOp Rooms and are responsible for ensuring the Nurse Assessment, preparatory care and associated documentation are complete for each patient assigned to those rooms. Activities include: validate the patient and surgical procedure, activate anesthesia, surgery and nursing orders, reconcile medications, administer medications, validate procedure and signature on the informed consent document, validate presence of a recent history and physical note is documented, ensure pre-operative testing is completed, coordinate outside resources, complete assessment questions with the patient (e.g., neurological assessment), collect patient contact information and post-surgery transportation plans and others. The RN’s last activity for each patient, is to complete a face-to-face verbal handoff to members of the patient’s surgical team, who then transport the patient to the operation room.

Each day, several nurses fill the role of Float RN. Unlike RNs, Float RNs are not assigned to rooms or patients. Instead, they provide assistance to the RN as needed. The assigned RN is responsible for the RN Assessment and required pre-operative documentation. Therefore, the Float RN will assist with non-documentation activities allowed by their training (e.g., place IV catheter and administer IV fluids).

The Desk RN and TeamLead RN roles are each filled daily by a nurse who is more experienced in the unit. The Desk RN manages patient flow through the PreOp unit and day-of-staffing schedules. Responsibilities include, assign patients to PreOp Rooms with consideration to the RN’s workload and the patient’s scheduled surgery start time,
identify patients who may not be prepared for the scheduled surgery start time and communicate with the surgical team (e.g., the patient arrives late to the unit), and coordinate patients and PreOp staff in response to emergent changes to the surgery schedule (e.g., contact patients to notify of change). Experience is important because the Desk RN needs to anticipate changes to the surgery schedule (e.g., based on a surgeon’s anticipated efficiency in prior procedures, the surgeon is likely to ask for a patient to be prepared of surgery earlier than scheduled) and resource constraints (availability of RNs and Float RNs). The TeamLead RN role is typically filled by one experienced RN. When this RN is not working, another experienced RN fills this role. Otherwise, the TeamLead RN’s daily responsibilities are similar to a Float RN where they assist PreOp RNs with safe and timely preparation of the patients for surgery.

*Patient-based workflow, captured through observation techniques*

Table 1 provides a schematic summary of observed patient-workflow in terms of patient’s physical location in PreOp setting, personnel involved in patient care, artifacts used (paper-based or electronic) and SurgiNet perioperative tracking status. The patient’s first interaction with the PreOp unit, is at the Surgery Waiting Room on the hospital’s second floor. Prior to this, upon arriving at the hospital, the patient checks in at the Admission Desk. In the Surgery Waiting Room, the patient checks in with the HUC at the check-in desk. Patients are asked to arrive two hours before their scheduled surgery. To inform other PreOp personnel that the patient has arrived, the HUC uses Perioperative Tracking in SurgiNet. Specifically, the HUC changes the patient’s tracking status to “Arrival Waiting Room”.

The Desk RN monitors the Perioperative Tracking display, which most often remains displayed on one of the person’s desktop monitors. When the Desk RN sees the indication that the patient is in the Waiting Room, the Desk RN retrieves the patient’s prepared chart from a file rack behind him/her, which holds prepared paper charts (PC) for all surgical cases planned for the day. The Desk RN interacts with the Perioperative Tracking display at SurgiNet to select the room and to type the name of the RN assigned to that patient case. To make the room assignment, the Desk RN views PreOp RN staff and the availability of their assigned rooms. The room number is also written on the front of the patient’s chart and on a small purple paper attached to the chart. PreOp RN staffing and their assigned rooms are viewed and managed on a locally-developed screen view, Staff Scheduler, that is always displayed on the Desk RN’s second desktop monitor.

### Table 1. Observed correlation of patient’s physical location in PreOp setting, personnel involved in patient care, artifacts used (paper-based or electronic) and SurgiNet perioperative tracking status.

<table>
<thead>
<tr>
<th>Patient’s Location</th>
<th>Surgery Waiting Room</th>
<th>PreOp Rm#</th>
<th>Operating Room</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Personnel/ Role</strong></td>
<td><strong>HUC</strong></td>
<td><strong>Desk RN</strong></td>
<td><strong>PCA</strong></td>
</tr>
<tr>
<td><strong>PreOp Location</strong></td>
<td>Surgery Waiting Room</td>
<td>Desk RN work space</td>
<td>Desk RN work space, Surgery Waiting Room, PreOp Rm</td>
</tr>
<tr>
<td><strong>Artifacts Used</strong></td>
<td>SurgiNet; Paper logbook</td>
<td>Staff Scheduler; SurgiNet; Patient chart (w/ stickers, purple paper)</td>
<td>Patient chart (w/ stickers, purple paper)</td>
</tr>
<tr>
<td><strong>Other SurgiNet updates</strong></td>
<td>“PreOp Rm#”</td>
<td>“[RN name]”</td>
<td></td>
</tr>
</tbody>
</table>

When a chart becomes available in the file rack, a PCA picks up the chart on her way to get the patient. The PCA meets the patient in the Surgical Waiting Room and guides the patient to the assigned PreOp room. On the way, they
stop at a scale and measure to get the patient’s height and weight, which the PCA records on the purple paper. To signal to the Desk RN that the patient is on the way to a room, the PCA takes a patient identification sticker off the chart and leaves it with the Desk RN. This triggers the Desk RN to change the patient status in Perioperative Tracking display in SurgiNet from “Arrival Waiting Room” to “In PreOp”. The patient is placed in the PreOp room one or two hours before the surgery.

In the PreOp room, the PCA collects additional patient vital signs (e.g., blood pressure, temperature, etc.). To signal to the RN when the patient is in the room, the PCA moves the rolling table in the room to the hallway just outside the room, and places the patient chart on the table. Before entering the room, the RN will often get the patient’s paper chart from the table and access patient data in the PowerChart to assess pre-operative care needs and initiate processes, such as activating surgical, anesthesia and nursing orders. The duration, number and type of system interactions the RN performs to complete the Nurse Assessment, preparatory care and associated documentation for each case, are affected by a number of factors, such as the patient’s acuity, surgical service and surgical procedure.

When finished with the RN Assessment and pre-operative care and the patient is ready to go to the operating room, the RN changes the patient’s Perioperative Tracking status in SurgiNet to “RN Assess Comp”. If the patient would like to see family who are waiting in the Surgery Waiting Room, the RN changes the patient’s status in SurgiNet to “Family” to signal to the RN when the patient is in the room. The HUCs monitor their Perioperative Tracking display for the brown “Family” status. When they see this, a HUC guides the family to the patient’s room. Upon returning to their desk, the HUC removes the “Family” status event from the Perioperative Tracking display, which returns the display back to the message “RN Assess Comp”.

After the RN completes a face-to-face verbal handoff with members of the patient’s surgical team, they transport the patient to the assigned operating room for surgery. On their way past the Desk RN, one of the surgical team members drops off the purple card with the Desk RN. This transfer of the purple card is used to signal to the Desk RN that the patient is leaving the PreOp unit. After several minutes have passed and the patient’s Perioperative Tracking status reads “In Surgery”, the Desk RN removes “In PreOp” from the patient’s tracking list. Of note, the purple card has the patient’s vital signs on it that were hand-written by the PCA. The Desk RN is responsible for entering these data into the patient’s chart in PowerChart.

Provider and patient-based workflows captured through process mining

Using a combination of process mining techniques, and outcomes from observational data collection we endeavored to answer four questions relevant for workflow analysis:

1. Can we validate observed handover-of-care based on PreOP personnel’s sequential interactions with patient cases?

We examined how an analysis of team interactions using the event logs compares to the above observational-based description of patient-based workflow. To achieve this, we created a sequential visualization and performed a social-network analysis of the event logs where team interactions are measured by personnel’s sequential access to a patient chart. For example, in the sequence of events associated with a patient chart, if there is an event associated with the HUC followed by an event associated with the Desk RN, then the sequential visualization would show a dot associated to a HUC preceding a dot associated to the Desk RN role, and the social-network analysis would identify a HUC-to-Desk RN interaction for this patient case.

We used the dotted chart PROM plug-in to automatically discover and visualize the sequence of personnel’s interactions for each patient case for Days 1 and 2. Dotted chart for Day 2 is shown in Figure 2. Of note, each row is for a single patient and the dot colors are associated to the personnel’s role. The HUC and the Desk RN usually generate the first event log for each patient. These events are often followed by an RN. This is consistent with the observed workflow.

The handover-of-work social network PROM plug-in provides an alternative visualization of team interactions on Days 1 and 2, in which the role-to-role interactions with SurgiNet are visualized as a relative probability. The event logs were filtered to contain four events types that best represented user interactions with SurgiNet for patient documentation or information review activities. These four events were among the 15 most-frequently occurring events for both Days 1 and 2. For example, “Build SDL” event represents personnel clicking check boxes on the screen when completing documentation or reviewing information, and “Load Set Events” event represents personnel changing the PreOp tracking status of the patient. The other two events, “Display Plans” and “Start Chart Search” indicate that the provider/staff has accessed a patient’s care plan and chart, respectively.

In Figure 3, the social network of handover-of-work analysis shows that generally all the roles reciprocate handover-of-work to one another, except for the TeamLead RN who only interacts with the Desk RN. The most common hand-offs occurred from the HUC-to-Desk RN and then back and forth between the Desk RN and the RNs (i.e., Desk RN-to-RN and RN-to-Desk RN). The next most common interactions occurred from the Desk RN–to-HUC, back-and-forth between the HUC and the RNs (i.e., HUC-to-RN and RN-to-HUC). Also, consistent with
observations there were fewer event logs generated by float RNs as they generally assist with non-documentation related care.

Figure 2: Dotted chart for Day 2. The x-axis is divided into hours of the day and each row represents a patient. The colored dots represent log file events from all the type of providers in PreOp. The colors are based on provider roles.

Figure 3: The social network of handover-of-work shows that generally all the roles reciprocate handover-of-work to one another except for the Team Lead RN only hands-off work to the Desk RN. The most common hand-offs occur from the HUC-to-Desk RN and then bi-directionally between the Desk RN and the RNs. Circle size correlates to the number of event logs associated with the role—a larger circle size indicates that role generated a larger number of event lows. Arrow thickness correlates to the number of interactions. A larger arrow indicates that interaction occurred more frequently when compared to other interactions.
2. How much time do personnel at PreOp spend using electronic information systems?

To answer this question, two things had to be quantified for each patient: a) how much time personnel spent in electronic information systems, and b) how much time personnel spent with a patient.

After manually reviewing event logs, we found that SurgiNet was almost exclusively the executed program recorded in the provided event logs. There were few events outside SurgiNet. For example, on Day 1, there were only 40 events of ~38,000, or 0.1%, not associated with SurgiNet. Two other programs identified, other than SurgiNet, were Smart Template Wizard and Schappt Book. We were not able to interpret these events, as they did not coincide with the Morae video capture. On the other hand, using Morae video recordings, we could associate an event from Day 2 labeled “web launch” with the launch of Synthesis. Synthesis is a home-grown system to facilitate the navigation of the EHR. It offers users a single, integrated view of information from many disparate sources. This log file event occurred an additional 11 times on Day 2 and only once on Day 1. These other “web launch” events from Days 1 and 2 did not coincide with Morae™ video recordings; therefore, we could not assess if this event name always referred to the user accessing the Synthesis system. Also, there were instances in the Morae™ video recordings where users launched other systems or tools, such as web browsers or email, and these events were not identifiable in the event logs. As a result, we concluded that event logs do not sufficiently identify when users access electronic systems outside SurgiNet; therefore, the event logs could not be analyzed to measure the amount of time personnel spent on other clinical information.

To quantify how much time personnel spent in electronic information systems, we assumed that all activity in the event logs was attributed to time spent using SurgiNet. We analyzed the amount of time the personnel spent in the system per patient by generating patient-centered workflows using ProM. The data from Days 1 and 2 was visualized from the patient perspective with the dotted chart PROM plug-in (Figure 4). From the 142 identified patients we selected a subset of 60 patients. For instance, we removed patients when we identified that personnel spent less than 15 minutes interacting with the information systems to review information of that patient. For the given example, the assumption was that the personnel were accessing patient charts for individuals not scheduled for the surgery that day. For each of the 60 patients we created the corresponding dotted chart and we discovered that the average amount of time a RN spent interacting the patient’s chart using SurgiNet was 44 minutes. On average the time the patient spent in the PreOp room, excluding time in the waiting room and waiting time for surgery, was 3 hours and 33 minutes and 27.2% of that time was spent by personnel accessing SurgiNet.

Figure 4: Dotted chart of Day 1. The x-axis is divided into hours of the day (4:00am to the end of the day is displayed) and each line represents one of the 77 patients. Each colored dot represents one of the 39,363 log file events from all the personnel and the colors are based on 274 unique even types. For instance, dark purple represents events where the doctor read clinical notes and light green corresponds to an ICU band load.
The time spent by personnel with a patient could be different from the time spent in electronic information systems. To quantify the time spent by personnel with a patient we excluded event logs generated by HUC, Desk RN and other personnel to eliminate variability in patient arrival times with respect to the PreOp schedule. The dataset was reduced to only include RNs and patients for whom we had Morae video recordings available, which was 6 patient cases. For that subset, total time in the system was found looking at the difference between the first event log from the RN and the last event log from the RN with respect to the patient of interest. This was assumed to closely represent the amount of time a patient spent in a PreOp room before being wheeled into surgery and did not include time spent in the waiting room. For the patient of interest, activity groups associated to the RN role were defined as two or more events that occurred within close temporal proximity to each other. Gaps of 5 minutes or more in the event logs indicated the beginning of a new activity group. The start time from the first event log in an activity group was subtracted from the end time of the last event log from the same activity group, to determine the total amount of time the RN spent reviewing/updating patient information in SurgiNet. The duration of each activity group was considered an approximate representation of the time the RN spent using SurgiNet, which was then compared to the total time in the system. We found that the average amount of time a patient spent in the PreOp Room, excluding time in the waiting room and waiting time for surgery, was 2 hours and 12 minutes. One hour and 12 minutes (55%) of that time was attributed to RNs using SurgiNet to record information relevant to the patient’s clinical case.

3. How much time do personnel at PreOp spend on documenting and/or reviewing data using paper based artifacts?

In Question 1, we used event logs to quantify the amount of time a patient is in PreOp and amount of time personnel use the computer-based artifact, SurgiNet. Morae™ video analysis was required to identify how much time personnel spent per patient reviewing and/or documenting with paper artifacts. The RN spent an average of 6 minutes and 40 seconds using paper-based artifacts per patient, which accounted for 5.2% of the average total time a patient is in PreOp. The remaining time (40%) was not spent using paper-based artifacts or interacting with SurgiNet.

4. Is there a correlation between care paths followed and the time spent in the PreOp by different patient groups?

In our case, patient groups could be defined in terms of several factors, such as the patient’s acuity, surgical service and surgical procedure. By filtering the event logs based on common PreOp patient tracking events described in Figure 1 (e.g., those associated with check-in and check-out events), it is possible to estimate how much time a patient spent in the PreOp unit. The amount of time each patient spent in PreOp was determined by first visualizing the filtered event logs (i.e., a sub-set that only contains the event type associated with tracker status update events for the patients with the dotted chart analysis), see Figure 4. The data was cleaned to only include personnel and patients for whom we had Morae video analysis available. Then the dotted chart plug-in was used to extract the time spent in PreOp in the same fashion that was done to find the amount of time personnel were in SurgiNet as in Figure 4. While we could estimate the care paths and time spent by patients at PreOp, correlating them with the patient’s clinical case requires a manual review of the patient’s chart, which is outside the scope of the methodological approaches chosen for the ROOT project. Therefore, while very relevant, this question remains unanswered.

Discussion

The application of process mining into healthcare is a relatively new, but growing area of research. Zheng et al investigated user’s interaction with an EHR by uncovering hidden navigational patterns in EHR event logs. They identified repeated feature access in successive action sequence. Various patterns deviated from optimal clinical management pathways. Kannampallil et al used event logs to compare the information-seeking strategies of clinicians in critical care settings. Specifically, they characterized how distributed information was searched, retrieved and used during clinical workflow. They found that there are costs (e.g., effort, time and cognitive load) associated with particular strategies. In our prior work, process mining was applied to characterize clinician’s patterns of information gathering. Kaufman et al. proposed the novel use of process mining techniques in combination with qualitative methods to help discover, analyze and visualize records of human computer interactions to lead to improved EHR designs. Applying similar methods, Furniss et al. found that participants’ EHR-interactive behavior was associated with patient case complexity, preferred information sources, and EHR default settings.

A 2016 review by Rojas et al. found 74 papers associated with the use of process mining in healthcare domains. In 2013 a review of the state of the art by Mans et al. indicated that only 40 publications referred to the use of process mining in real clinical contexts. From those, only one publication focused on the use of process mining in the context of PreOp processes. The review surfaced the five questions that are most frequently asked when process mining techniques are used in the healthcare context: Q1) What are the most commonly followed paths and what exceptional paths are followed? Q2) Where are the bottlenecks in the process? Q3) Is there compliance with external and internal guidelines? Q4) Are there any differences between care paths followed by different patient groups? Q5) What are the roles and social relationships between medical staff?. Besides the questions addressed in the Result section, the ROOT project also attempted to answer frequently asked questions Q1-Q5. A central goal of our research
was to apply a hybrid methodology for addressing similar questions. Below we explain in detail successes and challenges encountered trying to answer those questions. We also describe how questions that could not be fully addressed in this study can be investigated in future research.

Questions Q1-Q3: unfortunately, the event logs available in the ROOT project were created to quantify performance measurements. This constraint makes it difficult to accurately infer the clinical context of the related workflow activities. Therefore, applying process mining techniques to the available event logs did not lead to answering questions Q1-Q3.

Question Q4: the research teams wanted to know if there was a correlation between the care paths followed and the time spent in PreOp by different patient groups. Understanding patient’s clinical case would require having access to the patient’s chart. The ROOT project did not envision review of the patient’s medical records, and therefore we could not answer this question.

Question Q5: We successfully answered this question by validating results from observational techniques with the discovered handover-of-care (Figure 3). Consistent with observations, the social network shows that generally the roles reciprocate handover-of-work to one another except for the Team Lead RN who only hands-off work to the Desk RN. In addition, we compared the frequency of interactions between role types. The most common hand-offs occur from the HUC-to-Desk RN and then bi-directionally between the Desk RN and the RNs. This analysis could be replicated considering individual personnel or groups of personnel.

In the process of trying to answer frequently asked questions Q1-Q5 the ROOT project encountered some limitations:

- **Event logs were created for performance analysis and not for workflow analysis:** we only had access to event logs being recorded for tracking the performance of electronic information systems (e.g., system response time). Example: time required to load a new screen, time required to register a click in a box in a screen, time required to close a screen, etc. The event logs that we had access to did not contain the level of granularity and event types needed to answer some relevant clinical workflow questions posted by the research team, or to answer the four questions chosen in the Objective section using only process mining techniques. It would have been desirable to have access to event logs related to personnel activities as described in the observed provider-based and patient-based workflows. Example: check-in, assessment, upload vital signs, etc. What we have learned will help to better identify the type of event logs that we wish to collect in the new EHR ecosystem to better understand and discover clinical workflows and to guide workflow improvements.

- **Lack of access to event logs from all electronic information systems used:** We only had event logs from a single information system used in PreOp unit, the EHR-based SurgiNet system. While it was the primary system used for patient documentation and coordination activities in this setting, RNs referenced other information systems to complete key patient care activities. In addition, when applications outside the EHR system such as Synthesis, Outlook, Smart Template Wizard and Schappt Book were launched, there were few if any event logs created. As a result, we could only approximate the total time personnel spent with all HIT systems (Question 1). We are in the process of understanding the implications of these limitations.

- **Lack of access to patient’s medical records:** because the ROOT project has been framed as a quality improvement research, we had no access to EHR data for the observed patients. Access to this information could have provided additional insights on the discovered patient-based workflows.

The challenges and successes encountered while answering the questions described in the Result section and questions Q1-Q5 helped the ROOT team to identify lessons learned:

- Access to event log dictionaries was a very valuable initial resource to help interpreting logs
- Review of observational data could be time consuming but essential to further interpret event logs and outcomes from process mining. The need to access to observational data is based on the limitation that event logs available for the ROOT project were created for performance analysis, instead of for workflow analysis.
- Mine both patient-centered and provider-centered workflows provided complementary insights for the clinical workflow analysis.
- Understand differences between the type of events captured through the enterprise, as oppose to the type of desirable events could guide future event log capturing.

Conclusions and Future Work

The Mayo Clinic is embarking on a large-scale effort to standardize workflows across its four geographical clinical networks, and thus move from multiple disparate EHR platforms to a single EHR system.

Here we reported preliminary results that demonstrate the value of applying process mining techniques, in combination with observational techniques, to better understand clinical and electronic workflows before they
disappear due to the transition of Mayo Clinic to the new EHR. Most importantly, lessons learned on the limitations encountered during the process of mining existing log events to answer relevant workflow questions will guide future recommendations on type of information and levels of granularity required to be captured by log events to be generated by the new Mayo Clinic EHR.

We have conducted similar multi-method data collection observational techniques at four additional Mayo Clinic locations. In addition, we have completed and we are in the process of analyzing 35 interviews (21 of them involving leadership) across the enterprise. We plan to apply the methodologies described in this paper to answer similar questions, and to compare results between locations. Similarities and differences between identified workflows could inform the upcoming process of workflow standardization and future process improvements across the Mayo Clinic enterprise.

We suggest that determination of event types in system-generated event logs is made with consideration for clinical workflow analysis, such as for the goals of the ROOT project.

Acknowledgement
Mayo Clinic, for financial support through BMS0148.

References
Sharing Clinical Notes with Hospitalized Patients via an Acute Care Portal

Lisa V. Grossman¹, Ruth Masterson Creber, PhD, RN²,
Susan Restaino, MD³, David K. Vawdrey, PhD⁴

¹Department of Biomedical Informatics, Columbia University, New York, NY;
²School of Nursing, Columbia University, New York, NY;
³Department of Medicine, Columbia University, New York, NY;
⁴Value Institute, NewYork-Presbyterian Hospital, New York, NY

Abstract

Though several institutions offer hospitalized patients access to their medical records through acute care patient portals, no studies have assessed the potential impact of patients’ access to physicians’ notes through these systems. We employed a mixed-methods approach, including patient surveys, system usage log analysis, and qualitative interviews, to describe patients’ perspectives on receiving their clinical notes and usage of the clinical notes feature in an acute care patient portal. Patients visited the clinical notes feature more frequently and for longer durations than any other feature. In qualitative interviews, patients reported improved access to information, better insight into their conditions, decreased anxiety, increased appreciation for clinicians, improvements in health behaviors, and more engagement in care. Our results suggest that sharing notes with hospitalized patients is feasible and beneficial, although further studies should investigate the magnitude of benefit and explore the unintended negative consequences associated with increased transparency of clinical information.

Introduction

Increasingly, healthcare institutions offer inpatient access to the medical record through acute care patient portals. Several medical centers, including Brigham and Women’s Hospital, NewYork-Presbyterian Hospital / Columbia University, Michigan Medicine, the Ohio State University Wexner Medical Center, Northwestern Memorial Hospital, and the Veteran Affairs Medical Centers have trialed or currently maintain acute care patient portals. The Libretto Consortium is composed of five institutions that have implemented or plan to implement an acute care patient portal.

Acute care patient portals offer various features, including medical record viewing, patient-provider messaging, care plan and safety information, educational materials, and amenities such as food ordering. To our knowledge, no studies have evaluated the impact of a feature that enables inpatients to view their clinicians’ notes, despite evidence that most patients (over 90%) report wanting access to their complete medical record. In the outpatient setting, the OpenNotes consortium has enabled nearly 10 million patients to access their primary care physician’s office notes. Research suggests that OpenNotes increases patient satisfaction, trust and safety, and the project has attracted considerable media coverage.

Given the acceptance of OpenNotes in the outpatient setting and patients’ enthusiasm for increased transparency of their medical record, the expansion of note-sharing to the inpatient setting is a probable next step. To date, few data exist regarding the benefits and risks of inpatient note-sharing, because few institutions have embraced the practice. To address this gap, we designed a mixed-methods study to evaluate usage of a clinical notes feature in an acute care patient portal and describe inpatient perspectives regarding receiving their clinical notes.
Methods

Study Design

Using purposeful sampling, we recruited study participants from a single cardiac step-down unit at NewYork-Presbyterian Hospital / Columbia University Medical Center. Each participant received access to their hospital medical record via a previously described acute care patient portal. Participants accessed the portal using a hospital-provided iPad with internet. A new “Clinical Notes” feature provided participants with real-time access to all clinical notes written about them by either medical doctors or nurse practitioners. Patients maintained access to the portal throughout their hospital stay, even if moved to another unit. Each participant completed a survey as well as a semi-structured interview. Additionally, we recorded and analyzed system usage logs for each participant. The Columbia University Institutional Review Board approved the study.

Recruitment

Inclusion and Exclusion Criteria: We included adult patients admitted to the cardiac step-down unit. Although the portal offered information in both English and Spanish, we included only English-speaking patients because our clinicians write notes in English. We excluded patients with severe cognitive impairment or major psychiatric illness, patients unable to provide written informed consent, patients actively participating in another research study, and patients placed in contact isolation.

Recruitment Protocol: The research coordinator identified potential participants through discussion with attending cardiologists on the unit. The coordinator typically invited patients to participate within 1-2 days of admission. Patients agreed to provide survey data and to participate in an audio-recorded interview. After participants provided written informed consent, the coordinator conducted a brief training session to familiarize the participant with the iPad, the portal login procedure, and the portal layout. For the first five days following recruitment, the coordinator visited each participant daily to troubleshoot any issues with the network, iPad, or portal. After the first five days, the coordinator visited the participant twice weekly until discharge. For infection control purposes, the coordinator used antibacterial wipes to clean the iPad between participants.

Measurements

To collect participant demographics, assess technology literacy, and determine health literacy, we used our previously described patient survey. We used the Patient Activation Measure (PAM) to assess patients’ knowledge, skills, and confidence in managing their health and healthcare. To measure portal usage, we recorded each user action in a detailed system usage log.

A content expert used an iterative process to develop a semi-structured interview guide. Briefly, the content expert developed a preliminary five-question guide based on patient interviews from a separate study. Then, the content expert incorporated feedback from each study team member to develop the final interactive interview guide. The interview topics included general experience with the iPad and portal, and the usefulness, comprehension, emotions, and behavior changes associated with reviewing providers’ notes.

Data Collection and Management

The research coordinator (LVG) collected the patient survey and PAM responses at the time of recruitment. The semi-structured interview took place either two weeks after enrollment or one-to-two days before discharge. The research coordinator received training prior to interaction with participants and followed specific guidelines to ensure consistency of data collection. The coordinator continued enrollment until thematic saturation occurred, meaning participants discussed no novel themes.

Data Analysis

Quantitative: We conducted a descriptive analysis of the patient survey data, PAM data, and system usage log data in R Studio. For the purposes of comparison, we conducted a descriptive analysis of the age, sex, employment status, race, and ethnicity of all patients admitted on the cardiac step-down unit during the study period.
Qualitative: A professional service transcribed the audio-recordings of the interviews verbatim. Transcripts were imported into NVivo Version 11 (QSR International) for thematic analysis. The authors analyzed the transcripts in multiple steps using a qualitative descriptive approach to uncover common themes regarding patients' perspectives on receiving their medical record information, particularly their clinical notes. The analysis included three steps. First, two authors with training in qualitative methods (LVG, RMC) independently read each transcript, and defined codes in a dictionary for the remaining analysis. In addition to codes that emerged from the data, the dictionary included a priori codes based on the research questions, interview guide, and literature. Second, the two authors independently coded all transcripts using nodes corresponding to the coding dictionary. We conducted one round of inter-coder comparison queries in NVivo11. Kappa scores ranged from 0.74 to 0.94. The coders met to review, discuss, and arrive at consensus for the content coding. Third, additional experts reviewed codes for accuracy and to identify common themes. Themes emerging from four or more interviews were identified as common.

Rigor: One author (LVG) conducted all the interviews and confirmed the content of the audio recordings and transcripts to ensure accuracy. All authors reviewed the coding procedure to ensure dependability and credibility, and both coders stringently adhered to the coding procedure. To further enhance dependability and credibility, the coders shared an audit trail and notes on the application of each code with the research team to illustrate the process that led to their conclusions. The authors triangulated multiple sources with the interview data to confirm accuracy, including patients' comments through the portal and during daily visits.

Results

Study Population

Out of eleven patients approached to participate in the study, ten consented to and completed the study, including five congestive heart failure patients, four post-heart-transplant patients, and one post-kidney-transplant patient. The eleventh patient declined to participate due to feeling too ill. Participants spent an average of 13.3 days with access to the portal (range: 4-38). The mean age of participants was 49 years (range: 27-60), comprised predominantly of men (n=9). Table 1 compares the study participants' demographic characteristics with the overall demographics of the cardiac step-down unit. The study population was representative of the employment status, race, and ethnicity of the unit, but was more predominantly male and younger.

Table 1. Demographic Characteristics of the Study Population

<table>
<thead>
<tr>
<th></th>
<th>Study Participants</th>
<th>Entire Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age in years: median</td>
<td>49 (27-60)</td>
<td>65 (20-102)</td>
</tr>
<tr>
<td></td>
<td>(range)</td>
<td></td>
</tr>
<tr>
<td>Sex: n (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>1 (10.0)</td>
<td>247 (45.8)</td>
</tr>
<tr>
<td>Employment status: n (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Employed or Self-Employed</td>
<td>3 (30.0)</td>
<td>84 (15.5)</td>
</tr>
<tr>
<td>Retired</td>
<td>1 (10.0)</td>
<td>135 (25.0)</td>
</tr>
<tr>
<td>Unemployed</td>
<td>6 (60.0)</td>
<td>320 (59.4)</td>
</tr>
<tr>
<td>Race or Ethnicity: n (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>4 (40.0)</td>
<td>155 (28.7)</td>
</tr>
<tr>
<td>Black or African American</td>
<td>4 (40.0)</td>
<td>79 (14.7)</td>
</tr>
<tr>
<td>Asian or Pacific Islander</td>
<td>1 (10.0)</td>
<td>10 (0.02)</td>
</tr>
<tr>
<td>Hispanic, Latino, Spanish origin</td>
<td>1 (10.0)*</td>
<td>155 (28.7)</td>
</tr>
<tr>
<td>Education: n (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Some High School</td>
<td>1 (10.0)</td>
<td>--</td>
</tr>
<tr>
<td>High School Graduate</td>
<td>2 (20.0)</td>
<td>--</td>
</tr>
<tr>
<td>Associate's Degree</td>
<td>3 (30.0)</td>
<td>--</td>
</tr>
<tr>
<td>Bachelor's Degree</td>
<td>3 (30.0)</td>
<td>--</td>
</tr>
<tr>
<td>Master's Degree</td>
<td>1 (10.0)</td>
<td>--</td>
</tr>
</tbody>
</table>

* Non-English Speakers excluded from study because cannot read notes
-- Data not available for the unit
Technology and Health Literacy: The majority of participants reported previously using the internet to look up health information (n=7). All participants reported previously using the internet through another device, and half reported previously using an iPad or another tablet. Three participants reported needing frequent help to read and understand hospital materials or medical information.

Patient Activation: The Patient Activation Measure (PAM), which assesses individuals' knowledge, skills, and confidence in managing their health and healthcare,\textsuperscript{18, 41-56} categorizes patients into four distinct levels of activation: (1) disengaged and overwhelmed, (2) becoming aware, but still struggling, (3) taking action, (4) maintaining behaviors and pushing further. The PAM categorized our study participants as level 2 (n=3), level 3 (n=6), and level 4 (n=1). The high level of activation is consistent with the selection of five participants as transplant candidates.

Portal Usage

Based on system usage log analysis, all ten patients accessed the portal independently, without the research coordinator present. Eight out of ten patients viewed their clinical notes. Patients logged in an average of 2.19 times per day (range 0.375-7.2). Patients typically accessed the portal in the morning, prior to rounds, or in the afternoon, when new test results or clinical notes tended to appear.

Patients navigated to the clinical notes feature most frequently, followed by the laboratory test results and the medications. Patients spent the longest time viewing the clinical notes feature (mean: 7.2 minutes, max: 34.1 minutes), compared to the laboratory test results feature (mean: 3.0 minutes, max: 31.4 minutes) and the medication feature (mean: 1.25 minutes, max: 20 minutes).

Usefulness of Clinical Notes

Participants spoke of reading clinical notes as an informational experience that supplemented their verbal communications with their physicians. Participants felt the notes allowed for communication outside of potentially intimidating or rushed one-on-one situations. One participant discussed how important the written information felt to him, because he felt overwhelmed in one-on-one situations:

"A lot of times, the docs say all this stuff. It's really intimidating. I don't always understand everything they're saying. I get an idea, if things are good or bad, or if they're concerned or not concerned. But this is a different kind of communication that I've never seen before. It's bringing everybody [the patients] in, putting us on the same page." [Pt.1]

Another participant discussed how important the written information felt, given the short amount of time devoted to verbal communication with his physician:

"On average, I get to spend about a minute-and-a-half with my doctor a day, and so when I go back after the doctor writes the report, I get a more detailed account. Although it's not a personal one-on-one, it at least comes from her mind that she has seen me and these are the more detailed things." [Pt.7]

Participants viewed notes as objective indicators of both their health condition and their progress while in the hospital. Multiple participants (n=3) used the metaphor of seeing information in "black and white" through the notes. The participants explained:

"It's different, when you see the note. It's like a report card, I guess. When you see it; it hits you a different way. It's not as emotional, it's more objective. And it's there. It's right there in black and white so you can totally track like, "Oh, I'm good, I'm good, I'm good, I'm in trouble, ok now I got back on my feet." So I find that that's very good information." [Pt.1]

"To be honest, I can't always believe the doctors. I think that sometimes they sweet-talk you, excuse me for saying that. They don't want you to stress, and it's not like they lied to you, but sometimes they don't tell you the whole truth." [Pt.7]
Participants expressed a sense of ownership over their data, and wanted access independent of plans to read or use it. One participant who did not view his clinical notes still wanted the notes to be available:

"Even though I didn’t look at them [the notes], it’s just nice to have them. You can’t go wrong with it. More information is always better. Maybe if I had gotten sick, maybe it would have been different, or if something had changed, I would have looked." [Pt.5]

Half the participants (n=5) asked for access outside the hospital, unprompted. One participant even offered to pay for outside access. Two patients reported photographing information in the portal on their personal devices, to avoid fees associated with requesting their medical record from the hospital after discharge. Multiple participants (n=4) expressed the desire to review notes from prior hospitalizations.

Comprehension and Insight

Participants spoke of clinical notes as "truth tellers" that improved their insight about their progress and condition. One participant, after reading his notes, asked for a family meeting:

"My sister and I didn't realize that what I'm going through is as bad as it is. Basically, reading the notes, we felt it [my condition] was more serious. For some reason we weren't getting that in the room. We needed to know what our goals are. Which direction we're heading and how we're going to get there." [Pt.9]

After the family meeting, the patients' care team involved the palliative care service, who arranged the participant's discharge to home hospice care.

For another participant, the notes offered clarity about his condition's severity, and encouraged him to consider destination left ventricular assist device (LVAD) surgery:

"I was walking up and down that hall, going to the bathroom on my own. I really thought I was going to be able to go home, without any drugs, the Milrinone, all that stuff. But then you look at the notes. And they say it's the total opposite. I can see why I need this [LVAD]. I was glad that the doctors made the best decision... Every patient that walks through that door wants the raw deal of what's going on with their health situation." [Pt.10]

Participants spoke about "getting on the same page" as their physicians. Multiple participants (n=4) felt more able to converse with their physician, either because of better insight about their condition or increased comfort around the care team. The participants explained:

"It's very, very useful, because from the note we [I and my family] know exactly what's going on. And when we talked to the doctor, we were able to ask questions, and we know what the doctor is saying." [Pt.8]

"[The notes] help me keep track of who everybody is and who has seen me recently. Today, this older doc came in. And I was comfortable enough to be like, “Do I have to do a blood test every day?” And he was like, “No. We don’t have to do that.” And he totally changed it for me. If I was uncomfortable with him, I would never have broached that topic." [Pt.1]

Two participants reported that the portal answered questions they might otherwise ask their physician:

"It's better because you don’t have to call [the doctor] to ask questions; you just go there [the notes] directly." [Pt.6]

All participants who read their clinical notes reported struggling to understand medical terms and acronyms, excepting one participant who had a Master's degree in a medicine-related field. One participant explained his struggle with interpreting acronyms:

"There were a lot of acronyms. Some of these acronyms, I knew what they were talking about just because I know my history... there were a lot of line items [e.g., HPI, PMH, ROS], that I didn’t really get. But the parts that I could read and make my way through - that seemed pretty evident - I would just read." [Pt.1]
Emotional Reactions

Multiple participants (n=4) reported that viewing their notes decreased their anxiety. No participant expressed concern that seeing the notes increased anxiety. The notes reduced anxiety by providing new information, verifying known information, and offering additional processing time:

"Humans can make mistakes and even the input to these computers can be wrong. But when she [the doctor] tells me one thing and it’s backed up with visual, it makes me feel a little bit better about it... and you feel more confident, seeing that at least whatever episode or what you felt was bothering you at the moment has been notated, and at least it’s on their radar." [Pt.7]

"A lot of people get scared by this stuff. They’re intimidated. They’re in the hospital. They’re scared. It’s tough when there’s a guy in a white coat with stethoscope and glasses. Standing on top of you and telling you, “your CBC level is blah-blah-blah.” It’s rough. This [reading the notes] is not as intense. It’s a nice filter. It's like reading the paper." [Pt.1]

Participants reported increased appreciation for and trust in their clinicians. Seeing the amount of documentation prompted appreciation for the work physicians devoted to patient care:

"I don’t really understand everything in the notes, but it’s really amazing to see everything that goes into my care. I don’t think I really appreciated how much gets done for me before." [Pt.3]

"When I read the notes, I feel that the doctors are following and that they care for you." [Pt.8]

"[The notes] grounded me in my whole being here. At first, I just want to get out of here. But when I got this, I was like, “Wait a minute, they’re doing all this work for me.” I really looked at all this. It kind of laid it out for me. And I was like, 'Maybe I should just settle in. Let them do what they got to do.' " [Pt.1]

"Sometimes there are just so many different things, and if your gout doesn't clear up, you feel like you're not being taken care of sometimes. But when you read in the note, 'pending talking to a gout doctor,' it makes you feel better. They're really trying to care for you." [Pt.7]

Health Behavior Change

Multiple participants (n=4) reported changing their health behaviors after reading the notes, including one participant who called a family meeting, and another who considered LVAD surgery, described above. Another participant began drinking the nutrition supplement (Ensure) that his nutritionist had prescribed:

"I started drinking the Ensure. Honestly, I never really gave it a shot before. It was just that my dad had to drink it when he was sick, so it bummed me out. But once I saw everything, I felt like, the nutritionist is trying and is giving me this food for a reason, so I should try too, and do what I can to make my numbers as good as they can be." [Pt.1]

Another participant reported feeling more likely to take his antihypertensive medications:

"When I saw that my weight went really down, that was an eye-opener. I mean, you know it, but then when you actually see it, it’s different. Maybe when I get out, I’ll be a little more careful." [Pt.3]

Participants described how reading the notes increased their ability to engage with their care, because the written information allowed for processing time to overcome anxieties:

"I was able to read it and then process it, and then when the doctor came in and talked to me, it was easier to talk about." [Pt.9]

"At first, to be honest, I was a little bit weirded out. I was so excited to get my heart. And I think of that moment, and it's just like 'do I really want to know any more?' But once I got into it, it was like a good book. You just want to
Patient Safety

Participants reported incidents where access to their information improved their quality of care. One patient noticed a medication error:

"This morning I saw the prednisone had fallen off [of my medication list]. I showed the nurse, who agreed with me that something wasn't right. She called the doctor, and within a minute-and-a-half the prednisone was back on [my medication list]. And within another minute-and-a-half, my nurse was back with the [prednisone] pill." [Pt.7]

Another patient noticed a low calcium level which needed correction after plasmapheresis:

"In the afternoon something happened. It was about my calcium level. I went to the nurse, and explained to her, the normal is between this and this, and I'm here. The nurse wasn't sure, so I said look, I'm going to show you [with the portal]. She [the physician] doubled my calcium supplement. She said since I will have more plasmapheresis, we need the calcium to be higher. So [the portal], it saved me." [Pt.8]

Upon reviewing the chart, the participant's physician deferred his plasmapheresis until his serum calcium level returned to normal.

Discussion

A rapidly growing literature documents the benefits of acute care patient portals. Research suggests that bedside access to information may increase patient safety and satisfaction. To our knowledge, ours is the first study to elicit inpatients' perspectives on receiving their clinical notes through an acute care patient portal.

Although only ten patients participated, our results suggest that patients may benefit specifically from inpatient note-sharing. In our qualitative interviews, patients reported better access to information, increased insight into their condition, increased appreciation for their clinicians, changes in their health behaviors, and medical error prevention. The system usage log data demonstrated that patients used the clinical notes feature more frequently and for longer durations than any other feature. We achieved these results in a complex clinical practice setting, with variations in patients' conditions and treatments as well as heterogeneity among clinicians in education, experience, and motivation to share notes.

The lack of common negative themes surprised the research team. We expected that note-sharing might increase some patients' anxiety. However, we did not find evidence of increased anxiety among our participants, but we did find evidence of decreased anxiety. Given the study design, we do not know whether response bias existed which led to this conclusion. However, an alternative explanation is plausible, that information in any form counteracts the uncertainty and disempowerment patients feel when hospitalized. This explanation is consistent with prior OpenNotes research.

One striking result was the intervention's apparent influence on health behavior change and hospital outcomes. In a sample of only ten patients, four reported significant changes to their health behaviors, and two reported preventing possible medical errors. This included one patient who changed his entire course of end-of-life care, and another who opted to pursue surgery. Because completing an interview allows patients to reflect on and verbalize changes, further research should explore whether note-sharing actually translates to increased illness understanding, medical knowledge, health behavior change, and error prevention.

Some clinicians fear that note-sharing may prompt mistrust, force unwelcome changes to documentation practices, require increased time with patients to alleviate concerns, and increase legal liability risk. In our study, physicians did not appear to change their documentation practices. Patients expressed surprise and gratitude at receiving their notes, and after reading them, voiced increased appreciation for and trust in clinicians. In some instances, patients reported answering questions using the portal rather than asking their physicians. Although the patients' perspective...
cannot serve as a proxy for the impact on clinicians, our results encouraged us to remain open-minded about note-sharing. Future research should identify the unintended consequences of inpatient note-sharing, and investigate the effects of note-sharing on consultation time.

Our study had several limitations. First, we enlisted patients from a single clinical unit in an urban academic medical center. Though our sample included only ten patients, we achieved thematic saturation. Consistent with the sample size, we introduced stringent criteria for identifying common themes, namely that the theme must exist in four or more interviews. Second, we did not select participants randomly. Although the attending cardiologists purposefully selected patients representative of the unit, a selection bias for younger, more technology literate patients existed. Third, the study used five transplant patients as participants. Transplant patients must demonstrate high engagement prior to receiving their transplants, and may exhibit more interest in their health care than other patients. To confirm our findings, this study should be repeated with randomly selected participants and a more generalizable patient population than transplant patients.

Conclusion

To our knowledge, this is the first study to assess hospitalized patients' perspectives about receiving their clinical notes through an acute care patient portal. In qualitative interviews, patients who read their notes described better insight into their medical conditions, better access to information, decreased anxiety, increased appreciation for their clinicians, improvements to their health behaviors, and engagement with medical decision-making. Although our patient population is too unique to draw broadly generalizable conclusions, our results suggest that further studies to explore the potential benefits and unintended consequences of inpatient note-sharing are warranted.

Acknowledgements

For her valuable contributions the authors thank: Dianna L. Smith, MSN, NE-BC, RN-BC, CCRN-K, Patient Care Director for Milstein Hospital 5 Garden South Cardiac-Medical Telemetry Unit.

References

43. Landro L. How to get patients to take more control of their medical decisions. The Wall Street Journal. 2017.
When synonyms are not enough: Optimal parenthetical insertion for text simplification

Yang Gu\(^1\), Gondy Leroy, PhD MS\(^1\), David Kauchak, PhD\(^2\)

\(^1\)University of Arizona, Tucson, AZ; \(^2\)Pomona College, Claremont, CA

Abstract

As more patients use the Internet to answer health-related queries, simplifying medical information is becoming increasingly important. To simplify medical terms when synonyms are unavailable, we must add multi-word explanations. Following a data-driven approach, we conducted two user studies to determine the best formulation for adding explanatory content as parenthetical expressions. Study 1 focused on text with a single difficult term (N=260). We examined the effects of different types of text, types of content in parentheses, difficulty of the explanatory content, and position of the term in the sentence on actual difficulty, perceived difficulty, and reading time. We found significant support that enclosing the difficult term in parentheses is best for difficult text and enclosing the explanation in parentheses is best for simple text. Study 2 (N=116) focused on lists with multiple difficult terms. The same interaction is present although statistically insignificant, but parenthetical insertion can still significantly simplify text.

Introduction

Text-based information plays an important role in patient education. Doctors send leaflets and instructions home with patients so they can adequately care for themselves. More than half of American Internet users independently search for health-related information online\(^1\). Therefore, it is essential that medical information be written in language simple enough to be correctly interpreted by patients. Changing the words and grammar of a text can achieve text simplification, and appropriate modifications will enhance readability and comprehension.

In our prior research, we found that term familiarity, as approximated by term frequency, is a reliable indicator of text difficulty\(^2\). Term familiarity can identify and rank terms by difficulty. By substituting familiar synonyms for difficult terms, we were able to decrease both actual and perceived difficulty. However, many medical concepts do not have simple synonyms. For example, a single word is not enough to understand “reactogenecity”; the reader needs to make the connection between three concepts: a medical product, its adverse reaction, and the anticipated nature of the adverse reaction. For such concepts, we must insert an explanation in the text. Enclosing the new text in parentheses is one of the most natural ways to insert relevant but discontinuous information into text. Previous research has found that adding explanatory content enclosed in parentheses simplifies text\(^3\)\(^4\); however, the best way to formulate and insert the new information has not been investigated.

This research contributes to a semi-automated text simplification system for writers, which will automatically identify difficult text and suggest options for simplification. The human writer can choose to accept, refine or reject system suggestions to create coherent text. We examine four factors that may affect simplification using parentheticals: the difficulty of the original text, whether to place a term or an explanation in the parentheses, the difficulty of the explanatory content, and the position of the difficult term in the sentence. We conducted one study to investigate those four factors in sentences that contain a single difficult term. In a second study, we examine the extent to which parallelism in the formulation of parenthetical expressions will impact simplification of texts with multiple difficult terms in a list.

Background

Useful Existing Models to Guide Simplification

The fields of psychology and linguistics have explored the reading and discourse comprehension process, and there are many models of text comprehension. Some of the more well-known models include the Construction-Integration Model\(^5\), the Landscape Model\(^6\), and the Resonance Model\(^7\). Every model differs in its assumptions and computational model, but they agree on how the comprehension process works on a high level. During reading, the text provides the reader with cues, such as words and punctuation. Readers search their long-term memory to give meanings to these cues and construct propositions. Comprehension occurs when the reader is able to make connections between the propositions. In text simplification, we can facilitate comprehension by modifying the cues, such as words and syntactic markers, which will help readers make connections.
Explaining Difficult Concepts

Discovering and collecting a consumer-friendly medical vocabulary has been an ongoing effort in the health and medical informatics community. For example, there have been several efforts to collect vocabulary through web mining or automatically generating simple explanations based on relationships in the UMLS Metathesaurus.

Synonym replacement is an intuitive, effective, and popular approach for lexical simplification, but it is not always possible. When no synonyms are available, explanations need to be inserted. Kandula et al. made the distinction between “definitions” and “explanations” for terms. A “definition” is a description of the terminology that aims to be correct, accurate, and precise. For example, the Wikipedia page for “reactogenecity” provides a definition from the NIH (National Institute of Health), which states “Reactogenicity events are adverse events that are common and known to occur for the intervention/investigational product being studied.” An “explanation” is a description that is more abstract but also more understandable to the reader. It should not introduce more difficult terminology. In a sentence about anthrax vaccines, we can explain “reactogenecity” as “known side effects from vaccines.”

Some automated text simplification systems employ explanation generation. For example, Eom, Dickinson, and Sachs designed a system for second language learners that allows users to upload text in one panel and show the sense-specific definition for a selected vocabulary in a side panel. Damay et al.’s SimText system first uses a thesaurus to look for synonym replacements for difficult terms and then appends a definition when synonyms are unavailable. Kandula et al.’s system specializes in the medical domain and also first searches for consumer-friendly synonyms to medical terms in the Open Access and Collaborative Consumer Health Vocabulary. When that is unavailable, they generated simple explanations using high-level relationships in the UMLS.

Simplification through Parenthetical Expressions

A natural way to introduce new content into text is with parenthetical expressions. A parenthetical expression is any expression embedded in a host expression that makes no contribution to the structure of the host and makes the host expression discontinuous. In written text, the expression can be enclosed using a variety of punctuation characters, such as brackets, dashes, or commas. For example, consider “He told John – his best friend from college – all about it.” While it introduces more information, the parenthetical expression also disrupts the structure of the original sentence. We are interested in appositive parenthetical expressions, which reformulate parts of the original host expression. The reformulation provides additional information about the concept in the host expression. In the context of text simplification, we only consider appositive parenthetical expressions enclosed by literal parentheses in written texts.

We have two options for formulating parenthetical expressions to simplify difficult words. The common approach is to insert the explanatory content as a parenthetical expression behind the target term, as in “local and general reactogenicity (known side effects from vaccines) are expected.” This can be easily automated. The reader expects the parentheses to signal an interruption to the flow and structure of the entire sentence, so the language of the explanatory content need not blend perfectly with the host expression. However, the reader needs to read the parenthetical expression to understand the concept.

Alternatively, we can incorporate the explanatory content into the text and state the obscure medical term in a parenthetical expression, as in “local and general known side effects from vaccines (reactogenicity) are expected.” In this formulation, it becomes apparent immediately after the modifier clause that the clause describes the side effects of medicine. The reader can comprehend this information faster with less strain on their memory. Most reading comprehension models agree that readers keep snippets of information in their working memory, which has an approximate limit of seven plus/minus two units of information. Therefore, making information comprehensible as early as possible, and avoiding long interruptions and divergences in the text that must be stored in working memory, can be advantageous for comprehension. However, directly inserting definitions into text will create cohesion issues, especially when definitions are long, like those of many medical terms.

Because adding definitions within parentheses is widely used in written English, this formulation has been adopted as the default for parenthetical insertions. Nonetheless, there is little empirical support that one formulation is superior to the other.

Study Objectives

The goal of our studies is to discover combinations of features that would influence and optimize simplification when parenthetical insertion is needed. The first study focuses on simplifying sentences with one difficult “target” term. We investigate the impact of the difficulty of the source material, the placement of the parentheses, the position...
in the sentence, and the difficulty of the explanatory content. The second study focuses on simplifying sentences with multiple target terms in a single list. We investigate the impact of the difficulty of the source material and the parallelism of the parentheses. The target terms always remain in the text, so readers can learn the difficult vocabularies in their medical texts.

Methods

We conducted two user studies on the effects of parenthetical insertion on simplification. The subjects were asked to read short passages and answer questions about the actual and perceived difficulty of the text. Subjects were recruited on Amazon Mechanical Turk (MTurk), an online service that allows human workers to complete small tasks for monetary compensation. We restricted the workers to being in the United States with 95% approval ratings on previous tasks. MTurk has been used for tasks ranging from survey completion, data annotation or tagging, to user studies; study results have been shown to be as reliable as those from traditional approaches\(^1\),\(^2\).

Study Instrument. Each study used 16 passages identified from 16 documents. Eight documents are Wikipedia articles about medical topics and eight are PubMed abstracts of studies on various diseases. To select the passages, sentences were randomly selected and manually checked for suitability based on the following criteria:

1. The sentence does not contain difficult terms other than the target term, which we will augment with parentheses. The target term can be defined sufficiently in one sentence or less, and a reasonable, simple explanation can be created based on the target term.
2. The sentence can be edited to satisfy all experiment conditions, i.e., the definition and explanation for the term can be incorporated into the text of the sentence with minimal editing. Study 1 also requires that the target term can be moved to the end of the sentence with only minor changes to the words in the sentence.
3. The sentence is preceded and followed by one or two sentences from the same document that provide context for the passage. The context does not contain difficult terms or difficult terms can be removed.

The following is a passage from the Wikipedia article on Anthrax with the target word “reactogenicity”:

“Vaccines against anthrax for use in livestock and humans have had a prominent place in the history of medicine. The French scientist Louis [...] All currently used anthrax vaccines show considerable local and general reactogenicity (erythema, induration, soreness, fever) and serious adverse reactions occur in about 1% of recipients.”

We removed or simplified other difficult words besides the target word and split up sentences to make the context easier to understand. We also ensured that we did not introduce misinformation about the topic in our modifications. The edited passage used in our study is:

“Vaccines against anthrax in humans and livestock have had a prominent place in the history of medicine. Considerable local and general reactogenicity are expected from all currently used anthrax vaccines for humans. Common reactions include soreness and fever, and serious adverse reactions occur in about 1% of recipients.”

We conducted all statistical analyses with SPSS Statistics 24 with standard settings.

Study 1: Single Parenthetical Insertion

Design

This study focuses on simplifying sentences with one difficult target term. We use a 2x2x2x2 full-factorial within-subjects experimental design with the following factors:

1. Types of text: Wikipedia and PubMed. We use the source of the text as a proxy for difficulty. The Wikipedia articles represent easier text, while the PubMed research abstracts represent more difficult text. Wikipedia articles may contain difficult vocabulary and content, but they are meant to be descriptive and informative. Research articles have a different writing style and assumption of the reader’s knowledge, but they are also an important resource for expert patients and caregivers. We want to investigate how localized simplification that targets one single difficult concept may impact a generally difficult piece of text differently from a text that is only difficult at that single point.

2. Placement of Explanatory Content: Inside Parentheses and Outside Parentheses. Explanatory content placed inside parentheses is commonly used and easy to automate, but the alternative is less disruptive to the flow of the sentence and the reading process. Table 1 provides an example of how factors 2, 3, and 4 combine to form sentences.
3. Difficulty of the Explanatory Content: Simple (explanation) and Difficult (definition). Definitions can be extracted from established domain resources and provide more comprehensive information. Explanations tend to contain simpler language. For each target term, we create a “definition” using definitions from the term’s Wikipedia page or from a dictionary. Then, the definition is shortened and simplified manually to create an explanation. We did not use a single resource such as UMLS or Wikipedia for definitions because some terms are not defined concisely or understandably to a layperson in any one place. Since our system is designed for writers, they can intervene to refine existing definitions.

4. Position in Sentence: Middle and End. Since parenthetical expressions introduce an interruption in the flow of the text, we consider where this interruption occurs. If we insert the parenthetical expression at the end of the sentence, the interruption may be less disruptive than if we insert it in the middle.

Table 1. Demonstration of Study 1 Experiment Conditions

<table>
<thead>
<tr>
<th>Placement of Explanatory Content</th>
<th>Difficulty of Explanatory Content</th>
<th>Position in Sentence</th>
<th>Sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inside Parentheses</td>
<td>Simple (Explanation)</td>
<td>Middle</td>
<td>Considerable local and general reactogenicity (known side effects from vaccines) are expected from all currently used anthrax vaccines for humans.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>End</td>
<td>All currently used anthrax vaccines for humans are expected to produce considerable local and general reactogenicity (known side effects from vaccines).</td>
</tr>
<tr>
<td></td>
<td>Difficult (Definition)</td>
<td>Middle</td>
<td>Considerable local and general reactogenicity (adverse events that are common and known to occur for a medical intervention or investigational product) are expected from all currently used anthrax vaccines for humans.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>End</td>
<td>All currently used anthrax vaccines for humans are expected to produce considerable local and general reactogenicity (adverse events that are common and known to occur for a medical intervention or investigational product).</td>
</tr>
<tr>
<td>Outside Parentheses</td>
<td>Simple</td>
<td>Middle</td>
<td>Considerable local and general known side effects from vaccines (reactogenicity) are expected from all currently used anthrax vaccines for humans.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>End</td>
<td>All currently used anthrax vaccines for humans are expected to produce considerable local and general known side effects from vaccines (reactogenicity).</td>
</tr>
<tr>
<td></td>
<td>Difficult</td>
<td>Middle</td>
<td>Considerable local and general adverse events that are common and known to occur for a medical intervention or investigational product (reactogenicity) are expected from all currently used anthrax vaccines for humans.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>End</td>
<td>All currently used anthrax vaccines for humans are expected to produce considerable local and general adverse events that are common and known to occur for a medical intervention or investigational product (reactogenicity).</td>
</tr>
</tbody>
</table>

We evaluate simplification using the following metrics:

Actual Difficulty. Average accuracy on three multiple-choice questions about the content of the passage, which the subjects answer without the text present. One question specifically focuses on the meaning of the target term, which is most directly related to the parenthetical expression.

Perceived Difficulty. A 5-point Likert Scale (1 = Very Easy, 5 = Very Difficult) response to “How difficult would this passage look in a text” immediately after reading.

Reading Time. Time (in seconds) the respondents spent on the page containing the passage.

Procedure

Subjects recruited on MTurk were redirected to a survey on Qualtrics. We used Qualtrics’ randomization features to assign 8 passages to each subject. The subjects do not see more than one version of any single passage.

After reading each passage, subjects were taken to a new page with six questions. The first question asked about perceived difficulty. Then there were three multiple-choice questions about the content of the passage and an attention question. The final question always asks about the subject’s prior knowledge or experience with the condition: “How familiar are you with the topic of [topic of passage].” Response choices were: “I have never heard of it before,” “I have only heard of it in passing,” “I have studied it in detail for educational purposes,” “I know of someone with this condition,” or “I have (or have had) this condition.” We interpret scores of 1 or 2 (first two
options) as minimal or no knowledge of the condition and scores of 3 to 5 (last three options) as high knowledge, and use this metric to control for the impact of background knowledge on the results.

We also use an attention question to filter out respondents who have not read the passage. For example, for the earlier passage about Anthrax vaccines, we asked “Which word appeared in the passage?” To someone who read the passage, the correct answer is obvious among the choices: Anthrax, Arthur, Anthropology, and Antonym.

**Results**

We invited 56 MTurk workers to read and answer questions on 8 passages for compensation of $3 USD. We removed three subjects for having missed two or more attention questions and six subjects for spending an unreasonably long or short amount of time reading or answering the questions. Prior knowledge of the topic is significantly correlated with reading time (r = 0.102, p < 0.046), perceived difficulty (r = -0.212, p < 0.000), and actual difficulty (r = 0.1, p < 0.038), so we removed all responses from someone with high knowledge of the medical condition.

Table 2 summarizes the demographic characteristics of subjects whose responses were included in the analysis. The majority are under 40 years old. More than half have a bachelor’s degree and about one-quarter have a high school diploma. About three-quarters are white and most of the respondents speak only English at home. Females make up just about half of our respondents.

**Table 2. Demographics of Study 1 Subjects (N=47)**

<table>
<thead>
<tr>
<th>Age</th>
<th>Count (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;30</td>
<td>14 (0.298)</td>
</tr>
<tr>
<td>31 to 40</td>
<td>16 (0.340)</td>
</tr>
<tr>
<td>41 to 50</td>
<td>9 (0.192)</td>
</tr>
<tr>
<td>51 to 60</td>
<td>8 (0.170)</td>
</tr>
<tr>
<td>61 to 70</td>
<td>0 (0.0)</td>
</tr>
<tr>
<td>&gt;71</td>
<td>0 (0.0)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Gender</th>
<th>Count (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>27 (0.574)</td>
</tr>
<tr>
<td>Male</td>
<td>20 (0.426)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Race (multiple answers allowed) (N=49)</th>
</tr>
</thead>
<tbody>
<tr>
<td>American Indian/Alaska Native</td>
</tr>
<tr>
<td>Asian</td>
</tr>
<tr>
<td>Black</td>
</tr>
<tr>
<td>Native Hawaiian/Pacific Islander</td>
</tr>
<tr>
<td>White</td>
</tr>
<tr>
<td>More than one race</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Education</th>
</tr>
</thead>
<tbody>
<tr>
<td>Less than high school</td>
</tr>
<tr>
<td>High school diploma</td>
</tr>
<tr>
<td>Associate’s degree</td>
</tr>
<tr>
<td>Bachelor’s degree</td>
</tr>
<tr>
<td>Masters</td>
</tr>
<tr>
<td>Doctorate</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Language Spoken at Home</th>
</tr>
</thead>
<tbody>
<tr>
<td>Never/Rarely English</td>
</tr>
<tr>
<td>Half English</td>
</tr>
<tr>
<td>Mostly English</td>
</tr>
<tr>
<td>Only English</td>
</tr>
</tbody>
</table>

**Table 3. Means for Actual Difficulty, Perceived Difficulty, and Reading Time for each Experimental Condition**

<table>
<thead>
<tr>
<th>Source of Passage</th>
<th>Placement of Explanatory Content</th>
<th>Difficulty of Explanatory Content</th>
<th>Position in Sentence</th>
<th>N</th>
<th>Actual Difficulty (all) Mean</th>
<th>Actual Difficulty (term) Mean</th>
<th>Perceived Difficulty Mean</th>
<th>Reading Time Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Actual Difficulty (all) Mean</td>
<td>Actual Difficulty (term) Mean</td>
<td>Perceived Difficulty Mean</td>
<td>Reading Time Mean</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>Baseline</td>
<td>-</td>
<td>-</td>
<td>13</td>
<td>0.641</td>
<td>0.214</td>
<td>0.462</td>
<td>0.519</td>
</tr>
<tr>
<td></td>
<td>Inside Parentheses</td>
<td>Simple (explanation)</td>
<td>Middle</td>
<td>16</td>
<td>0.792</td>
<td>0.269</td>
<td>0.688</td>
<td>0.479</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Difficult (Definition)</td>
<td>Middle</td>
<td>14</td>
<td>0.714</td>
<td>0.410</td>
<td>0.786</td>
<td>0.426</td>
</tr>
<tr>
<td></td>
<td>Outside parentheses</td>
<td>Simple</td>
<td>End</td>
<td>16</td>
<td>0.667</td>
<td>0.312</td>
<td>0.588</td>
<td>0.507</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Difficult</td>
<td>End</td>
<td>18</td>
<td>0.611</td>
<td>0.308</td>
<td>0.556</td>
<td>0.511</td>
</tr>
<tr>
<td>PubMed</td>
<td>Baseline</td>
<td>-</td>
<td>-</td>
<td>15</td>
<td>0.556</td>
<td>0.371</td>
<td>0.600</td>
<td>0.507</td>
</tr>
<tr>
<td></td>
<td>Inside Parentheses</td>
<td>Simple</td>
<td>Middle</td>
<td>16</td>
<td>0.750</td>
<td>0.285</td>
<td>0.625</td>
<td>0.500</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Difficult</td>
<td>Middle</td>
<td>19</td>
<td>0.684</td>
<td>0.304</td>
<td>0.684</td>
<td>0.478</td>
</tr>
<tr>
<td></td>
<td>Outside parentheses</td>
<td>Simple</td>
<td>Middle</td>
<td>13</td>
<td>0.744</td>
<td>0.277</td>
<td>0.615</td>
<td>0.506</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Difficult</td>
<td>Middle</td>
<td>19</td>
<td>0.685</td>
<td>0.267</td>
<td>0.667</td>
<td>0.485</td>
</tr>
</tbody>
</table>

814
Table 3 summarizes the means for all experimental conditions. Actual difficulty is measured by accuracy, so lower scores denote more difficult texts. Consistent with conventional wisdom, inserting any explanatory material in any formulation reduces actual difficulty compared to the baseline condition, with a few exceptions for when difficult definitions are inserted into the text. For Wikipedia texts, the improvement from baseline is the largest when we insert simple explanations in parentheses at the end of the sentence ($t_{35} = 2.241, p < 0.034$). Other configurations do show improvement as well, though not statistically significant, though this may be due to the small number of responses we have in each condition.

When we aggregate data based on features of interest, we do observe significant interactions, though, some parenthetical formulations are more successful at simplification than others. The best formulation for Wikipedia passages (simple explanation inside parentheses at the end of sentences) yields significantly higher accuracy (i.e. easier documents) than the worst formulation (difficult explanation outside parentheses the middle of the sentence; $t_{31} = 2.242, p < 0.032$). Similarly, the best configuration for PubMed (explanation inside parentheses in the middle) is significantly different than the worst configuration (definition inside parentheses in the middle; $t_{30} = 2.236, p < 0.033$). It is important that we understand which combinations of features can affect the effectiveness of each formulation.

We performed a 4-way ANOVA on the experimental condition for actual difficulty, perceived difficulty, and reading time ($N = 260$). We exclude the baseline from this discussion because it does not contain any explanatory content.

**Actual Difficulty**: We found a main effect for the difficulty of explanatory content on the accuracy of the content questions ($f_{1,244} = 6.501, p < 0.011$). Readers presented with a technical definition correctly answered 62.6% of the content questions, while readers presented with the explanation achieved higher accuracy and were able to correctly answer 72.1% of the time. The difficulty of the explanatory content also has a significant effect ($f_{1,244} = 4.056, p < 0.045$) on the accuracy of the questions focusing on the meaning of the target term. Readers presented with a technical definition were able to correctly answer the question about meaning of the target term 56.4% of the time, while readers presented with the explanation were able to correctly answer 68.7% of the time.

Figure 1.A shows the significant interaction between placement of the parentheses and the source material of the text ($f_{1,244} = 5.759, p < 0.017$). If the text is from Wikipedia, adding explanatory content inside the parentheses (with the difficult term outside the parentheses) results in better accuracy than adding the new content outside the parentheses (with the difficult term inside the parentheses). However, if the text is from PubMed, which is more difficult and less familiar to the general audience, the results are reversed but the magnitude of change is much smaller. This suggests that the best approach for inserting new explanatory content into text depends on the difficulty, or at least type of text. Using the wrong approach could reduce comprehension and increase difficulty.

We also found a nearly statistically significant trend ($f_{1,244} = 3.633, p < 0.058$) showing a similar interaction between the placement of parentheses based on type of explanatory content. As seen in Figure 1.B, given a simple explanation, adding explanatory content inside parentheses resulted in a high retention of content compared to adding the explanatory content outside. However, given a technical definition, adding it inside parentheses resulted in lower retention compared to keeping the explanatory content outside of parentheses and integrated with the text.

![Figure 1](image-url)

**Figure 1. Interaction Effects from Placement of Parentheses and Difficulty of Source or Explanatory Content**
**Perceived Difficulty.** While perceived difficulty is significantly correlated with actual difficulty ($r = -0.243$, $p < 0.001$), we do not see a significant difference between individual experiment conditions and the baseline due to the small sample size. However, if we aggregate the data based on specific features, we find that the difficulty of the explanatory content has a significant effect ($f_{1, 244} = 4.053$, $p < 0.045$) on perceived difficulty. Readers presented with a technical definition report the average difficulty of the text as 3.143, while readers presented with a simple explanation report a lower difficulty of 2.860.

**Reading Time.** One concern is that retention of the text material would be influenced by the time spent reading it. In our experiment, we kept wording consistent between conditions as much as possible, only changing the position and placement of the target term and associated vocabulary. We ran correlations between our difficulty scores and reading time. We found that reading time is not significantly correlated with perceived difficulty ($r = -0.027$, $p < 0.666$), term question ($r = 0.083$, $p < 0.185$), and actual difficulty ($r = 0.042$, $p < 0.490$), which rules out reading time as an alternative explanation to the effects we observe in our data.

Table 3 shows two outliers for mean reading time of Wikipedia passages. Three subjects in the baseline condition took about six minutes each. Only the baseline condition had multiple subjects needing more than two minutes. This behavior can be explained by the subjects becoming confused by passages that were lexically simple (Wikipedia text) but incomprehensible without the explanatory content. The other outlier condition is caused by one single subject whose time pattern suggests he/she took a break during the experiment and returned later. This reflected plausible behavior of a subject who was committed to the task, so we retained the data. (We removed all data from subjects who took multiple long breaks). Without this subject, the mean reading time is 36.547 seconds.

**Study 2: Multiple Parenthetical Insertions**

**Design**

Study 2 focuses on the cases where there are several target terms that form a list in a sentence to understand whether using parallel parentheses insertion, which would provide structure to the text, simplifies the text. Based on the results from Study 1, position of target terms in the sentence is not considered, and we use only explanations. We also included a baseline condition containing the original sentence. We use a 2x3 full-factorial within-subjects experimental design with the following factors and levels:

1. **Types of text:** *Wikipedia and PubMed.* Identical to Study 1.

2. **Parallelism of Parentheses:** *All Inside, Mismatched, and All Outside.* We expect parallelism to provide more cues about the structure and organization of the information in the passage, which should help with simplification. Table 4 provides an example for each level.

<table>
<thead>
<tr>
<th>Parallelism of Parentheses</th>
<th>Sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>All inside</td>
<td>Heart murmurs are most frequently categorized into systolic (when the heart is contracting) and diastolic (when the heart is expanding) heart murmurs, differing in the part of the heartbeat on which they can be heard.</td>
</tr>
<tr>
<td>Mismatched</td>
<td>Heart murmurs are most frequently categorized into systolic (when the heart is contracting) and diastolic (when the heart is expanding) heart murmurs produced when the heart is expanding (diastolic), differing in the part of the heartbeat on which they can be heard.</td>
</tr>
<tr>
<td>All outside</td>
<td>Heart murmurs are most frequently categorized into heart murmurs produced when the heart is contracting (systolic) and heart murmurs produced when the heart is expanding (diastolic), differing in the part of the heartbeat on which they can be heard.</td>
</tr>
</tbody>
</table>

As in Study 1, we evaluate simplification using actual difficulty, perceived difficulty, and reading time using the same scales.

In this study, each subject was assigned 4 passages (since we have fewer conditions).

**Results**

We recruited 48 workers for our study for $2 each. Similar to Study 1, we removed 6 workers for spending an unreasonably long or short amount of time reading or answering the questions and those with high background knowledge.
The majority of our respondents are under 40 years old (see Table 5). Just over half hold a bachelor’s or associate’s degree. Over three-quarters are white and most speak only English at home. Females make up just above half of our respondents.

Table 6 summarizes the mean values of our major metrics for each experiment condition. The reduction in actual difficulty of Wikipedia text is significant when we insert all explanations in parentheses ($t_{34} = 3.647$, $p < .001$) or insert some explanations in a mismatched manner ($t_{31} = 3.647$, $p < .010$). In general, the baseline condition is most difficult. Mismatched parentheses are an improvement over the baseline, but are still more difficult than any parallel formulation. Curiously, for PubMed passages, inserting explanatory content increased perceived and actual difficulty (decreased content question accuracy). However, results for Wikipedia passages appear consistent with our expectation, though not statistically significant.

We performed 2-way ANOVA for actual difficulty, perceived difficulty, and reading time ($N = 116$ without baseline condition).

**Actual Difficulty.** Readers scored much higher on Wikipedia (81.6%) than on PubMed passages (66.4%). The difference in difficulty is significant ($f_{1.90} = 4.222$, $p < 0.043$). No significant effects were observed for term-specific questions.

Contrary to our expectations, actual difficulty is not significantly affected by the misalignment of parentheses. However, the interaction effect between different source material and placement of parentheses is still evident, though not statistically significant. Figure 2 illustrates this. For passages from PubMed, integrating all explanatory content into the text yields higher accuracy scores, but for passages from Wikipedia, keeping explanatory content in parentheses yielded higher accuracy scores.

**Perceived Difficulty.** PubMed passages are perceived to be more difficult than Wikipedia passages ($f_{1.92} = 7.333$, $p < 0.008$). PubMed passages had an average perceived difficulty of 3.572, while Wikipedia passages’ was 3.106.

**Reading Time.** None of our experiment variables were significantly related to reading time. As in Study 1, reading time did not impact our results and was not correlated with actual difficulty ($r = 0.107$, $p < 0.254$), term comprehension ($r = 0.017$, $p < 0.858$), or perceived difficulty ($r = 0.098$, $p < 0.294$).

Table 5. Demographics of Study 2 Subjects ($N = 42$)

<table>
<thead>
<tr>
<th>Age</th>
<th>Count (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;30</td>
<td>13 (0.309)</td>
</tr>
<tr>
<td>31 to 40</td>
<td>17 (0.405)</td>
</tr>
<tr>
<td>41 to 50</td>
<td>9 (0.214)</td>
</tr>
<tr>
<td>51 to 60</td>
<td>1 (0.024)</td>
</tr>
<tr>
<td>61 to 70</td>
<td>2 (0.048)</td>
</tr>
<tr>
<td>&gt;71</td>
<td>0 (0.0)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Gender</th>
<th>Count (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>22 (0.524)</td>
</tr>
<tr>
<td>Male</td>
<td>20 (0.476)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Race (multiple answers allowed)</th>
<th>Count (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>American Indian/Alaska Native</td>
<td>0 (0.0)</td>
</tr>
<tr>
<td>Asian</td>
<td>5 (0.116)</td>
</tr>
<tr>
<td>Black</td>
<td>3 (0.070)</td>
</tr>
<tr>
<td>Native Hawaiian/Pacific Islander</td>
<td>0 (0.0)</td>
</tr>
<tr>
<td>White</td>
<td>35 (0.814)</td>
</tr>
<tr>
<td>More than one race</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Education</th>
<th>Count (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Less than high school</td>
<td>0 (0.0)</td>
</tr>
<tr>
<td>High school diploma</td>
<td>12 (0.286)</td>
</tr>
<tr>
<td>Associate's degree</td>
<td>9 (0.214)</td>
</tr>
<tr>
<td>Bachelor's degree</td>
<td>15 (0.357)</td>
</tr>
<tr>
<td>Masters</td>
<td>5 (0.119)</td>
</tr>
<tr>
<td>Doctorate</td>
<td>1 (0.024)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Language Spoken at Home</th>
<th>Count (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Never/Rarely English</td>
<td>0 (0.0)</td>
</tr>
<tr>
<td>Half English</td>
<td>0 (0.0)</td>
</tr>
<tr>
<td>Mostly English</td>
<td>2 (0.048)</td>
</tr>
<tr>
<td>Only English</td>
<td>40 (0.952)</td>
</tr>
</tbody>
</table>

Table 6. Means for Actual Difficulty, Perceived Difficulty, and Reading Time for each Experimental Condition

<table>
<thead>
<tr>
<th>Document Source</th>
<th>Placement of Parentheses</th>
<th>N</th>
<th>Actual Difficulty (all)</th>
<th>Actual Difficulty (term)</th>
<th>Perceived Difficulty</th>
<th>Reading Time</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Mean</td>
<td>Std.Dev</td>
<td>Mean</td>
<td>Std.Dev</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>Baseline</td>
<td>15</td>
<td>0.556</td>
<td>0.272</td>
<td>0.600</td>
<td>0.507</td>
</tr>
<tr>
<td></td>
<td>All inside</td>
<td>21</td>
<td>0.873</td>
<td>0.247</td>
<td>0.857</td>
<td>0.359</td>
</tr>
<tr>
<td></td>
<td>Mismatched</td>
<td>23</td>
<td>0.710</td>
<td>0.307</td>
<td>0.739</td>
<td>0.449</td>
</tr>
<tr>
<td></td>
<td>All outside</td>
<td>20</td>
<td>0.783</td>
<td>0.224</td>
<td>0.850</td>
<td>0.366</td>
</tr>
<tr>
<td>PubMed</td>
<td>Baseline</td>
<td>17</td>
<td>0.686</td>
<td>0.322</td>
<td>0.647</td>
<td>0.493</td>
</tr>
<tr>
<td></td>
<td>All inside</td>
<td>18</td>
<td>0.593</td>
<td>0.334</td>
<td>0.611</td>
<td>0.502</td>
</tr>
<tr>
<td></td>
<td>Mismatched</td>
<td>18</td>
<td>0.685</td>
<td>0.242</td>
<td>0.833</td>
<td>0.383</td>
</tr>
<tr>
<td></td>
<td>All outside</td>
<td>16</td>
<td>0.771</td>
<td>0.315</td>
<td>0.750</td>
<td>0.447</td>
</tr>
</tbody>
</table>
Although we were able to significantly reduce the actual difficulty of the text using optimal strategies identified in Study 1, Study 2 did not provide conclusive results about the effect of parallel parentheses for simplifying a list of difficult terms. Therefore, in a post hoc analysis, we examined the properties of the lists in our study. Six of the eight Wikipedia passages included lists of similar terms (terms are semantically related, such as “systolic” and “diastolic”), while five of the PubMed passages were lists with dissimilar terms. Parallel parentheses may highlight the similarity between related terms and make reading easier. The significant main effects we observed in Study 2, in which Wikipedia passages are actually and perceived to be easier than PubMed material, may be partially attributed to the similarity between the terms in the lists. However, in this study, similarity is also confounded with the difficulty of the source material.

Limitations

The small scale of the study, with only 16 passages in each experiment, is a limitation. We selected and edited the passages to restrict the number of difficult words and the position in the sentence. This allowed us to precisely control different parameters in the experiment. To draw more generalizable conclusions, future research should use passages that were originally published with difficult terms in positions of interest in the sentence. Additionally, we used Wikipedia articles and PubMed abstracts to represent simple and difficult texts, respectively. However, in addition to lexical difficulty, these texts also differ in writing style, which is a nuanced concept we can’t address in the scope of this study.

Conclusion and Lessons Learned

Through our studies, we found that parenthetical expressions can simplify text, but only under the correct conditions. Augmenting the text with simple explanations for difficult terms can reduce the difficulty of the text. However, there exists a significant interaction between difficulty of the text overall and whether to enclose the difficult term or its explanation in the text. Modifying the text without taking these factors into account will be ineffective or even counter-productive to text simplification.

The following text simplification lessons result from our two studies:

1. Difficult content benefited from putting the target term (i.e., difficult term) inside parentheses and incorporating the explanatory content into the text. In Study 1, PubMed passages with the term inside the parentheses had lower actual difficulty, while Wikipedia passages with an explanation in the parentheses had lower actual difficulty. The same pattern is observed in Study 2, though it is not statistically significant. Also in Study 1, in difficult texts, the texts were easier when a definition (i.e., difficult text) was incorporated in the text rather than in parentheses. A possible explanation is that parentheses provide a clear boundary for snippets of text. When the text is more difficult, the reader becomes unwilling to read the difficult material and will simply skip to the end of the parentheses. By incorporating explanatory content into difficult text, or difficult definitions into any text, readers are encouraged to read the difficult content and benefited from absorbing the information.

2. If we want to add explanatory content to a text, adding a simple explanation helps more than a detailed, technical explanation. Practically, we need to generate more resources with consumer-friendly explanations of medical terminology, which is part of our ongoing research.

3. The position of the target term does not significantly affect actual or perceived difficulty of the text. Even though parenthetical expressions interrupt the flow of a sentence, the location of the interruption is not significant. Its impact on the reader’s processing of the sentence is overshadowed by the difficulty of the content.

4. We have preliminary but non-significant evidence suggesting that parallelism in a list may play a role in simplification. We can consider parallelism as the similarity between difficult concepts in a list, which indicates a consistency in the theme or semantic types. Another aspect of parallelism is using parentheses and other cohesion markers in a consistent manner. This is a topic for future research.
5. In general, simplification strategies are inconsistent across different types of text. For example, PubMed passages in Study 2 became more difficult after employing parenthetical insertion in non-optimal configurations. From patients to caregivers to health professionals, the medical domain produces many types of text, ranging from electronic health records to scientific research to general guidebooks. Simplification strategies need to be tailored to the task at hand to be effective.

Acknowledgements

Research reported in this publication was approved by the Institutional Review Board at the University of Arizona, and supported by the National Library of Medicine of the National Institutes of Health, under Award Number R01LM011975. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

Understanding the Patterns of Health Information Dissemination on Social Media during the Zika Outbreak

Xinning Gui, MA¹, Yue Wang, MS², Yubo Kou, PhD³, Tera Leigh Reynolds, MPH, MA¹, Yunan Chen, PhD¹, Qiaozhu Mei, PhD², Kai Zheng, PhD¹
¹University of California, Irvine, Irvine, CA; ²University of Michigan, Ann Arbor, MI; ³Purdue University, West Lafayette, IN

Abstract

Social media are important platforms for risk communication during public health crises. Effective dissemination of accurate, relevant, and up-to-date health information is important for the public to raise awareness and develop risk management strategies. This study investigates Zika virus-related information circulated on Twitter, identifying the patterns of dissemination of popular tweets and tweets from public health authorities such as the CDC. We leveraged a large corpus of Twitter data covering the entire year of 2016. We analyzed the data using quantitative and qualitative content analyses, followed by machine learning to scale the manual content analyses to the corpus. The results revealed possible discrepancies between what the general public was most interested in, or concerned about, and what public health authorities provided during the Zika outbreak. We provide implications for public health authorities to improve risk communication through better alignment with the general public’s information needs during public health crises.

Introduction

Social media have become important information hubs, where individuals and organizations create and disseminate real-time content beyond their personal social networks and physical location. As a result, social media have become a powerful channel for information seeking and sharing, especially in times when timely information is critical, e.g., during crisis breakout⁴. An increasing number of public health authorities are seizing this opportunity to improve risk communication during public health emergencies (e.g., disease outbreaks, natural disasters)⁵,⁶. For instance, during the H1N1 influenza pandemic, governmental institutions such as the U.S. Centers for Disease Control and Prevention (CDC) and Department of Health & Human Services (HHS) disseminated information to the general public through social media and other websites⁵,⁶. Similarly, during the 2011 South East Queensland floods crisis, state and local authorities shared information and updates through this avenue⁴. The existing research assessing using social media for risk communication in practice, suggests that social media can be a powerful channel for communication during crises. For example, several studies have shown that using social media is highly beneficial for effective and fast risk communication during disease outbreaks and disasters⁷,⁸.

Although many public health organizations use social media in risk communication, it is unclear whether the general public actually receive such information properly⁹. More evaluations of the effectiveness of crisis-specific risk communication efforts are needed for developing best practices in social media use⁹. In this sense, research that investigates the distribution and determinants of public health messages on social media has important implications. First, public health organizations can monitor and evaluate the general public’s attitudinal and behavioral reactions during emergency situations. Second, public health organizations and practitioners can learn to adopt more effective communication strategies to improve the dissemination of health-related knowledge, announcements, and recommendations⁹. For example, by identifying a surge of misinformation, public health professionals can adapt their communication efforts to counterbalance the misinformation on social media⁹. Similarly, investigating surges in information demand can provide insights on common information gaps that exist and, thus, inform what kinds of information to disseminate⁹.

In this study, we analyzed risk communications during a 2016 public health emergency, the Zika virus outbreak, on one social media platform, Twitter. Our goal was to understand how public health authorities, and the general public, communicated on Twitter during the Zika virus outbreak, and to provide implications for effective risk communication efforts. Most people, at least outside of infectious disease physicians and researchers, had not heard of the Zika virus prior to the recent, highly publicized outbreak. On February 1, 2016, the World Health Organization (WHO) declared a cluster of infants born with microcephaly and other neurologic disorders in Brazil to be a public health emergency of international concern (PHEIC)⁹. WHO noted that, while Zika virus infection was strongly suspected to be the cause, it had not yet been scientifically proven⁹; thus, declaring the cluster to be a PHEIC signaled the need for an urgent international effort to fill this research gap⁹,¹⁰. By November 18, 2016, WHO
reported that there was now sufficient evidence that Zika virus infection was the cause, and ended this PHEIC12. While this likely resulted in reduced media and public attention, it does not mean that the Zika virus is no longer a public health threat or priority. On the contrary, WHO described the consequences of Zika virus infection as “a highly significant long-term problem” that requires a “robust longer-term response mechanism,” and stated that WHO was shifting focus and resources accordingly12. Therefore, public health organizations will most likely continue to communicate risk information related to the Zika virus.

Based on existing research, Twitter is the best social media platform for our research goal. Among the various platforms, Twitter has frequently been used by public health authorities, such as WHO, the U.S. National Institutes of Health (NIH), and local health departments, to publish information about disease outbreaks14–16. Therefore, it is very likely that many public health organizations also disseminated risk information during the Zika virus outbreak. Recent research also shows the great potential of mining and analyzing Twitter data that can be used as “a proxy measure of the effectiveness of public health messaging or public health campaigns”17.

We collected a 10% sample of all tweets that contain “Zika” and were circulated on Twitter throughout 2016 (from January 1st to December 31st), as this was the time period when Zika obtained the most international attention (due to WHO’s announcement). We analyzed the typology of these tweets and their relation to tweets that were the most retweeted. Using the same strategy, we also analyzed tweets that came from authoritative accounts. Authoritative accounts are those that are affiliated to public health organizations. Unlike in noncontagious diseases, epidemics and pandemics necessitate effective and timely communications from public health organizations, as they need the general public’s cooperation to mitigate the situation. Our case study of risk communication during the Zika virus outbreak can shed light on how public health organizations could (1) improve their ongoing Zika-related communications, and (2) better leverage Twitter to target the general public’s information needs, communicate more effectively and reach more people in shorter amount of time during public health crises.

Methods

We used a mixed-methods approach that combines quantitative and qualitative content analysis methods to analyze a sample of tweets. To scale up the mixed approach to a large collection of tweets, we leveraged machine learning to assist in qualitative coding. Our general analysis flow contains seven steps (see Figure 1 for a flow chart):

**Figure 1.** Flow of our methodology. Each block is described in detail in text.

(a) Through Twitter’s gardenhose API, we can access 10% of the entire Twitter data stream. We prepared our dataset by selecting tweets that contained the keyword “Zika” in the year 2016. By doing this, we yielded a dataset consisting of 1,495,480 tweets. Only English tweets were used in the manual content analysis. English tweets account for 54% of all Zika-related tweets.

(b.1) We defined “popular tweets” to be those that had been retweeted at least 100 times. We identified popular tweets based on Twitter’s “retweet” data field. This gave us 3,581 unique tweets, which we refer to as the “popular set.”

(b.2) We sorted all the accounts ranked by “# of mentions” (using @ token), and manually selected 17 authoritative accounts that are dedicated to public health issues, and represent public health agencies, medical experts, and institutions from top 300 accounts. Though these accounts themselves do not publish many tweets, they are frequently mentioned (using @ token). Table 1 shows the top mentioned authoritative accounts. During the accounts selecting process, we excluded some indirect accounts such as POTUS (the U.S. president’s Twitter account) and JuanOrlandoH (Official account of the President of the Republic of Honduras). These 17 authoritative accounts gave us 1,227 unique Zika-related tweets, which we refer to as the “authoritative set.”
(c) We sampled 200 tweets from the popular set. From the authoritative set, we sampled five tweets from each of the 17 authoritative accounts. In total, we had 285 tweets. Employing a grounded-theory approach\(^\text{18}\), each of the two authors coded the content of these 285 tweets. They first read the tweets to get a general sense of the content. After that, each author followed the open coding process to code the 285 tweets and generate a set of initial codes independently. Through several rounds of reading, coding, and comparing emerging data to existing themes, the two authors generated a set of codes that emerged naturally to describe the tweet content. The two authors discussed and consolidate the codes through the axial coding process. The final codebook contains a set of 8 codes for the sample of 285 tweets.

<table>
<thead>
<tr>
<th>Table 1: Taxonomy of 17 authoritative accounts we selected.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Category</strong></td>
</tr>
<tr>
<td>---------------</td>
</tr>
<tr>
<td>CDC official</td>
</tr>
<tr>
<td>WHO official</td>
</tr>
<tr>
<td>Federal official</td>
</tr>
<tr>
<td>Academic or medical expert/Institution</td>
</tr>
<tr>
<td>State or local official</td>
</tr>
</tbody>
</table>

(d) We used a machine learning approach to generalize from the manually coded tweets and assign codes to the remaining tweets. Using coded tweets as the training set, we developed a text classification model. The model was a one-versus-rest multiclass logistic regression classifier, taking in the content of a tweet, and predicting the most likely category. The preprocessing steps included lowercasing, removing common punctuations (".", "," and ",") and normalizing URLs (replacing by the token "URL"). Each tweet was represented as bag-of-unigram feature vectors with binary values. The classifier was trained using LibLinear package, with default regularization weight \(C = 1\)^19.

Table 2 shows each code’s definition, example, frequency in the sample, frequency in the whole dataset, and top informative content keywords with highest information gain in each category.

<table>
<thead>
<tr>
<th>Table 2: Typology of tweet content related to the Zika virus (continued).</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Category</strong></td>
</tr>
<tr>
<td>---------------</td>
</tr>
<tr>
<td>Joke</td>
</tr>
<tr>
<td>Policy</td>
</tr>
<tr>
<td>Research Progress</td>
</tr>
<tr>
<td>Infection Update</td>
</tr>
<tr>
<td>Sports Events</td>
</tr>
</tbody>
</table>
(e) We manually checked randomly sampled tweets with those applied codes to ensure the accuracy of the prediction. To ensure that the text classification model generates accurate predictions, we also conducted 10-fold cross-validation on the 285 coded tweets to evaluate the quality of the text classification model. Cross-validation is a commonly used statistical procedure for estimating the quality of machine learning models on uncoded data instances. For multi-class classification, the accuracy, macro-averaged precision, and macro-averaged recall are defined as

$$\text{accuracy} = \frac{\sum_{y \in Y} TP_y}{N}$$
$$\text{macro-averaged precision} = \frac{1}{|Y|} \sum_{y \in Y} \frac{TP_y}{P_y}$$
$$\text{macro-averaged recall} = \frac{1}{|Y|} \sum_{y \in Y} \frac{TP_y}{N_y}$$

Where $Y$ is the set of 8 categories; $TP_y$ is the number of tweets correctly predicted to belong to category $y$; $N$ is the total number of tweets; $P_y$ is the number of tweets predicted to belong to category $y$; $N_y$ is the number of tweets actually belong to category $y$. Our text classifier achieved 79.6% accuracy, 75.2% macro-averaged precision, and 74.0% macro-averaged recall on the multi-class classification task, which is reasonably good. These predicted codes were used as a basis for further content analysis.

(f) We generated and compared typological difference between popular tweets and authoritative tweets.

**Results**

**Temporal and Geographical Characteristics of Zika-related Tweets**

We first describe general characteristics of the whole zika related tweets. As the word “Zika” is shared across languages, these tweets are sent from regions speaking diverse languages, where English (54%), Spanish (27%), and Portuguese (12%) are among the top three. Figure 2 shows the timeline of these tweets, annotated with important events. To cross-validate our data collection, we also show the Google Trend of “Zika” for the entire year. The two timelines are consistent. We found that tweets peaked often right after a major event. The tweet amount peaked around late January, during which locally transmitted infection had been reported from more than 20 countries. The severity of the outbreak triggered a massive amount of discussions on Twitter. In July and August, Twitter users paid more attention to Zika again, as the 2016 Olympic Games took place in August in Brazil, the place where the Zika outbreak originated from. Interestingly, the trajectory of Tweet amount rose and dropped sharply during the first peak, possibly because the initial shocking news lasted very briefly. By contrast, the trajectory of Tweet amount during the Olympic Games covers a much longer period of time with gradual increase and decrease, possibly because the general public’s attention to the Olympics spanned over longer time and discussion continued prior, during and after it, since the Olympic Games was a major event that lasted for weeks.

Using 1% of these tweets that have GPS information, we visualize the worldwide discussion of Zika in an evolving heatmap. Due to the scope and space of this paper, we only show two screenshots from mid-January and mid-February (see Figure 3).
The discussion was heated in the first three months of 2016. It originates from countries nearby the tropical region, especially South American countries. At the end of January, it has attracted attention in many countries, and swept the world in February. We notice that the spots on heatmap correlates with monthly infection update worldwide. By mid-February, Zika cases had been reported in Latin and Central America, the United States, Africa, Southeast Asia, among other countries and regions \(^{20}\).

Typology of Zika-related Tweet Content

We classified all the English tweets into eight categories. Within the typology, joke and sports events are the least relevant to the Zika virus itself, compared to other categories that directly point to scientific discovery and effective countermeasures. A closer read of tweets in these two categories suggested that joke was prevalent possibly because of the characteristics of Twitter as a source of entertainment; and sports events was also popular mostly because the 2016 Olympic Games were held in Brazil, where the Zika started to spread. We posit that during a health crisis, critical information such as infection update, knowledge, and progress should be disseminated as broadly as possible. Knowing the basic types of tweet content, it is crucial to understand whether critical information about health crisis is related to the tweet’s popularity, in terms of retweet number and like number. Based on the typology, we explore descriptive statistics of these categories to understand each category’s distinct characteristics. Table 4 shows each type’s length, retweet number, and like number.

Table 4 shows that jokes were the most popular type of content on Twitter, in terms of retweet (avg=1362) and like numbers (avg=1425). We suggest that it is possibly because Twitter’s one primary function is to entertain. This finding correlates with other work reporting humor as a major type of content on social media\(^ {21}\). Such information might have negative consequences to Zika-related public communication, as the amusement caused by these
widespread jokes might dilute the general public’s awareness and caution of the severity of Zika infection. For example, the top retweeted tweet was an English translation of a Portuguese tweet:

BrazilStats2: USA goalkeeper joked about Brazil being dangerous due to Zika. Every time she hit the ball, Brazil fans scream ZIKA. https://t.co/2dZnY1qY4R

Table 4. Descriptive Statistics of Top Retweeted Tweets (# of retweets >= 100).

<table>
<thead>
<tr>
<th>Category</th>
<th>Retweet count average (std. dev.)</th>
<th>Like count average (std. dev.)</th>
<th>Percentage (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Joke</td>
<td>1362 (3706)</td>
<td>1425 (3040)</td>
<td>1.6</td>
</tr>
<tr>
<td>Policy</td>
<td>289 (341)</td>
<td>376.8 (683)</td>
<td>34.2</td>
</tr>
<tr>
<td>Research Progress</td>
<td>419 (672)</td>
<td>301.4 (529)</td>
<td>9.5</td>
</tr>
<tr>
<td>Infection Update</td>
<td>228 (260)</td>
<td>126.6 (196)</td>
<td>31.9</td>
</tr>
<tr>
<td>Sports Events</td>
<td>336 (452)</td>
<td>317.0 (311)</td>
<td>0.4</td>
</tr>
<tr>
<td>Consequence</td>
<td>704 (922)</td>
<td>450 (680)</td>
<td>1.5</td>
</tr>
<tr>
<td>Scientific Knowledge</td>
<td>253 (381)</td>
<td>197 (399)</td>
<td>20.6</td>
</tr>
<tr>
<td>Pharmaceutical Progress</td>
<td>1192 (1509)</td>
<td>607 (787)</td>
<td>0.4</td>
</tr>
</tbody>
</table>

The Twitter account then only had 15771 followers, but the tweet was retweeted 21982 times. Noticeably, among these top retweeted tweets, factual information related to policy, infection update, and scientific knowledge had the least popularity. Each of these three received averagely less than 400 retweets and less than 570 likes. For example, an authoritative “infection update” tweet from CDCMMWR that was aimed to disseminate latest report but only retweeted twice, possibly because the reports are lengthy and require sufficient health literacy to understand:

CDCMMWR: Read @CDCgov's latest #Zika reports on the @CDCMMWR website: https://t.co/lwLgL1RQKol
https://t.co/ysP6vhlur

Such huge discrepancy in popularity between different types of content signifies Twitter’s ineffectiveness in disseminating critical health information during a public health crisis.

After discussing popular tweets’ characteristics, now we examine authoritative tweets. In Table 5 we discuss the tweets from 17 authoritative accounts. Among the 1,277 authoritative tweets, only 27 are retweeted more than 100 times (“popular”). We found that authoritative tweets do not contain content related to joke and sports events. This phenomenon correlates with our previous interpretation that joke and sports are the least relevant to Zika itself. Public health authorities might consider it inappropriate and does not match their authoritative identity.

Table 5. Descriptive Statistics of Authoritative Tweets (continued).

<table>
<thead>
<tr>
<th>Category</th>
<th>Retweet count average (std. dev.)</th>
<th>Like count average (std. dev.)</th>
<th>Percentage (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Joke</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>0</td>
</tr>
<tr>
<td>Policy</td>
<td>26 (38)</td>
<td>19 (30)</td>
<td>17.0</td>
</tr>
<tr>
<td>Research Progress</td>
<td>17 (32)</td>
<td>12 (24)</td>
<td>9.1</td>
</tr>
<tr>
<td>Infection Update</td>
<td>25 (45)</td>
<td>87 (56)</td>
<td>24.8</td>
</tr>
<tr>
<td>Sports Events</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>0</td>
</tr>
<tr>
<td>Consequence</td>
<td>34 (73)</td>
<td>14 (23)</td>
<td>0.7</td>
</tr>
<tr>
<td>Scientific Knowledge</td>
<td>43 (177)</td>
<td>22 (65)</td>
<td>47.6</td>
</tr>
</tbody>
</table>
Comparing Table 4 and Table 5, we noticed categorical differences between popular tweets and authoritative tweets. People retweet jokes and sports events, which can be problematic in conveying the critical information about health crisis, especially when those jokes and events include unverified information. Other than jokes and sports events, the general public was interested in research progress, consequence, and pharmaceutical progress, in terms of numbers of retweet and like numbers. However, most of the authoritative tweets fall into the categories of policy, infection update, and scientific knowledge. This difference suggests that in terms of public communication on Twitter, it seems that the general public’s attention and the information that authoritative accounts attempt to disseminate diverge significantly: the general public is seeking and broadcasting information that tells them the results of Zika infection and what can be done against it. However, public health authorities seem to tend to publish objective, verified information in a scientific, rational tone. This discrepancy and its influence are worthy of future research.

Temporal Development between Popular and Authoritative Tweets

We compare the difference between popular and authoritative tweets in terms of their typological distribution. Figure 4 shows each type of tweets’ development across time, where popular tweets are highlighted in blue and authoritative tweets in red.

As shown in Figure 4, popular tweets mostly correlate with the overall trajectory in Figure 2, peaking from January to early February and in early August. This is understandable because social media oftentimes mirror the general public’s attitude, concern, and interest in real time. However, authoritative tweets do not correlate with the temporal trajectory of major Zika events (See Figure 2.). Public health authorities broadcasted information at a regular pace that did not match the general public’s interest and concern. This can be easily observed in the diagrams of policy, research progress, infection update, scientific knowledge, and pharmaceutical progress. Public health authorities also tweeted more about consequence at the beginning of 2016, possibly because such information was still new to the general public at that time. However, they stopped broadcasting such information later. This suggests that when public health authorities used Twitter, it seems that they showed little timely sensitivity to major news and events. It is also possibly because public health authorities possess limited capacity in generating popular, or “viral,” content.

Discussion

Our results indicate that Twitter does support widely and real-time communication during public health emergencies. It was widely used by the general public to discuss the Zika virus outbreak on a timely manner, suggesting that it is an important channel during public health emergencies.

By analyzing Zika-related tweets, we found that both the general public and public health authorities actively used Twitter as a platform to circulate Zika-related information. However, authoritative accounts constantly published factual information such as infection update and established scientific knowledge, and their tweets rarely gained as

<table>
<thead>
<tr>
<th>Category</th>
<th>Retweet count average (std. dev.)</th>
<th>Like count average (std. dev.)</th>
<th>Percentage (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pharmaceutical Progress</td>
<td>73 (94)</td>
<td>34 (44)</td>
<td>0.2</td>
</tr>
</tbody>
</table>
much popularity as tweets published by the general public, suggesting there may be discrepancies between what the general public is most interested in, or concerned about, and what public health authorities are providing during the risk communication processes. Indeed, the two had vastly different interests in terms of content, with the former being more interested in joke, sports, research and pharmaceutical progress, and consequence, while the latter focused on infection update, policy, and scientific knowledge. One possible reason for these differences is that, contrary to the general public, public health authorities have a responsibility to provide the most accurate information possible, are accountable for the content of their tweets and, often, must obtain approval (from multiple people) before content is released to the general public. Any error could cause unexpected harm, and erode public confidence in the institution. While these concerns are very real, our results suggest that Twitter messages from public health authorities are not reaching as broad of an audience as Twitter affords. Public agencies such as CDC use multiple ways (e.g., dedicated websites and press releases) to broadcast authoritative information, the traffic of which our collection of Twitter data cannot capture. However, the limited social media interaction and communication between the general public and public health agencies in our data might have negative consequences considering the popularity of Twitter, such as general Twitter users spreading rumors about public health situations, or even referencing pseudoscience. Previous studies of social media communication found that authorities often do not correct rumors quickly enough—at that point, they have already been widely transmitted\textsuperscript{22–25}. During our manual coding process, we indeed identified a set of unverified information from the popular tweets published by the general public. This suggests that while social media is often imagined as a singular space where information flows freely, divisions exists, and some types of information (e.g., infection updates) only circulates in specific populations (e.g., authoritative accounts’ followers), which can be a barrier to effective risk communication during health crises.

As people increasingly seek health information on social media, previous research has demonstrated the benefits and the necessity of directing people to trustworthy information sources\textsuperscript{26–28}. Therefore, it is important to promote the accounts of public health authorities on social media, so that accurate information triumphs over unreliable information, i.e., unverified and speculative. By analyzing the dissemination of Zika-related health information, we identified implications that may improve the reach and effectiveness of public health authorities’ future Zika-related communications, and public health emergencies more broadly.

First, our results show that popular tweets are often not directly relevant to Zika itself and contains noise such as jokes. This finding indicates that social media users may prefer entertaining and engaging content. Thus, rather than mainly doing one-way information dissemination, public health authorities may conduct more interactive communication strategies, such as building conversations with the general public, answering questions from ordinary citizens, and produce more engaging messages. For example, even with consideration of previously identified challenges, it is possible for public health authorities to produce more engaging content, as evidence by CDC’s viral Zombie Apocalypse campaign which teaches about emergency preparedness\textsuperscript{29}. Also, another possible reason that public health authority messages did not gain popularity is that the messages (e.g., scientific reports) may be lengthy and requires sufficient health literacy to understand. When publishing messages on social media, authoritative accounts may consider restating scientific messages in plain language. Furthermore, our study shows that, the general public cared more about Zika’s research progress, consequence, and pharmaceutical progress, probably because these are pragmatically related to their everyday lives, countermeasures, and future plans. Similarly, a study of travel decision making during the Zika crisis shows that people actively sought information from both authorities, local residents, and previous travelers in order to make informed travel decisions\textsuperscript{30}. To better attend to the general public’s information needs, authoritative accounts can engage in such conversations to both interact with the general public and correct rumors, providing more details about the research and pharmaceutical progress.

Second, based on our analysis of temporal patterns of zika-related tweets, people’s interests in Zika correlated with the latest news and major events. This suggests that public health authorities may consider publishing more timely content related to influential news in order to ease the panic, and to proactively provide information regarding the general public’s foreseeable concerns, such as major international sports events. Additionally, noting that Zika attracted the largest number of tweets from January to February, 2016, possibly because of its newness, we suggest that public health authorities should pay special attention to these moments as opportunities to engage and educate the public. Moreover, we found that public health authorities focused on information that had been verified or clinically proven, such as policy, infection update, and scientific knowledge. However, such information often takes time to be developed after an event. Thus, we suggest that public health authorities publish previous verified or just
other related information right after influential news was published or the major event happened, which may draw more public interests and also likely to get them more engaged with the information.

Third, our analysis suggests that monitoring information dissemination trends on social media can be an effective way of understanding the general public’s interests and concerns. Public health authorities might consider improving capacity in social media monitoring, gaining familiarity with major conversations and debates that take place among the general public. In this way, public health authorities may have better chances in getting involved in these public discussions and broadcasting necessary knowledge.

Fourth, our study also indicates a disconnection between the public’s understanding of a public health situation and available, scientific information sources provided by public health agencies in various ways. The dissemination of rumors highlights the disconnection. It is an urgent question as to how public health authorities can raise the public’s health literacy and eHealth literacy in taking full advantage of authoritative, scientific information to understand public health issues and make informed health-related decisions.

Limitation

We based our analysis and drew conclusions on a portion of data from one social media platform. Therefore, our findings and suggestions may not generalize to other social media, websites, or traditional media, or other crisis situations. Our study is formative rather than definitive. Further hypothesis tests are necessary to examine our findings.

Conclusion

This research concerns risk communication during a public health crisis induced by the Zika virus. We employed mixed-method approach and machine learning to understand the content of both popular tweets and authoritative tweets. We discussed the information gap between the general public and public health authorities and practitioners. We suggested implications that may help improve public health authorities’ risk communication strategies, including providing more engaging and straightforward health message contents that attend to people’s information needs, adopting more interactive communication strategies, delivering messages timely after related news and major events, conducting social media surveillance and raising the public’s health literacy and eHealth literacy.

References


19. Machine Learning Group at National Taiwan University. LIBLINEAR -- A Library for Large Linear Classification. 2015. Available from: https://www.csie.ntu.edu.tw/~cjlin/liblinear/


StressHacker: Towards Practical Stress Monitoring in the Wild with Smartwatches

Tian Hao, Ph.D.1, Kimberly N. Walter, Ph.D.1, Marion J. Ball, Ed.D.1, Hung-Yang Chang, Ph.D.1, Si Sun, Ph.D.1, Xinxin Zhu, M.D., Ph.D.1
1IBM T.J. Watson Research Center, Yorktown Heights, NY 10598

Abstract

In modern life, the nonstop and pervasive stress tends to keep us on long-lasting high alert, which over time, could lead to a broad range of health problems from depression, metabolic disorders to heart diseases. However, there is a stunning lack of practical tools for effective stress management that can help people navigate through their daily stress. This paper presents the feasibility evaluation of StressHacker, a smartwatch-based system designed to continuously and passively monitor one’s stress level using bio-signals obtained from the on-board sensors. With the proliferation of smartwatches, StressHacker is highly accessible and suited for daily use. Our preliminary evaluation is based on 300 hours of data collected in a real-life setting (12 subjects, 29 days). The result suggests that StressHacker is capable of reliably capturing daily stress dynamics (precision = 86.1%, recall = 91.2%), thus with great potential to enable seamless and personalized stress management.

Introduction

There is mounting evidence that psychological stress from work-related pressures, financial concerns, and family responsibilities is an important public health issue in the United States1. A majority of Americans report frequently feeling stressed and that stress negatively impacts their physical and mental well-being2. Further, psychological stress is associated with adverse health outcomes such as cardiovascular disease, diabetes, depression, and substance use disorders3–4. Despite the evidence that stress negatively affects health, there are few effective and practical stress management solutions that can be seamlessly integrated into an individual’s daily life. In order to develop truly personalized, targeted stress management solutions, we need to be able to accurately and reliably identify occasions of stress in real-world settings.

Daily stress processes are typically assessed via self-report measures and/or biomarkers of stress system functioning, which are limited in their ability to capture the dynamic nature of daily stress. Self-reported stress cannot be assessed continuously and consequently may not capture all stressful events5. Further, it is subject to recall issues and self-report bias. Biomarkers of stress commonly used in real-world settings include cortisol, ambulatory blood pressure, and heart rate variability (HRV). Neither cortisol nor blood pressure can be measured continuously and they are also fairly invasive6, 7. Although HRV can be assessed passively and continuously using commercially available wrist and chest sensors, these wearable sensors do not provide automated information about stress system functioning in real-time. Recently, numerous research efforts8–10 have been made to advance the technologies for assessing stress in real-world scenarios using a variety of sensors including ECG (electrocardiography), RIP (respiratory inductance plethysmography), camera, microphone, etc. However, these technologies have limited practicality as they either require custom devices8 or are cumbersome and therefore not suited for daily use9, 10.

The present study aims to address these limitations by evaluating the feasibility of using a novel HRV-based stress monitoring system to passively, continuously, and accurately capture occasions of stress in a person’s daily life. HRV is a commonly used measure of autonomic nervous system (ANS) functioning, which is one of the two primary stress systems – the other being the hypothalamic-pituitary-adrenal axis11. The ANS consists of the sympathetic nervous system (SNS; “fight or flight”) and the parasympathetic nervous system (PNS; “rest and digest”), which work in opposition of each other to control stress reactivity and recovery12. HRV measures the variation in time between two consecutive heart beats13 and numerous studies have demonstrated that HRV is a valid measure of reactivity to acute psychological stressors14–16. Our HRV-based stress monitoring system, called StressHacker, uses photoplethysmography (PPG) sensor data that is available in commercially available wearable devices. StressHacker does not require the use of a custom biosensor devices. Rather, it can be easily integrated into commercially available smartwatches allowing for seamless integration into individuals’ daily lives. Thus, StressHacker has the potential to make automated and
personalized stress management a reality by detecting stress and offering real-time stress reduction recommendations in real-world settings.

In this paper we present preliminary data from an in-field study evaluating StressHacker’s ability to monitor stress in people’s daily lives. In summary, StressHacker is a system empowered by mobile sensing technologies that can translate continuous PPG signal or RR intervals into stress index – an indicator that is reflective of the change of stress level. At the current stage, we evaluate StressHacker using PPG data collected from wrist-worn wearables. However, the underlining algorithms of StressHacker can be tailored to take similar signals from other devices as input (e.g., electrocardiogram from chest band sensor). Ultimately, our hope is that through a set of data-driven analytical tools, StressHacker will be able to uncover powerful insights about daily stress processes by identifying sources of stress, recognizing patterns, and making personalized stress reduction recommendations. Possible real-world applications of StressHacker include corporate wellness programs, telemedicine in populations highly vulnerable to stress (e.g., chronic health conditions, substance use disorders, depression, anxiety), and automobiles to increase safety and enhance the driver’s experience.

In-field Experiment

The primary goal of the in-field study is to assess the feasibility of continuously monitoring daily stress using smartwatches. To this end, we have recruited 12 subjects who are IBM employees, among which 8 are female and 4 are male. Most of the subjects are young adults. All the subjects voluntarily agreed to contribute to the data collection, and signed a consent form.

In order to simulate real-life scenarios, all of the experiments were un-controlled with the following procedure. Prior to data collection, the researchers briefly informed the subject of the purpose of the experiment, types of data to be collected, and their responsibilities. Then a wrist-worn wearable device (Empatica E4 wristband) was provided to the subject. The device is designed to record bio-signals for research and development purposes, and equipped with the same sensors we intend to use in smartwatches. The device’s extended battery life allows a continuous data collection for up to 2 days. The recorded data was stored locally in the flash memory of the device, and was later downloaded for further analysis. We demonstrated to each subject how to start and stop the recording. The subjects were allowed to start and stop the recording anytime, however, they were informed that a data collection that covers most of their daily activities is preferred (e.g., a recording starting in the morning before work and ending in the evening after work or before sleep).

After the recording was started, data including pulse wave and acceleration was continuously sampled from the PPG heart rate sensor and accelerometer, respectively, and stored locally for off-line examination. In the current stage of our evaluation, the subjects did not get any feedback from or have any interaction with the system during data collection. To simulate real-life scenarios, the subject was informed to conduct their daily activities as they normally do during data collection. After the data collection was completed, the subject was asked to provide a log containing a list of their daily activities covered by the recording. Each entry should include a time period and a brief description of the

<table>
<thead>
<tr>
<th>Time</th>
<th>Activity</th>
<th>Stress Information</th>
<th>PSL</th>
</tr>
</thead>
<tbody>
<tr>
<td>6:30-7:30</td>
<td>Getting ready for work.</td>
<td>No particular emotions, routine</td>
<td>1</td>
</tr>
<tr>
<td>7:30-8:56</td>
<td>Driving, 1-hour traffic jam on the bridge</td>
<td>Anxious about being late.</td>
<td>2</td>
</tr>
<tr>
<td>8:56-11:06</td>
<td>Working</td>
<td>Attentive, stuck at a problem.</td>
<td>2</td>
</tr>
<tr>
<td>11:06-11:31</td>
<td>Lunch and responding to emails</td>
<td>Relaxed</td>
<td>1</td>
</tr>
<tr>
<td>11:31-12:53</td>
<td>Working</td>
<td>Attentive</td>
<td>2</td>
</tr>
<tr>
<td>12:53-14:15</td>
<td>At a seminar and working</td>
<td>Attentive. Stressed due to the multitasking.</td>
<td>4</td>
</tr>
<tr>
<td>14:15-15:00</td>
<td>Coffee break</td>
<td>Relaxed but still thinking about things.</td>
<td>2</td>
</tr>
<tr>
<td>15:00-17:00</td>
<td>Meeting</td>
<td>Attentive. The first half more stressed.</td>
<td>3</td>
</tr>
</tbody>
</table>

Table 1: An example of the logs provided by the subjects reporting their daily activities and the associated stress information (optional). The stress information can be given in the form of a short descriptive sentence and/or perceived stress level (PSL) on a scale from 1 to 10. In the log shown in this table, both descriptions and PSLs are provided.
activity, for example, “1:00-2:30 group meeting.” In addition, to help us better evaluate StressHacker, we also asked the subjects to report stress information associated with each reported activity in the form of a short description and/or perceived stress level (PSL) on a scale from 1 to 10. In the evaluation, such stress information was considered as references of their daily stress dynamics, and compared with StressHacker’s result. Table 1 shows a log provided by one of the subjects as an example. Note that the stress information is optional, meaning the subjects can choose not to include stress information for certain or all of the activities in the log.

Figure 1: An example of StressHacker’s stress level output based on a 10-hour recording with annotated activities and stress information from the log shown in Table 1.

Dataset

We have conducted in-field experiments across 29 days, collecting recordings from 12 subjects. The durations of the recordings range from 3.78 hours to 46.6 hours (see Table 2 for details). The total duration of all recordings is 306.2 hours, containing a total of 237 reported activities, 130 of which were reported with associated stress information. Among the 130 activities with stress information, 102 of them were labeled as stressful.

After the recording device was started, it continuously recorded pulse wave signal (i.e., blood volume pulse) from PPG heart rate sensor (sampling rate = 64 Hz), and acceleration from accelerometer (sampling rate = 32 Hz), along with their timestamps. The pulse wave collected from the PPG sensor was used to calculate HRV-based features, which in turn were used to derive stress levels. The collected motion data was used to estimate the motion level, which was primarily used to gauge the quality of the PPG signal. Specifically, we first extracted RR intervals from pulse wave. Then the RR intervals contained within a 5-min window were used to calculate short-term HRV features that have been studied and suggested to be a good indicator of stress level\(^{18,19}\). We adopted the time-domain methods\(^{20}\) to calculate two HRV features that reflect the total power and the high frequency power of HRV. Finally, we fed the resulted features into a model that maps the features to a stress level ranging from -1 to 1, where -1 indicates very relaxed, and 1 represents very stressed.

Figure 1 demonstrates StressHacker’s continuous stress level output using data from a typical in-field experiment as an example. The final output is a time series of stress levels ([-1, 1]) with a 5-minute interval. The stress level is categorized into four states: Relaxed [-1, -0.2], Less Relaxed (-0.2, 0], Slightly Stressed (0, 0.2] and Stressed (0.2, 1]. Note that in the current stage of evaluation, we use a fixed mapping from HRV features to stress level, and a pre-defined categorization of stress states. However, in practice, the system will be able to adjust the related parameters over time according to the user’s feedback.
Table 2: A list of detailed information for each individual experiment. N(Activity) is the number of activities reported by the subjects. N(StressInfo) indicates the number of reported activities with stress information. N(Stress) is the number of activities that have stress information and are labeled as stressful activity. The evaluation results are represented in TP (true positive), FN (false negative), TN (true negative) and FP (false positive). They are obtained by comparing StressHacker’s stress level output against the stress information reported by the subject (precision=86.1%, recall=91.2%, F-measure=88.6%).

<table>
<thead>
<tr>
<th>Exp ID</th>
<th>Duration (Hr)</th>
<th>N(Activity)</th>
<th>N(StressInfo)</th>
<th>N(Stress)</th>
<th>TP</th>
<th>FN</th>
<th>TN</th>
<th>FP</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>30.85</td>
<td>17</td>
<td>10</td>
<td>8</td>
<td>8</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>B1</td>
<td>11.77</td>
<td>7</td>
<td>6</td>
<td>5</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>C1</td>
<td>7.57</td>
<td>11</td>
<td>9</td>
<td>6</td>
<td>6</td>
<td>0</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>C2</td>
<td>7.41</td>
<td>7</td>
<td>6</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>D1</td>
<td>7.42</td>
<td>18</td>
<td>10</td>
<td>10</td>
<td>9</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>E1</td>
<td>5.74</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>E2</td>
<td>8.11</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>E3</td>
<td>6.37</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>F1</td>
<td>14.29</td>
<td>9</td>
<td>8</td>
<td>5</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>F2</td>
<td>7.5</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>F3</td>
<td>46.6</td>
<td>20</td>
<td>16</td>
<td>13</td>
<td>11</td>
<td>2</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>G1</td>
<td>8</td>
<td>9</td>
<td>6</td>
<td>4</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>G2</td>
<td>8.13</td>
<td>5</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>H1</td>
<td>22.31</td>
<td>9</td>
<td>6</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>H2</td>
<td>21.79</td>
<td>9</td>
<td>6</td>
<td>5</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>I1</td>
<td>15.49</td>
<td>18</td>
<td>12</td>
<td>6</td>
<td>6</td>
<td>0</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>I2</td>
<td>10.11</td>
<td>8</td>
<td>7</td>
<td>6</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>J1</td>
<td>8.88</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>J2</td>
<td>3.78</td>
<td>5</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>J3</td>
<td>9.54</td>
<td>6</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>K1</td>
<td>6.68</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>K2</td>
<td>7.15</td>
<td>6</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>K3</td>
<td>7.4</td>
<td>6</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>K4</td>
<td>8.1</td>
<td>8</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>L1</td>
<td>7.46</td>
<td>12</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>L2</td>
<td>7.74</td>
<td>12</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>total</td>
<td>306.19</td>
<td>237</td>
<td>132</td>
<td>102</td>
<td>93</td>
<td>9</td>
<td>15</td>
<td>15</td>
</tr>
</tbody>
</table>

Result

In this section, we evaluate the feasibility of continuously monitoring daily stress using smartwatches using quantitative metrics that reflect the correlation between StressHacker’s output and subject-reported stress information. Therefore, the evaluation is based on the 20 in-field experiments in which the subject provided stress information (i.e., experiments in Table 2 with N(StressInfo)>0).

First, we identify “stressful” activities based on the reported stress information. Specifically, we classify an activity as “stressful” when its associated stress information satisfies one of the following two conditions: (1) it contains word(s) such as “stressful”, “stressed”, “nervous” or “anxious”; (2) its associated perceived stress level (PSL) is larger than the baseline 1 (in a few cases, subjects reported PSL=2 as relaxed, thus we used 2 as baseline in such cases).

Next, for the purpose of calculating the correlation, we determine a similar binary label (“stressful” or “not stressful”) for each activity using StressHacker’s output within the activity’s associated time period. Specifically, an activity will be labeled as “stressful” if the stress levels within its time period satisfy one of the following two conditions: (1) at least one of the stress levels are categorized as “Stressed” (0.2, 1]; (2) at least 10% of the stress levels are categorized...
Lastly, we compare the binary labels determined by the self-report stress information and StressHacker’s output. The detailed result of the comparison is shown in Table 2. We can see StressHacker labels 108 activities (93TP+15FP) as “stressful,” with 93 classifications being true positives, leading to a precision of 86.1%. Out of 102 “stressful” activities reported in the subjects’ logs, 93 activities are correctly labeled by StressHacker, resulting in a recall (sensitivity) of 91.2%. Overall, StressHacker achieves an accuracy of 81.8% in the binary classification of “stressful” activity. We believe that such evaluation results are sufficient to suggest that the stress level information provided by StressHacker can be used to reliably monitor the dynamics of one’s daily stress.

Discussion

In this section, we look beyond the scope of quantitative evaluation and discuss StressHacker’s potential for offering valuable and timely insights that could fuel a broad range of decision makings in not only stress management at an individual level but also mental wellness at an organization or community level.

![Figure 2: Comparing StressHacker’s stress level output across three typical work scenarios.](image)

We first look at StressHacker’s stress level output in three typical work scenarios reported in the in-field experiment, which are working individually (83 hours), passively attending a meeting (20 hours) and actively engaging in a meeting (52 hours). As all of the in-field experiments were conducted during week days, the combination of these three scenarios accounts for more than half of the overall data collection. Figure 2 shows the comparison of stress levels. A key observation is that, overall, the subjects were most stressed while actively engaging in a meeting (40% of the time), compared with while working individually (26% of the time) and during a passive meeting (15%). Such insight can also be supported by the percentages of the “relaxed” state detected by StressHacker in each scenario, which suggests that the subjects were less relaxed during active meetings (only 6% of the time), compared with 15% of the time when working individually and 20% of the time during passive meetings. As all of the 12 subjects are employees at IBM Research, we speculate that the result above could be reflective, to some extent, of the overall “stress profile” of the organization under different working scenarios.

In addition to the work scenarios, we have also captured a typical non-work scenario that was labeled as “getting ready in the morning” (reported by 4 subjects in 6 days). Figure 3 shows the percentages of different stress levels. We can see that, although it has higher percentage of time being relaxed, the overall percentage of time being stressed (41%) is similar to that in the scenario of actively engaging in a meeting. Such insights offered by StressHacker could potentially
be used to identify major stressors and provide personalized recommendations to help users navigate through daily stress. For instance, when significant stress has been detected on every Monday morning, the system could recommend trying to get up half an hour earlier.

**Figure 3:** Percentages of stress levels detected by StressHacker in a typical non-work scenario labeled as “getting ready in the morning.”

**Figure 4:** Demonstrating StressHacker’s ability to reveal fine-grained and insightful stress dynamics using two daily events captured in the in-field experiments as examples. The continuous stress level detected by StressHacker and the self-report activity/stress information are plotted over time.

Next, we zoom in on StressHacker’s continuous stress level output, and use two examples captured in the in-field study to showcase that StressHacker’s result allows us to take a closer look at the dynamics of stress associated with an event of interest, which in turn could potentially help us make more informed decision to minimize the negative impacts of stress. Figure 4 (a) shows that StressHacker captures the subject’s stress level changes before, during and after
her presentation. During the time when she was preparing for the meeting, we can observe the transition from being relaxed to slightly stressed, which could be the effect of working under a close deadline and/or the anticipation of the coming presentation. Interestingly, around the 200th minute, StressHacker’s result shows that she recovers from being slightly stressed to less relaxed, which could be because that she had finished the preparation resulting in a reduction in her stress from working under deadline. Unsurprisingly, StressHacker’s result shows a substantial increase in her stress level especially in the first half of her presentation, which is confirmed by the self-reported stress information where the subject reported as being “nervous”. In the course of the following group discussion, we can see the subject stayed at being slightly stressed for about 10 minutes before she started recovering and became relaxed at around the 260th minute.

The second case shown in Figure 4 (b) captures the subject’s stress dynamics while watching a close basketball game. From StressHacker’s result, we can see that overall, the subject was stressed throughout the game, with a brief recovery during the half-time break. Interestingly, StressHacker’s result reveals the subject’s transition from being slightly stressed to stressed in both the first and the second half of the game, which might reflect the effect of stress accumulation. Another interesting observation is that, in the second half of the game, the subject only stayed being slightly stressed for about 15 minutes (from around the 425th to the 440th minute) until his stress elevated to a higher level (shown in dark bars), whereas the duration of which in the first half of the game was about twice as longer (from around the 350th to the 380th minute).

We believe that StressHacker’s ability to offer such fine-grained and real-time stress monitoring can significantly advance the automation of personalized stress management. To put it into perspective, we compare StressHacker with cStress, a stress monitoring system proposed in a recent research study. Compared with StressHacker, cStress relies on data collected from a larger set of sensors including inductive plethysmography (RIP), a two-lead electrocardiograph (ECG) and 3-axis accelerometers, which are integrated into a chest band that users need to wear all the time. An in-field study had been conducted to evaluate cStress in real-life scenarios based on a similar task to the one used in our study (identifying “stressful” and “not stressful” activities and comparing with subjects’ self-report data). The evaluation result suggests that cStress achieves a median accuracy of 72%, whereas StressHacker’s overall accuracy is 81.8%. Therefore, although only relying on data from PPG sensor, a less reliable heart rate sensor compared with ECG, StressHacker is still able to provide reliable results that are sufficient for daily stress monitoring.

Limitations and Future Work

In the paper, we focus on evaluating the feasibility of using bio-signals obtained from off-the-shelf wearables to infer daily stress. A major cause of the limited 81% accuracy is that data collected from PPG sensor is susceptible to various interferences. We have examined the collected data in attempts to identify factors that might affect the robustness of the system. Such factors include daily activities that involve relatively high level of motion (e.g., walking up/down stairs). As PPG sensor is susceptible to motion artifacts, such activities can lead to unreliable or even corrupted sensor data, which in turn, result in inaccurate stress detection results. As part of our future work, we aim to resolve this limitation by developing algorithms that can identify and mitigate the impacts of motion artifacts. To further improve the robustness, we plan to leverage the user’s historical data to better understand the user’s response to a certain context. For instance, the system might be able to learn over time that the user rarely gets stressed when s/he is driving home, and therefore classifies an isolated stress state with a short duration incorrectly detected in that scenario as false alarm.

As a next step of this study, we will investigate how users can address their stress experiences in real time and how not to distract users during high concentration activities. We are in the process of conducting a study aimed at providing users with real time data on their stress levels via a smartwatch. We intend to make these devices adaptive and personalized to deliver context aware notifications and stress management recommendations. This means that the notification and recommendations should not be delivered when users are performing high concentration tasks or activities. The printout of stress level output shown in Figure 1 is the first step toward such a real-time stress feedback and management system. In future studies we will investigate if the notifications help individuals become more mindful of stress from common daily activities and if increased mindfulness is associated with improved well-being. Further, we will evaluate if completion of real-time stress management recommendations (e.g., meditation, deep breathing) is associated with improved well-being.
Conclusion

In this paper, we present the result of our uncontrolled in-field study that aims to evaluate the feasibility of continuous daily stress monitoring using smartwatches. Based on more than 300 hours of data collected in real-life setting, the quantitative results of the evaluation suggest that the proposed system, StressHacker, can be used as a practical tool for effective and personalized stress management. Moreover, we have also presented analyses at both individual and community levels to showcase StressHacker’s potential to offer valuable data and insights that can be used to advance the automation and personalization of stress management and promote mental wellness in general.

References


A mobile system for the improvement of heart failure management: Evaluation of a prototype

Sarah C. Haynes, MPH1, Katherine K. Kim, PhD, MPH, MBA1
1University of California Davis, Sacramento, California

Abstract

Management of heart failure is complex, often involving interaction with multiple providers, monitoring of symptoms, and numerous medications. Employing principles of user-centered design, we developed a high-fidelity prototype of a mobile system for heart failure self-management and care coordination. Participants, including both heart failure patients and health care providers, tested the mobile system during a one-hour one-on-one session with a facilitator. The facilitator interviewed participants about the strengths and weaknesses of the prototype, necessary features, and willingness to use the technology. We performed a qualitative content analysis using the transcripts of these interviews. Fourteen distinct themes were identified in the analysis. Of these themes, integration, technology literacy, memory, and organization were the most common. Privacy was the least common theme. Our study suggests that this integration is essential for adoption of a mobile system for chronic disease management and care coordination.

Introduction

Heart failure is a major public health concern, affecting more than 5.7 million Americans1 and projected to increase 46% by 2030.2 Heart failure is the most common cause of hospitalization among adults over age 653 and accounts for more than 30 billion dollars in health care spending each year.1 Management of heart failure is complex, often involving interaction with multiple providers, continuous monitoring of symptoms and measurements, and numerous medications. Newly diagnosed heart failure patients are faced with the challenge of adapting to these major lifestyle changes and often struggle to successfully self-manage.4 As a result, heart failure readmissions are common. More than 50% of patients are readmitted within six months of initial discharge.1 Additionally, heart failure patients often suffer from one or more comorbidities; a large national study of Medicare patients found that 40% of heart failure patients also had diabetes, and around one-third had chronic obstructive pulmonary disease.5 A recent systematic review of depression in heart failure patients found an average prevalence of depression of 29%.6 Thus, patients with heart failure often have the task of managing multiple chronic conditions, adding further complexity to self-management.

Several heart failure management interventions have been associated with decreased readmissions and mortality.7-10 Characteristics of these interventions associated with improved outcomes include face-to-face contact,8 use of multidisciplinary teams,8,9 and longer intervention duration.10 However, because these interventions are extremely resource-intensive scalability is limited. A 2016 systematic review found 34 existing mobile systems for heart failure self-management.11 While data collected from these types of systems have the potential to inform healthcare decisions, existing applications are largely limited to the collection of physical measurements and symptom tracking and lack integration with the health system.11 Furthermore, most existing mobile systems for care management do not facilitate two-way communication, which has been found to be preferred by patients using heart failure monitoring systems.12 Finally, most self-management applications are focused on a single disease, rather than allowing patients to manage multiple conditions. These limitations reflect a lack of attention to patient-centeredness in the design of current mobile systems. Patient-centered applications are defined as “systems that enable a partnership among practitioners, patients, and their families (when appropriate) to ensure that procedures and decisions respect patients’ needs and preferences.”13 Such systems can facilitate care coordination, in which all members of the care team work collaboratively with the patient to make decisions about care and facilitate delivery of services.14 This type of shared decision-making requires that patients have a comprehensive understanding of their health condition, treatment options, and the roles and responsibilities of each care team member.15,17 User-centered design can help to create patient-centered applications for care coordination by understanding how patients would use the system in their daily lives and building features and functionalities around patient preferences and use cases.13,18,19 User-centered design may also increase the likelihood of adoption and satisfaction with the system.20,21 The OnPoint study was initiated to improve understanding of how patient-centered mobile
systems may improve care coordination with for people with multiple chronic conditions, such as those with heart failure and other comorbidities.

**Objectives**

The objectives of this study were to 1) assess the major strengths and weaknesses of a high-fidelity prototype of the OnPoint application, a mobile system for the management of complex chronic disease, with patients with heart failure and key members of the care team and 2) identify additional functions and features that would support adoption of the application by patients and members of the care team.

**Materials and methods**

We used principles of user-centered design to design and evaluate a high-fidelity prototype of OnPoint, a patient-centered mobile application for collaborative chronic disease management. A high-fidelity prototype is an application that is similar to the intended final product, allowing the user to experience realistic interactions; this is in contrast to a low-fidelity prototype, which may be rendered on paper or screenshots. In a previous paper we reported on interviews with key informants, including patients, clinicians, caregivers, and health coaches that resulted in a list of key user functions and user scenarios, around which we created the prototype. The prototype focused on the medication reconciliation and medication taking, priorities identified by both patients and providers: constructing a comprehensive medication list, creating a medication schedule, filling a pillbox, and tracking medications. Other features included care team contacts, tracking weight/blood pressure, managing appointments, and screening for symptoms. When possible, we created digital representations that complemented the tools that patients use in daily life for care coordination (the user-interface metaphor). For instance, because patients often use pillboxes to manage medications, we created an interface that evokes a pillbox metaphor to connect the digital representation on the tablet to the patient’s physical pillbox. The novel medication management features allow a user to take a picture of the medication bottle label which scans the data into the application, inserts a picture of the pill itself, sorts medications into appropriate day/time compartment in the rendered physical pillbox, which all serve to help the patient organize medications accurately and efficiently. The high-fidelity prototype was developed using Ionic, an open-source software development kit that allows creation of hybrid mobile applications. Adobe PhoneGap was deployed to allow development of a cross-platform application that would run on both iOS and Android. Figure 1 shows several screens from the high-fidelity prototype.

We recruited participants from the UC Davis Health System to test the high-fidelity prototype. We selected patient participants who had experience living with or managing heart failure along with another health condition as well as clinicians or health coaches who provide care for heart failure patients. Participants tested the mobile system on a 7.9-inch internet-connected iPad mini tablet during a one-hour one-on-one session with a facilitator. Following a structured patient scenario, the facilitator instructed the participant to attempt each feature of the prototype. For 1) the medication management feature, participants were asked to a) add several medications, b) create a schedule of medications, c) fill a pillbox using the application, d) record the taking and skipping of medications, and e) view medication history. For 2) the measurements feature, participants were asked to a) set up new measurements (weight, blood pressure, and heart rate), b) add measurements to the schedule, c) record measurements, and d) view measurement history. For 3) the appointments feature, participants were asked to enter new appointments and view appointments in the timeline. Finally, participants were asked to use 4) the comments feature of the timeline to record notes and questions to use share with other members of the care team. The symptom management feature was not ready to be tested at the time of the interviews, and was therefore not included in this evaluation. We applied a think-aloud methodology to elicit participants’ thoughts and perceptions of use of the application. The facilitator took notes capturing both verbalized feedback and visually observed challenges with using the application. The facilitator then interviewed participants about the strengths and weaknesses of the prototype, additional necessary features and functionalities, and willingness to use the technology, following a semi-structured interview guide.

Field notes and interview notes were analyzed using a three-phase qualitative content analysis process: 1) immersion, consisting of numerous close readings of the transcripts, 2) reduction, in which themes were identified, defined, and coded, and 3) interpretation, in which themes were organized and conclusions were drawn and verified. Initial codes were inductively developed by one investigator (SH) on analysis of the
first three interview transcripts. The codes and their associated themes were discussed in detail with a second investigator (KK) and revised. The revised coding scheme was applied to the remaining transcripts. A final review of transcripts and codes was completed by the two investigators with discussion of any discrepancies with the goal of consensus on issues identified. A basic quantitative content analysis consisting of counting the number of times a theme was mentioned by any participant was calculated to provide a high level indication of the themes that are “top of mind” among the participant group. The UC Davis Institutional Review Board concluded that this study was not human subjects research (IRB# 782917-1).

Figure 1. Selected medication features of OnPoint high-fidelity prototype, including A. Medication setup, B. Medication scheduling, C. Pillbox filling feature, and D. Timeline view showing list of today’s tasks
Results

Eight individuals participated in the testing of the prototype. These included five patients with heart failure who also manage multiple health conditions, including cancer, diabetes, hypertension, and depression. Three care team members participated including a cardiologist, a cardiac care nurse, and a primary care health coach. Themes fell into three main categories: strengths and potential benefits, potential barriers to use, and features and functions for future development. Fourteen distinct themes were identified during the content analysis. These themes are shown in Table 1 along with a count of the number of times the theme was mentioned, and the proportion (%) of patient and care team members who mentioned the theme at least once.

Table 1. Definitions, overall number of mentions, and percent of participants who mentioned the theme, for each theme identified.

<table>
<thead>
<tr>
<th>Theme</th>
<th>Definition used for coding</th>
<th>Number of mentions</th>
<th>% of patients</th>
<th>% of care team</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strengths and Potential Benefits to Use of OnPoint Prototype</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Memory</td>
<td>Remembering tasks and activities</td>
<td>12</td>
<td>80.0</td>
<td>66.7</td>
</tr>
<tr>
<td>Organization</td>
<td>Arranging tasks and activities; structuring and storing information in a standard way</td>
<td>12</td>
<td>60.0</td>
<td>66.7</td>
</tr>
<tr>
<td>Convenience</td>
<td>Ease of using the system</td>
<td>7</td>
<td>60.0</td>
<td>33.3</td>
</tr>
<tr>
<td>Sharing</td>
<td>Sharing of information with others, including clinicians, caregivers, family members, and other members of the care team</td>
<td>7</td>
<td>40.0</td>
<td>100.0</td>
</tr>
<tr>
<td>Efficiency</td>
<td>Speed of tasks and activities</td>
<td>2</td>
<td>20.0</td>
<td>33.3</td>
</tr>
</tbody>
</table>

Potential Barriers to Use of OnPoint Prototype | | | | |
| Technology literacy | Affordances for users with different levels of familiarity with technology to use technology effectively | 16 | 80.0 | 100.0 |
| Burden | Challenges facing users that may inhibit effective use of the system | 10 | 40.0 | 66.7 |
| User types | Differences between user personalities and profiles that may affect how the system is used | 4 | 40.0 | 33.3 |
| Connectivity | Reliance of a technology system on wireless Internet | 3 | 0.0 | 66.7 |
| Privacy | Protection of personal information | 1 | 0.0 | 33.3 |

Features and Functions for Future Development | | | | |
| Integration | Linkage with other data sources, applications, or devices | 20 | 80.0 | 100.0 |
| Personalization | Tailoring of the system to meet an individual’s specific needs | 9 | 60.0 | 66.7 |
| Communication | Exchange of information between patients and providers | 6 | 40.0 | 66.7 |
| Tracking | Recording information (patient-generated) | 5 | 60.0 | 0.0 |

Integration was the most common theme overall, mentioned by 80% of patient participants and all of the clinician participants. Participants stressed that the usefulness of a mobile system depends upon its ability to integrate with other important sources of information. These sources include other applications such as calendars and applications that allow users to track salt, fluids, and other nutrition information; they also include devices such as physical activity trackers and Wi-Fi-enabled scales and heart rate monitors. The ability to link the application to clinical data, including lab results and prescription information, was
especially important to patient participants, particularly for those taking medications like warfarin that require continuous monitoring. Integration may also help to decrease the burden on patients that may be associated with setup and management of a mobile application:

“The application would need to be smart about how to pull data from multiple places; I wouldn’t want to enter the same information in two different places.” – Patient participant

Technology literacy, the second most common theme, was mentioned by both patient and clinician participants as a potential challenge to implementation. They expressed concern that those with low technology literacy, particularly many elderly patients, would need substantial support from others to set up and use the system. For this reason, participants thought that this group of people might be reluctant to adopt the technology:

“Setting up the application by yourself would be really difficult, particularly if you’re not familiar with technology.” – Health coach

Memory was the next most common theme. Several participants described memory as a common problem for heart failure patients. Participants believed that the system would improve the ability to remember both routine and infrequent tasks:

“A lot of people have brain fog from heart failure- mental reminders are helpful so you know which pills have been taken and which haven’t.” – Cardiac nurse

Participants felt that the system would improve the ability to organize data in one place so that information was available when it was needed. Patient participants reported that they currently used a variety of methods and systems to manage information; patients felt that this was a burden for daily self-management as well as for sharing information:

“Right now, I have tracking on lots of different devices, papers, calendars… it would be much easier if it was all in one place.” – Patient participant

“If you go to the ER, you could just show all of the information. It would be so much easier to provide your information when someone needed it because it’s all in one place.” – Patient participant

Importantly, participants also saw the benefit of the mobile system as something that could help them better understand their condition and their care plan, an important component of shared decision-making.

“This would be useful for familiarizing yourself with your medications… I realize I need a closer look at what’s going on with my meds.” – Patient participant

“I like that you can see the purpose of all the [medications] – you can really know what you’re taking and why.” – Patient participant

Several themes were more important to either patient participants or care team participants. Tracking, or the recording of patient-generated information, was mentioned by 60% of patient participants but mentioned by none of the care team participants. Conversely, several care team participants spoke about connectivity issues, while none of the patient participants mentioned connectivity as a potential barrier to use.

Privacy was the least common theme, mentioned by only one of the care team participants:

“Personal information protection is a concern - not everyone wants to have their wife or husband or adult children know what they’re taking.” – Cardiologist
Discussion

In their review of 34 mobile applications for heart failure, Masterson Creber et al. found that most applications are focused on building healthy habits, rather than on shared disease management. One reason for this may be the difficulty of integration with existing health care systems. Currently, incompatibility between systems and data sharing policies inhibit integration with mobile applications. In our study, the voice of patients themselves brings attention to a key challenge for chronic disease management: the challenge of unifying multiple data streams to create a single personalized system tailored to each individual patient’s needs. This type of system requires not only the ability to gather and organize patient data but also the capacity to learn and respond to information. Integration with the electronic health record (EHR) is one example of this. Clinicians commonly rely on EHR systems and patient portals for managing patient data and communicating with patients. Health system regulations, in addition to feasibility and workflow considerations, prevent adoption of a system that is separate from the EHR. Because shared information and communication are crucial components of successful disease management and effective care coordination, integration with the EHR should be a priority for any chronic disease management system. Our study suggests that integration is also crucial to patients’ adoption of a mobile system for chronic disease management. Because chronic disease patients often manage multiple comorbidities and have varying nutrition and activity requirements, fluid restrictions, or other specific measurements, integration with other devices and applications is necessary to ensure completeness of a shared care plan. A useful application will also have the ability to integrate with the patient’s existing management strategies, such as calendar or pharmacy applications. This is in line with previous findings on user preferences for self-management applications.

Our study suggests that integration of an application with existing external data sources and systems is more important than developing additional features; this is an important implication for prioritizing future work in this area. In addition, our findings indicate that there is a need for organized summary data for all types of users: patients, caregivers, and care team. Each type of user requires different types of information to make informed decisions. Development of data visualizations for these various users will require in-depth exploration into how users make decisions about care, and how data collected by a mobile system can assist in the decision-making process.

To create a usable application for chronic disease management, concerns about technology literacy must be addressed. Many previous studies have used mobile systems to deliver chronic disease management interventions. However, these studies often suffer from low usage of the technology, which may be in part a function of low technology literacy. To date, there have been few studies examining technology literacy in the context of mobile applications for disease management. Principles of user-centered design can help to create a system that can be used effectively even by those with little previous experience with mobile applications.

Although privacy was not mentioned by patient participants in this study, concerns about privacy are often mentioned with respect to use of technology. Some studies that have found low levels of concern about privacy threats with regard to mobile technologies. In a national public opinion survey of radio frequency identification (RFID) technology for healthcare, Katz and Rice found only a minority of participants who were concerned about threats to privacy. In a qualitative study by Pinnock et al., neither patients nor healthcare professionals considered privacy to be a major issue with regard to adoption of a healthcare management technology. Nevertheless, a recent statewide survey found high levels of concern regarding privacy in sharing of electronic health data for provision of healthcare and willingness to agree to share these data. Hence, assurance of privacy should continue to be considered in development and implementation of mobile care coordination solutions.

Findings of our study suggest several implications for building systems for care coordination. First, while reminders and tracking can be useful management tools, they are not sufficient for facilitating shared care. The patient participants in our study desired a thorough understanding of the health condition and care plan. Improved understanding of medications was seen as especially useful for patients. This suggests that patient-centered mobile systems should have the capacity to enrich a patient’s understanding of the disease and the treatment plans, instead of simply recording it. These findings align with existing literature on shared care and shared decision-making, which emphasize the informed participation of the patient in the creation of a care plan. Second, our findings indicate that the ability to share information quickly and easily is an
important part of a mobile system for chronic disease management. Access to the same information by everyone involved in the patient’s care will likely allow for improved coordination and collaboration among the care team, characteristics of care found to be associated with improved health outcomes.\textsuperscript{45}

There are several limitations to our study. First, we interviewed a small convenience sample of potential users, recruited from a single academic medical center in California. Although appropriate for a user-centered design study,\textsuperscript{46} the sample was not large enough to quantify differences between types of users or to discuss how preferences differed by age, disease severity, familiarity with technology, or other characteristics. Second, four of the participants (three patients participants and the cardiologist) were also involved in previous design phases, where they provided details on how they manage care and what aspects of care management and coordination are most important to them. For this reason, they may have been more inclined to provide positive feedback on the high-fidelity prototype, biasing the study results. As is common for qualitative studies such as the one present in this paper, the findings are not intended to be generalizable.

Next steps

We are currently using the feedback provided by patient and care team participants to develop the OnPoint prototype into a fully functional application for chronic disease management. The working version of the system will include additional features and integration with other systems, including specifications for EHR integration. We plan to use a participatory design process to test the application with patient and clinician participants in simulated home and clinical environments; we expect that this will provide insight into how a system can be optimized for different types of users. Our next steps will also include the creation and testing of data visualizations, to explore how patient-generated data collected by the application can be used for care coordination and shared decision-making. Understanding how these types of data will be shared and used is essential for building a system that can support coordination and management of multiple complex chronic diseases.

Conclusion

Although many mobile applications have been developed for the management of chronic disease, none has achieved widespread adoption. Our study provides insight into potential reasons for this by highlighting key barriers and facilitators of adoption of a mobile system for heart failure care coordination. The findings of this project emphasize the need for an integrated solution. Such a system will have the potential to improve self-management for patients managing multiple chronic conditions. It will also facilitate collaboration between patients and their care team, by allowing for the development and management of a shared care plan. Integration of the resulting care plan with the electronic health record is crucial for meaningful sharing of patient-generated information between patients and healthcare providers. The unique requirements, restrictions, and circumstances facing each patient require a flexible system that has the ability to capture and summarize information from multiple sources.

Acknowledgements

We thank Bjoern Hartmann, Dan Gillette, Angela Hsueh, Bill Kim, and Amy Wang for collaborating on the development of the prototype. This project was funded by University of California, Center for Information Technology in the Interest of Society (CITRIS).
References


Comparing and Contrasting A Priori and A Posteriori Generalizability Assessment of Clinical Trials on Type 2 Diabetes Mellitus

Zhe He, PhD1, Arturo Gonzalez-Izquierdo, PhD2, Spiros Denaxas, PhD2, Andrei Sura, BS3, Yi Guo, PhD3, William R. Hogan, MD3, Elizabeth Shenkman, PhD3, Jiang Bian, PhD3
1Florida State University, Tallahassee, FL, USA; 2University of College London, London, UK; 3University of Florida, Gainesville, FL, USA

Abstract

Clinical trials are indispensable tools for evidence-based medicine. However, they are often criticized for poor generalizability. Traditional trial generalizability assessment can only be done after the trial results are published, which compares the enrolled patients with a convenience sample of real-world patients. However, the proliferation of electronic data in clinical trial registries and clinical data warehouses offer a great opportunity to assess the priori generalizability during the design phase of a new trial. In this work, we compared and contrasted a priori (based on eligibility criteria) and a posteriori (based on enrolled patients) generalizability of Type 2 diabetes clinical trials. Further, we showed that comparing the study population selected by the clinical trial eligibility criteria to the real-world patient population is a good indicator of the generalizability of trials. Our findings demonstrate that the a priori generalizability of a trial is comparable to its a posteriori generalizability in identifying restrictive quantitative eligibility criteria.

Introduction

Clinical trials, which test the efficacy and safety of an intervention (e.g., medication, device, procedure, and behavioral change), are indispensable tools for evidence-based medicine [1]. However, the generalizability of clinical research studies has long been a concern [2]. For example, elderly patients are reported to be underrepresented in clinical trials across major medical conditions, including cardiovascular diseases [3], cancers [4], dementia [5], and diabetes [6]. Most research on the generalizability of clinical studies has focused on the a posteriori generalizability (i.e., the representativeness of enrolled participants), which compares the characteristics of enrolled patients of a study with a convenience sample of a real-world patient population [7] or those in other studies [8]. For example, van der Water et al. evaluated the external validity of a cancer clinical trial by comparing the socioeconomic status, number of comorbidities, treatments, and various stage information between the enrolled patients and the patients in the Netherlands Cancer Registry [7]. Cahan and colleagues proposed an a posteriori generalizability score that incorporates demographic information, clinical attributes, and clinical settings to compare a trial to multiple target clinical scenarios in other trials [8]. However, with such an a posteriori approach, the generalizability issue is not detected before the conclusion of the studies. Typical clinical trials cost over hundreds of millions dollars or more and take 10 - 17 years to complete [9]. Thus, it is crucial to assess a clinical trial’s generalizability before conducting the trial. Nevertheless, existing methods for assessing the a priori generalizability (i.e., the representativeness of eligible participants) have historically been scarce and laborious [10].

The rapidly growing amount of electronic patient data such as electronic health records (EHR) data, presents an unprecedented opportunity for optimizing eligibility criteria in the design phase of a new trial towards balanced internal and external validity (i.e., generalizability) [11]. In recent years, a suite of informatics methods has been introduced to quantify the population representativeness of clinical studies and characterize underrepresented population subgroups [12-14]. Notably, the Generalizability Index on Study Traits (GIST) metric quantifies the a priori generalizability of clinical trials with respect to selected quantitative eligibility criteria that specify a permissible value range (e.g., HbA1c > 7%), one at a time [12]. The extension of GIST, mGIST [15], can quantify the population representativeness of clinical trials with joint use of multiple criteria of interest. Both GIST and mGIST focus on the generalizability assessment at the disease domain level (i.e., assessing the generalizability of trials targeting the same disease). Recently, Sen and colleagues introduced GIST 2.0 as a scalable multivariate metric for quantifying the population representativeness of individual clinical studies by explicitly modeling the dependencies among all eligibility criteria [16]. However, to the best of our knowledge, no work has compared the a priori and the a posteriori generalizability assessment of trials.

Diabetes, recognized as an important public health problem by the World Health Organization [17], has caused 1.5 million deaths in 2012 alone and may, over time, lead to serious damage to the heart, blood vessels, eyes, kidneys,
and nerves. More than 400 million people live with diabetes [18]. Type 2 diabetes mellitus (T2DM), which can be developed at any age, accounts for 90% - 95% of people who have diabetes [19]. Many countries, including United States (US) and United Kingdom (UK), have invested heavily in research on treating and controlling diabetes [20, 21]. In this study, we compared and contrasted \textit{a priori} (i.e., using GIST, Weng et al. [12]) and \textit{a posteriori} (i.e., using van de Water et al. [7]) generalizability of T2DM clinical trials. We hypothesize that the \textit{a priori} generalizability of a trial is comparable to its \textit{a posteriori} generalizability in identifying certain restrictive quantitative eligibility criteria. To enable such a comparison, we will use univariate GIST metric to assess \textit{a priori} generalizability of T2DM trials with respect to three most frequently used quantitative eligibility criteria, age, HbA1c, and BMI. We chose GIST rather than its multivariable extensions, mGIST, as we want to compare individual variables' \textit{a priori} generalizability with their \textit{a posteriori} generalizability independently. We used the eligibility criteria of T2DM trials registered on ClinicalTrials.gov to profile the study populations, and extracted the published summary-level statistics of the enrolled patients. We compared the \textit{a priori} (based on the study populations) and the \textit{a posteriori} (based on the enrolled patients) generalizability of US-based T2DM trials using the target population profiled by the T2DM patients in the OneFlorida Data Trust [22], and further validated the results using UK-based T2DM trials with the target population profiled by the patients from the CALIBER research platform [23]. The OneFlorida Clinical Research Consortium (CRC) is one of the 13 Clinical Data Research Networks (CDRN) in the United States funded by the Patient-Centered Outcome Research Institute (PCORI) as part of the National Patient-Centered Clinical Research Network (PCORnet). The CALIBER resource generates and investigates deep, longitudinal phenotypic data from linked electronic health records for people registered in participating clinical practices in UK. The four main data sources include primary care EHR, hospital billing data, and death certificate records. Our work will inform the research community of the difference of \textit{a priori} and \textit{a posteriori} generalizability of T2DM trials with respect to quantitative eligibility features.

\section*{Background}

\textbf{ClinicalTrials.gov and the COMPACT Database}

ClinicalTrials.gov, created and maintained by the National Library of Medicine, is a clinical study registry in the United States. Since September 2007, all the United States-based clinical trials of FDA-regulated drugs, devices, and biologic products must be registered in ClinicalTrials.gov prior to participant recruitment. In September 2016, the United States Department of Health and Human Services issued the final rule of that expands the regulatory procedure for trial registration and summary results reporting [24]. Mandated by the final rule, trial sponsors are required to report summary statistics on race, ethnicity, and other measures assessed at baseline that are used in study descriptors such as study phase (i.e., Phase I, II, III, and IV), intervention type (e.g., drug, device, biologic product), locations, are stored in structured fields, whereas eligibility criteria are largely free-text.

To facilitate \textit{a priori} generalizability analysis, we leveraged the numeric expression extraction tool “Valx” [26, 27] and the frequent tag mining tool [28] to transform study summaries in ClinicalTrials.gov into a relational database, and we named it “COMPACT” [29]. COMPACT contains various study descriptors (e.g., study phase, intervention), and numeric eligibility features. With Valx, different names for the same variable in the eligibility criteria, such as “hemoglobin A1c”, “HbA1c”, “Glycohemoglobin”, are unified. Different measurement units are also unified. COMPACT indexes trials by medical conditions.

\textbf{OneFlorida Clinical Research Consortium and Data Trust}

The OneFlorida CRC is a collaborative statewide network that seeks to improve health research capacity and opportunities in the State of Florida through the facilitation of clinical and translational research in communities and health care settings. OneFlorida includes nine unique health systems that provides care for \textasciitilde{}9.7M or 48% of all Floridians through 4,100 physician providers, 1,240 clinic/practice settings and 22 hospitals with a catchment area covering all 67 Florida counties. In 2015, OneFlorida became one of 13 PCORI-funded clinical data research networks in the US.

The OneFlorida Data Trust is the centerpiece of the OneFlorida CRC and is the informatics infrastructure that supports pragmatic trials; comparative effectiveness research, implementation science, and other research in OneFlorida. The OneFlorida Data Trust currently contains collated EHR, health care claims, and other data on a
broad-based, unselected population of ~10 million people in Florida. The data are limited to a Health Insurance Portability and Accountability Act (HIPAA) Limited Data Set (LDS), which restricts the types of protected health information (PHI) to only dates (e.g., birthdates and dates of service) and location (to the zip code level).

**CALIBER (Clinical research using Linked Bespoke studies and Electronic health Records)**

CALIBER is a unique research platform consisting of ‘research ready’ variables extracted from linked EHR from primary and secondary care, social deprivation information and cause-specific mortality data in UK. Led from the University College London Institute of Health Informatics and the Farr Institute of Health Informatics Research, London, CALIBER enables researchers to recreate the longitudinal journey of patients through care pathways to study disease onset and progression. This research platform accesses linked electronic health records and recreates the healthcare pathways of approximately 10 million patients with 400 million person years of follow up. The aim of CALIBER is to foster an open community developing methods and tools to accelerate replicable science across all clinical and scientific disciplines spanning the translational cycle (from drug discovery through to public health). The resource consists of disease and risk factor phenotyping algorithms, methods [30], tools and scripts, specialized infrastructure and training and support. All finalized EHR phenotyping algorithms are provided in an open-access Portal (https://www.caliberresearch.org/portal) for researchers to extent and re-use.

**Methods**

We first define the three patient populations for a clinical trial:

- **Target population**: patients to whom the results of the clinical study are intended to apply. The target population can only be approximated with available patient data.
- **Study population**: patients who are eligible for the study based on the study inclusion and exclusion criteria.
- **Enrolled patients**: patients who are enrolled in the clinical study. It is a subset of the study population.

Figure 1 illustrates the analytical workflow of this study. We first retrieved interventional clinical studies on T2DM in the US and UK from the COMPACT database. In this work, we chose to focus on three major quantitative eligibility criteria in T2DM studies: age, HbA1, and BMI, which are used in the free-text eligibility criteria of 49.1%, 48.5%, and 43.98% T2DM trials, respectively [13]. Age is also a required field in the eligibility criteria section of a trial summary in ClinicalTrials.gov. We therefore used the structured age field. To profile the target population of T2DM, we identified all the T2DM patients in the OneFlorida Data Trust using both ICD-9-CM and ICD-10-CM codes. To validate the findings of US-based T2DM trials on UK-based trials, we identified T2DM patients from the CALIBER data with the ICD-10 codes in the hospital data and the Readcodes in the primary care data. We applied the GIST metric to assess the a priori generalizability of T2DM trials in the US and UK using the target populations derived from OneFlorida Data Trust and CALIBER, respectively. We also stratified the analysis by trial phase. We assessed the a posteriori generalizability by comparing the characteristics of the enrolled patients reported in ClinicalTrials.gov with the T2DM patients in OneFlorida Data Trust and CALIBER, respectively.

Figure 1. Analytical workflow of this study

**Dataset Preparation**

1) **Processing clinical trial summaries.** From the COMPACT database, we identified interventional clinical trials on T2DM with a study start date between January 2005 and September 2016. There are 1,671 such studies in the United States and 209 such studies in the United Kingdom.
2) Identifying T2DM Patients in OneFlorida Data Trust. Following existing literature [12], we identified patients with T2DM in OneFlorida Data Trust using the following criteria, where the patient (1) needs to have at least two diagnoses of Type 2 diabetes; (2) needs NOT have any Type 1 diabetes diagnoses; and (3) should have at least one HbA1c measurement regardless of their temporal relationships to diagnosis times. Diagnoses of Type 1 and 2 diabetes were identified with ICD-9-CM and ICD-10-CM diagnosis codes.

3) Identifying T2DM Patients in CALIBER. We utilized descriptive data from a previous study [31] using a deterministic to identify patients with Type 2 diabetes in CALIBER using diagnostic codes (Read codes in primary care, ICD-10 in secondary care).

Assessing a Priori Generalizability

To quantify the population representativeness of studies based on a single quantitative criterion (i.e., age, HbA1c, and BMI in this study), we calculated the univariate GIST scores for trial sets of different study phases [12]. The GIST score is the sum across all consecutive non-overlapping value intervals of the percentage of studies that recruit patients in that interval, multiplied by the percentage of patients observed in that interval:

\[
GIST = \frac{\sum_{j=1}^{N} \frac{I([i_{low}, i_{high}] \subseteq w_j)}{T} \times \frac{\sum_{i=1}^{P} I(y_i \leq i_{low} < i_{high})}{P}}{T}
\]

where \(N\) is the number of distinct value intervals of the quantitative feature, \(T\) is the number of trials, \(P\) is the number of patients, \(w_j\) is the inclusion value interval of the quantitative feature for the \(j^{th}\) study, such that indicator \(I\) can be defined as \(j^{th}\) study interval subsumes the \(j^{th}\) interval low and high boundary values, and \(y_i\) is the observed value of the quantitative feature for the \(k^{th}\) patient such that an indicator \(I\) can be defined when \(k^{th}\) patient has a value of the quantitative feature falls within the \(j^{th}\) interval. The GIST score ranges from 0 to 1, with 0 being not generalizable and 1 being perfectly generalizable. It characterizes the proportion of patients potentially eligible across trials. For more detailed explanation of GIST, see [12, 14]. We have previously evaluated the validity of the GIST metric in quantifying the population representativeness of trials using simulated patient populations [32]. Compared with mGIST which gives an overall score for multiple variables, GIST gives variable-specific scores, allowing us to compare them with the results of a posteriori generalizability assessment which are also variable-specific.

Assessing a Posteriori Generalizability

To compare enrolled patients with the target population, we identified all the US and UK T2DM trials that have reported results in ClinicalTrials.gov and extracted the baseline measures including the number of participants, their gender and race, and the mean and standard deviation (SD) values of the three major measures: age, BMI, and HbA1c. We aggregated the mean and SD for age, HbA1C, and BMI separately for all the trials that report both mean and race distributions with the real world population of T2DM patients in OneFlorida and CALIBER. We used the two-sample t-test to assess differences in quantitative variables (i.e., age, HbA1c, and BMI) and chi-square test to assess difference in categorical variables (i.e., race and gender) between the target population and enrolled patients.

Results

Basic Characteristics of T2DM Trials

Basic characteristics of the clinical trials on T2DM included in our analysis are shown in Table 1. Even though the number of trials differed significantly between the US and the UK, they exhibit similar characteristics. The rates of missing data for study phase are 23.6% and 27.3% in trials in the US and the UK respectively. A majority of trials are sponsored by industry (63.4% and 64.1%). Drugs were the most common interventions (72.1% and 72.7%).
followed by behavioral interventions (11.6% and 8.1%). Treatment was the primary purpose for the majority of the trials (75.9% and 77.0%). Most of the trials were randomized (86.9% and 90.1%).

Table 1. Characteristics of T2DM trials in the US and the UK

<table>
<thead>
<tr>
<th>Study Characteristics</th>
<th># of US-Based Trials (%) (N = 1,671)</th>
<th># of UK-Based Trials (%) (N = 209)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Study phase</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Phase 0</td>
<td>9 (0.5%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>Phase 1</td>
<td>299 (17.9%)</td>
<td>22 (10.5%)</td>
</tr>
<tr>
<td>Phase 2</td>
<td>349 (20.9%)</td>
<td>23 (11.0%)</td>
</tr>
<tr>
<td>Phase 3</td>
<td>495 (29.6%)</td>
<td>77 (36.8%)</td>
</tr>
<tr>
<td>Phase 4</td>
<td>189 (11.3%)</td>
<td>33 (15.8%)</td>
</tr>
<tr>
<td>Unspecified</td>
<td>394 (23.6%)</td>
<td>57 (27.3%)</td>
</tr>
<tr>
<td><strong>Sponsor type</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NIH</td>
<td>23 (1.4%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>Industry</td>
<td>1,060 (63.4%)</td>
<td>134 (64.1%)</td>
</tr>
<tr>
<td>Other U.S. Federal Agency</td>
<td>30 (1.8%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>Other</td>
<td>558 (33.4%)</td>
<td>75 (35.9%)</td>
</tr>
<tr>
<td><strong>Intervention type</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Drug</td>
<td>1,204 (72.1%)</td>
<td>152 (72.7%)</td>
</tr>
<tr>
<td>Procedure</td>
<td>34 (1.9%)</td>
<td>2 (1.0%)</td>
</tr>
<tr>
<td>Biological</td>
<td>32 (1.9%)</td>
<td>4 (1.9%)</td>
</tr>
<tr>
<td>Device</td>
<td>49 (2.9%)</td>
<td>9 (4.3%)</td>
</tr>
<tr>
<td>Behavioral</td>
<td>194 (11.6%)</td>
<td>17 (8.1%)</td>
</tr>
<tr>
<td>Dietary supplement</td>
<td>49 (2.9%)</td>
<td>14 (6.7%)</td>
</tr>
<tr>
<td>Genetic</td>
<td>1 (0.1%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>Radiation</td>
<td>1 (0.1%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>Other</td>
<td>107 (6.4%)</td>
<td>11 (5.3%)</td>
</tr>
<tr>
<td><strong>Primary purpose</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Basic science</td>
<td>108 (6.5%)</td>
<td>14 (6.7%)</td>
</tr>
<tr>
<td>Diagnostic</td>
<td>20 (1.2%)</td>
<td>2 (2.0%)</td>
</tr>
<tr>
<td>Education/Counseling/Training</td>
<td>1 (0.1%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>Health services research</td>
<td>40 (2.4%)</td>
<td>3 (1.4%)</td>
</tr>
<tr>
<td>Prevention</td>
<td>99 (5.9%)</td>
<td>17 (8.1%)</td>
</tr>
<tr>
<td>Screening</td>
<td>4 (0.2%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>Supportive care</td>
<td>33 (2.0%)</td>
<td>5 (2.4%)</td>
</tr>
<tr>
<td>Treatment</td>
<td>1,268 (75.9%)</td>
<td>161 (77.0%)</td>
</tr>
<tr>
<td>Unspecified</td>
<td>98 (5.9%)</td>
<td>7 (3.3%)</td>
</tr>
<tr>
<td><strong>Allocation</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Randomized</td>
<td>1,452 (86.9%)</td>
<td>190 (90.1%)</td>
</tr>
<tr>
<td>Non-Randomized</td>
<td>107 (6.4%)</td>
<td>3 (1.4%)</td>
</tr>
<tr>
<td>Unspecified</td>
<td>112 (6.7%)</td>
<td>16 (7.7%)</td>
</tr>
</tbody>
</table>

A Posteriori Generalizability of T2DM Trials

Among all the T2DM trials in our analysis, 428 (25.6%, out of 1,671) US trials and 86 (41.1%, out of 209) UK trials report summary-level results (e.g., baseline characteristics and outcome measures) in ClinicalTrials.gov. Table 2 illustrates the number of T2DM trials that report mean values and standard deviation values for age, HbA1c, BMI, as well as race and gender. Within the trials that provided these statistics, most of them provided the mean and standard deviation of age for all the enrolled patients. A higher percentage of UK-based trials reported mean and standard deviation values of HbA1c and BMI than the US-based trials. The primary reasons that the remaining trials did not report any results in ClinicalTrials.gov include (1) still under recruitment, (2) completed before December 6, 2007 and thus not required to submit results, and (3) pending results (results of applicable trials of FDA-regulated drugs, biologic, and device must be submitted within 12 months of trial completion [34]).
Table 2. Number of T2DM trials that reported results in ClinicalTrials.gov

<table>
<thead>
<tr>
<th>Results</th>
<th># of Trials in the US / Total # (%) (N = 1,671)</th>
<th># of Trials in the UK / Total # (%) (N = 209)</th>
</tr>
</thead>
<tbody>
<tr>
<td># of trials with any results</td>
<td>428/1,671 (25.6%)</td>
<td>86/209 (41.1%)</td>
</tr>
<tr>
<td># of trials reporting mean value of age</td>
<td>400/428 (93.5%)</td>
<td>80/86 (93.0%)</td>
</tr>
<tr>
<td># of trials reporting standard deviation of age</td>
<td>388/428 (90.7%)</td>
<td>79/86 (91.9%)</td>
</tr>
<tr>
<td># of trials reporting mean value of HbA1c</td>
<td>131/428 (30.6%)</td>
<td>38/86 (44.2%)</td>
</tr>
<tr>
<td># of trials reporting standard deviation of HbA1c</td>
<td>128/428 (29.9%)</td>
<td>38/86 (44.2%)</td>
</tr>
<tr>
<td># of trials reporting mean value of BMI</td>
<td>109/428 (25.5%)</td>
<td>30/86 (34.9%)</td>
</tr>
<tr>
<td># of trials reporting standard deviation of BMI</td>
<td>104/428 (24.3%)</td>
<td>30/86 (34.9%)</td>
</tr>
<tr>
<td># of trials reporting race</td>
<td>159/428 (37.1%)</td>
<td>34/86 (39.5%)</td>
</tr>
<tr>
<td># of trials reporting gender</td>
<td>426/428 (99.5%)</td>
<td>86/86 (100.0%)</td>
</tr>
</tbody>
</table>

We extracted the baseline characteristics of the patient's enrolled in T2DM trials in the trial summaries on ClinicalTrials.gov. Table 3 reports the number of trials that provide the results of each baseline characteristic. We used formula (2) and (3) to calculate the weighted mean and standard deviation values for quantitative variables (i.e., age, BMI, and HbA1c) separately.

Table 3. Number of T2DM Trials that provided the results for the baseline measures

<table>
<thead>
<tr>
<th>Reported Results</th>
<th># of US-Based T2DM Trials (# of patients)</th>
<th># of UK-Based T2DM Trials (# of patients)</th>
<th>Two-Tailed P Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Any baseline measures</td>
<td>428 (193,345)</td>
<td>86 (90,026)</td>
<td>--</td>
</tr>
<tr>
<td>Mean and standard deviation of age</td>
<td>388 (185,560)</td>
<td>79 (87,441)</td>
<td>P &lt; 0.0001</td>
</tr>
<tr>
<td>Mean and standard deviation of HbA1c</td>
<td>128 (65,445)</td>
<td>38 (32,314)</td>
<td>P &lt; 0.0001</td>
</tr>
<tr>
<td>Mean and standard deviation of BMI</td>
<td>104 (68,515)</td>
<td>30 (36,768)</td>
<td>P &lt; 0.0001</td>
</tr>
<tr>
<td>Race (white, black, Asian, other)</td>
<td>159 (96,518)</td>
<td>34 (53,016)</td>
<td>P &lt; 0.0001</td>
</tr>
<tr>
<td>Gender (male, female)</td>
<td>426 (192,721)</td>
<td>86 (89,791)</td>
<td>P &lt; 0.0001</td>
</tr>
</tbody>
</table>

Table 4 compares the characteristics of the patients enrolled in T2DM trials and the T2DM patients in the OneFlorida Data Trust and CALIBER. The differences of mean values of age between the patients in OneFlorida and the patients enrolled in US-based T2DM trials of Phase I, II, II and are 21.9, 13.0, and 11.0, respectively, showing an increasing a posteriori generalizability of US-based trials regarding age. The patients enrolled in both US and UK-based T2DM trials were younger than the T2DM patients in the clinical data warehouses (two-tailed P < 0.0001). Regarding race, Caucasian/White and Asian were overrepresented, whereas women, black, and other races were underrepresented. The differences between the target populations and the enrolled patients in US and UK-based trials are both statistically significant with respect to race (two-tailed P < 0.0001). Regarding gender, female patients were underrepresented in both US-and UK-based trials (two-tailed p < 0.0001). The patients enrolled in both US and UK-based trials have higher HbA1c values than the target populations (two-tailed p < 0.0001). The patients enrolled in US-based trials have a slightly lower BMI (two-tailed p < 0.0001), whereas the patients enrolled in UK-based trials have a slightly higher BMI (two-tailed p < 0.0001).

Table 4. Characteristics of patients enrolled in T2DM trials and T2DM patients in OneFlorida Data Trust and CALIBER

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>United States</th>
<th>United Kingdom</th>
<th>United States</th>
<th>United Kingdom</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total # of T2DM patients</td>
<td>148,970</td>
<td>--</td>
<td>--</td>
<td>150,665</td>
</tr>
<tr>
<td>Age (mean ± SD)</td>
<td>68.5 ± 14.2</td>
<td>58.1 ± 9.6</td>
<td>-10.4</td>
<td>64.9 ± 13.88</td>
</tr>
<tr>
<td>Race</td>
<td>148,970</td>
<td>96,518</td>
<td>--</td>
<td>57,192</td>
</tr>
<tr>
<td>Caucasian / White (%)</td>
<td>88,213 (59.2%)</td>
<td>72,647 (75.3%)</td>
<td>16.1%</td>
<td>28,629 (50.1%)</td>
</tr>
<tr>
<td>Black (%)</td>
<td>40,186 (27.0%)</td>
<td>6,604 (6.8%)</td>
<td>-20.2%</td>
<td>2,757 (9.6%)</td>
</tr>
<tr>
<td>Asian (%)</td>
<td>2,594 (1.7%)</td>
<td>10,323 (10.7%)</td>
<td>9.0%</td>
<td>5,018 (8.8%)</td>
</tr>
<tr>
<td>Other (%)</td>
<td>17,977 (12.1%)</td>
<td>6,825 (7.1%)</td>
<td>-5.0%</td>
<td>20,788 (36.3%)</td>
</tr>
<tr>
<td>Gender</td>
<td>Male (%)</td>
<td>Female (%)</td>
<td>Male (%)</td>
<td>Female (%)</td>
</tr>
<tr>
<td>--------</td>
<td>----------</td>
<td>------------</td>
<td>----------</td>
<td>------------</td>
</tr>
<tr>
<td></td>
<td>148,969</td>
<td>192,721</td>
<td>150,658</td>
<td>89,791</td>
</tr>
<tr>
<td>Male (%)</td>
<td>69,221 (46.5%)</td>
<td>106,821 (55.4%)</td>
<td>8.9%</td>
<td>81,312 (54.0%)</td>
</tr>
<tr>
<td>Female (%)</td>
<td>79,748 (53.5%)</td>
<td>85,900 (44.6%)</td>
<td>-8.9%</td>
<td>69,346 (46.0%)</td>
</tr>
</tbody>
</table>

BMI, kg/m² (mean ± SD)

<table>
<thead>
<tr>
<th></th>
<th>Male (%)</th>
<th>Female (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>32.3 ± 8.2</td>
<td>31.7 ± 5.5</td>
</tr>
<tr>
<td>Male (%)</td>
<td>29.3 ± 6.0</td>
<td>29.3 ± 6.0</td>
</tr>
<tr>
<td>Female (%)</td>
<td>30.9 ± 5.6</td>
<td>30.9 ± 5.6</td>
</tr>
<tr>
<td>HbA1c, % (mean ± SD)</td>
<td>7.5 ± 2.9</td>
<td>8.2 ± 1.0</td>
</tr>
</tbody>
</table>

* The difference between the patients enrolled in the trials and the patients in the target population.

Population Representativeness of T2DM Trials

Figure 2 visualizes the target population, study population, and the enrolled patients in the US and UK-based T2DM trials for each of the three major quantitative eligibility criteria: age, BMI, and HbA1c. The green curves represent the study populations of the eligible patients, i.e., the percentage of trials that allow a certain value of the variable. The red curves represent the distribution of patients enrolled in the T2DM trials over the value spectrum of the variable. The blue curves represent the target population. In general, the visualization of these populations for US and UK trials exhibited similar characteristics. For example, there is a noticeable gap between the patients enrolled in the T2DM trials and the target population with respect to all the three quantitative criteria. The patients enrolled in T2DM trials were younger and had lower HbA1c and BMI values than those in the target populations (red curves vs. blue curves). The T2DM trials usually permit patients of a wide range of age values. The gap between the study population and the target population is obvious for HbA1c (green curves vs. blue curves). The trends of the study populations and enrolled patients are similar (green curves vs. red curves).

Figure 2. Visualization of (a) age, (b) BMI, and (c) HbA1c in the target populations and study populations of US and UK-based T2DM trials, respectively.

We used the GIST metric to assess the population representativeness of US and UK-based T2DM trials with respect to the three major quantitative eligibility criteria: age, BMI, and HbA1c (Table 5). The GIST scores for the US and UK-based trials were calculated using the target populations derived from the OneFlorida Data Trust data and CALIBER, respectively. We stratified the analysis by study phase. As shown in Table 5, US and UK trials have
similar overall GIST scores for all three variables. The GIST scores of age in both the US and UK trials increased from Phase I to Phase III, which is consistent with the results from two previous studies using the T2DM patients data in the Columbia University [12] and a national survey [14] as the target populations. It is also consistent with our finding of the \textit{a posteriori} generalizability assessment conducted in this work. Phase I studies in the UK had lower GIST scores of age than Phase I studies in the US. The GIST scores of 0.27-0.36 indicate a serious population representativeness issue of UK T2DM Phase I trials. However, Phase II trials in the UK had higher GIST scores of age than Phase II trials in the US. With respect to BMI, the UK trials had slightly higher GIST scores than the US trials of all phases. With respect to HbA1c, the GIST scores of the US trials decreased from Phase I to Phase III, which is also consistent with the two previously mentioned studies [12, 14]. The HbA1c’s GIST scores of the UK trials decreased from Phase I to Phase II, while the UK trials had similar GIST scores of HbA1c as US trials. The fact that \textit{a priori} generalizability of HbA1c is the lowest among the three criteria is consistent with the visualization shown in Figure 2, as well as the \textit{a posteriori} generalizability assessment.

Table 5. GIST scores of age, BMI, and HbA1c of T2DM trials in different phases.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Reading</th>
<th>US-Based T2DM Trials</th>
<th>UK-Based T2DM Trials</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>All</td>
<td>Phase I</td>
<td>Phase II</td>
</tr>
<tr>
<td>N</td>
<td>1,671</td>
<td>299</td>
<td>349</td>
</tr>
<tr>
<td>Age</td>
<td>Earliest</td>
<td>0.74</td>
<td>0.55</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>0.74</td>
<td>0.55</td>
</tr>
<tr>
<td></td>
<td>Median</td>
<td>0.74</td>
<td>0.54</td>
</tr>
<tr>
<td></td>
<td>Middle</td>
<td>0.73</td>
<td>0.54</td>
</tr>
<tr>
<td></td>
<td>Latest</td>
<td>0.73</td>
<td>0.53</td>
</tr>
<tr>
<td>BMI</td>
<td>Earliest</td>
<td>0.87</td>
<td>0.78</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>0.87</td>
<td>0.79</td>
</tr>
<tr>
<td></td>
<td>Median</td>
<td>0.87</td>
<td>0.79</td>
</tr>
<tr>
<td></td>
<td>Middle</td>
<td>0.87</td>
<td>0.79</td>
</tr>
<tr>
<td></td>
<td>Latest</td>
<td>0.87</td>
<td>0.79</td>
</tr>
<tr>
<td>HbA1c</td>
<td>Earliest</td>
<td>0.73</td>
<td>0.83</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>0.74</td>
<td>0.84</td>
</tr>
<tr>
<td></td>
<td>Median</td>
<td>0.73</td>
<td>0.84</td>
</tr>
<tr>
<td></td>
<td>Middle</td>
<td>0.73</td>
<td>0.84</td>
</tr>
<tr>
<td></td>
<td>Latest</td>
<td>0.73</td>
<td>0.84</td>
</tr>
</tbody>
</table>

Discussion and Conclusions

In this study, we used real-world patient data in the target population to assess the generalizability of T2DM clinical trials in the US and UK. As shown in Table 5, US and UK-based T2DM trials have similar \textit{a priori} generalizability of age, HbA1c, and BMI. However, the GIST scores for age in trials of different phases differ between US and UK-based trials. While GIST provides a quantitative metric for comparing the population representativeness of different sets of trials, visualization of different populations can reveal the systematically omitted or overly included population subgroups. The results of \textit{a priori} generalizability showed that the US and UK-based trials exhibit similar issues with respect to the three most frequently used quantitative criteria. The results of the \textit{a posteriori} generalizability assessment showed that males, whites, and Asians are overrepresented in both US and UK-based T2DM trials while females, blacks, and other races are underrepresented.

Compared to the \textit{a posteriori} generalizability assessment, the use of the GIST metric to assess the \textit{a priori} generalizability assessment has a few advantages. First, it can be performed during the trial design phase, which would help reveal issues of eligibility criteria that are biased towards certain population subgroups, and help trial designers optimize the balance between internal and external validity. Trial designers can fine-tune the criteria without diminishing the internal validity. For example, UK T2DM Phase I trials have a very low GIST score of age. Trial designers should thus adjust the restrictive age criterion in Phase I trials in the future to improve their population representativeness. Second, GIST quantifies the difference of the distributions of the eligible patients and the target population over a variable, whereas the \textit{a posteriori} generalizability compares the mean difference of a variable. It can be done using cost-effective informatics tools. Meanwhile, the \textit{a priori} generalizability assessment has a few disadvantages. First, it does not take into account the practical issues in the trial recruitment phase, such as
geographic locations, accessibility of trial information, and consideration of comorbidities. For example, due to real-world complications, most studies failed to recruit representative samples of their study population as specified in trial eligibility criteria [10]. On the other hand, the \textit{a posteriori} generalizability assessment, which compares the enrolled patients with the target population, can provide a more accurate assessment of the population representativeness. The issues of gender disparity and race disparity, as well as other disease-specific outcome measures can be accurately detected. Both \textit{a priori} and \textit{a posteriori} generalizability results showed increasing generalizability of US-based T2DM trials from Phase I to Phase II with respect to age, which confirmed our hypothesis that \textit{a priori} generalizability is comparable to \textit{a posteriori} generalizability in identifying certain restrictive quantitative eligibility criteria. It should be a common practice to assess both \textit{a priori} generalizability based on trial design factors such as eligibility criteria before patient recruitment as well as \textit{a posteriori} generalizability \textit{post hoc} based on enrolled patients. The trial design issues that are found in \textit{a priori} generalizability can be addressed before patient recruitment, thereby improving the \textit{a posteriori} generalizability and the cost-benefit ratio of the trials. Nevertheless, clinical trial investigators should also consider practical issues in the trial recruitment phase. For example, most trials still use a traditional hospital-based recruitment strategy. Thus, the trial designers of these studies should carefully choose recruitment sites, and take into account the population characteristics of these sites’ catchment areas.

A number of limitations should be noted in our study. First, less than 40% of T2DM trials reported results in ClinicalTrials.gov. Therefore, the aggregate results of the enrolled patients represent merely a convenient sample. Second, the OneFlorida Data Trust contains data of patients who have visited healthcare organizations in the state of Florida, and thus might not be representative. Even though we only used patient data from the state of Florida in the US, Florida is the third most populous state (~19.9 million) in the US. Third, it is possible that not all clinical trials conducted in the UK are registered on ClinicalTrials.gov. We identified trials based on the study site. Some trials are conducted in multiple countries. Therefore, some UK-based trials may also have study sites in other countries.

\textbf{Acknowledgments}

We show our gratitude to Sarah Mixon at Florida State University Department of Chemistry for her hard work to extract results from Type 2 diabetes trials in the United Kingdom. The development of the COMPACT database and the VITTA tool was supported by U.S. National Library of Medicine Grant R01LM009886 (PI: Weng) and the National Center for Advancing Translational Science (NCATS) Award UL1TR000040 (PI: Ginsberg). This work was partially supported by an Amazon Web Service in Education Research Grant Award (PI: He) and the Planning Grant of the Institute of Successful Longevity at Florida State University. The work was also partially supported by National Center for Advancing Translational Sciences under the Clinical and Translational Science Award UL1TR001427 (PI: Nelson & Shenkman). The content is solely the responsibility of the authors and does not represent the official view of the National Institutes of Health.

\textbf{References}

Bar charts detection and analysis in biomedical literature of PubMed Central

Ying He, BD¹, Xiaohan Yu, MD¹, Yangjing Gan, BD¹, Tujin Zhu, Shengwu Xiong, PhD¹, Jing Peng, PhD¹, Lun Hu PhD¹, Guang Xu, PhD² Xiaohui Yuan, PhD¹
¹School of Computer Science and Technology, Wuhan University of Technology, Wuhan, China; ² Hubei Co-Innovation Center of Basic Education Information Technology Services, College of Computer, Hubei University of Education, Wuhan, China

Abstract

Bar charts are crucial to summarize and present multi-faceted data sets in biomedical publications. Quantitative information carried by bar charts is of great interest to scientists and practitioners, which make it valuable to parse bar charts. This fact together with the abundance of bar chart images and their shared common patterns gives us a good candidates for automated image mining and parsing. We demonstrate a workflow to analyze bar charts and give a few feasible solutions to apply it. We are able to detect bar segments and panels with a promising performance in terms of both accuracy and recall, and we also perform extensive experiments to identify the entities of bar charts in the images of biomedical literature collected from PubMed Central. While we cannot provide a complete instance of the application using our method, we present evidence that this kind of image mining is feasible.

Introduction

Bar charts are necessary resources for scientists and practitioners to describe experimental results such as comparisons among categories or grouped data in published biomedical literature. Recently, there has an enormous increase in the amount of open-access heterogeneous biomedical image production and publication, and a trend in the area of literature mining is allowing users to query the figures of biomedical articles which are otherwise not readily accessible.¹, ², ³ These researches focus mostly on image retrieval, image classification and making text within image available. However, biomedical images contain much structured information and quantitative information, which is not yet accessible through search. Below, we present our approach to detect and access bar charts in biomedical publication. For the purpose of further study, we also illustrate some possible situations which take advantage of our bar charts mining method.

Image types are of great importance for image mining. Kuhn et al. manually annotate the segment sub-graphs in the constructed corpus of 3000 images and classify them into five basic categories: Experimental/Microscopy, Graph, Diagram, Clinical, and Picture. The results show that bar chart is the most common types of subfigure accounting for 12.4% of entire set of images,⁴ which reinforces our determination to study bar charts. Furthermore, the bar charts share the common patterns that have at least two straight and visible axes that one represents specific categories being compared, and the other represents a discrete value, and uses rectangular bars with lengths proportional to the values that they represent.

A closer look at bar chart reveals that they follow regular patterns to encode their semantic relations. Since the priority target of our approach is to automatically extract the relation of the corresponding quantitative proportion to the categorical data, we focus on the text information (the axes labels and the legends) as well as the length proportion of

Figure 1. Four typical examples of bar charts.
the bars. Figure 1 shows four typical bar chart images. The slight differences about the text and the bars should be noticed. Panel (a) shows the most common bar chart, which has a horizontal x-label, y-label and legends. Panel (b) shows a figure with a horizontal grid and a slanted x-label. Panel (c) shows bars on both sides of the x-axis, and the x-label is vertical. Panel (d) shows the condition in which the bars are filled with slash. These kinds of bar charts represent the most typical bar chart images.

Figure 2 shows two common bar charts together with a table representation of the extracted information. The first one shows the expression patterns of the GmIFR genes in various tissues. The x-label represent the category of the tissues while the y-label represent the expression level of the gene. A bar chart can be considered a kind of matrix with pictures of experimental artifacts as content. The tables to the right illustrate the semantic relations encoded in the bar charts. Each relation instance consists of a condition, a measurement and a result. Gene is the entity being measured under the conditions of the different tissues. The result is a certain level of expression indicated by the lengths proportion to the bars. Second example is a slightly more complex one. The mRNA transcript level of the GmPAD4, GmEDS1 and GmPR1 gene on AtPAD4-overexpressing and control roots. More than one gene are tested against each other in a way that involves more than two dimensions. In this case, legend is the major technique to denote the different possible combinations of a number of conditions.

Background and related work

Image classification and retrieval are two of the most active areas in biological image mining research. Rafkind et al.\textsuperscript{5} proposed a classification system using coherence and frequency features to divide figures into five sets. Then, a retrieval method,\textsuperscript{6} which had better flexibility than the system of Rafkind et al.\textsuperscript{3}, was proposed to retrieve figure types defined conceptually by taking advantage of principles in image understanding, text mining, and optical character recognition (OCR). For searching information in text embedded in figures, YIF (Yale Image Finder)\textsuperscript{7} has been proposed to retrieve biological figures and associated papers in which the retrieved images allow users to find related papers by linking to their source papers. Hearst et al.\textsuperscript{8} developed the BioText Search Engine, which is a freely available web-based application for searching and browsing figures in articles as well as their captions. The study of image processing also includes image segmentation, optical character recognition (OCR) and interpretation of figure captions. Li et al.\textsuperscript{9} proposed an algorithm to segment images that consisted of multiple subfigure into single images and then

![Figure 2](image)

**Figure 2.** Two examples of bar charts from biomedical publications (PMC4655237 and PMC3648381) with tables showing the relations that could be extracted from them.
extract the title and tag information from the graph. Lopez et al. developed a robust image segmentation algorithm in order to perform text retrieval based on images. Kim et al. developed an image and text extraction tool (figtext) through the combination of image preprocessing, character recognition, and text correction to improve the performance of OCR tools.

Research involving biological images is currently subject to some specific limitations. Only a small amount of research has been conducted into the mining of information from image data in biological research or into the construction of a structured biological knowledge base. SLIF (Structured Literature Image Finder) has been developed for automated information extraction from fluorescence microscopy image information, and a series of follow-up studies have been reported for promotion. Their project included image classification, figure title understanding, figure segmentation, and the relationships between sub-images and sub-captions. Also, there is a large amount of existing work on how to process bar charts. Zhou et al. proposed a modified probabilistic Hough transform algorithm to detect and recognize bar charts, which is, however, just the first step in locating them and analyzing their labels and their structure. Al-Zaidy et al. attempts to extract information from bar charts automatically. However, the method takes plain bar chart as input, which are not readily accessible from biomedical papers, because they make up just parts of the figures. Furthermore, the method mentioned above is designed for researchers who want to analyze their bar charts and not to read bar charts that have already been analyzed and annotated by a researcher. Therefore, these approaches do not tackle the problem of recognizing and analyzing bar charts. Some attempt to classify and redesign biomedical images include bar chart, but they are only tested on small data sets that do not satisfy the diversity of images in biological documents. To the best of our knowledge, there is little research into the mining of bar charts, which is a frequently used means of demonstrating the differences among data.

Method

Our approach to image mining from bar charts consists of 6 components: figure extraction, image preprocessing, bar segment detection, in-image text recognition, panel segmentation, and quantitative information extraction (Figure 3). Figure extraction module that extract images accessed in PubMed. Image preprocessing module that removes non-informative figures to reduce computing cost. For bar segment detection, we use a detection procedure based on hand-coded rules and convolutional neural network (CNN) method to detect bar segments. In-Image Text Processing module identifies the corresponding text region in the image by performing a text location method. Additionally, the Panel Segmentation module combines the results obtained from the previous two modules to first estimate. Finally, we extract the quantitative information among the entities.

Figure 3. The procedure of our bar charts mining method. (from PMC4482714)
**Figure extraction.** The data interface (OA web service) are manually accessed in PubMed, which makes it convenient to download article folds consist of documents with different formats (PDF/NXML format articles, GIF/JPEG format images, etc.). We obtain a subset for our future use case by searching the keyword combination of “soybean,” “gene” and “expression.” We only use this subset so far, which makes it easy for us to extract figure. Although it would be definitely feasible, it would also be more trivial for automatic extraction of figures from biomedical articles in PDF format.

**Image preprocessing.** The download figures from PubMed Central typically incorporate a significant number of non-informative figures (e.g., conference and journal logos, formula), which lead to additional computational processing cost. We separate the objects from the background image by using pixel level features. When converting an image into binary image, we give a fixed threshold value of 0.9 instead of automatic threshold value because we want to get a more complete bar segment. Three factors are considered here to filter out figures that are definitely not bar charts: (1) The number of connected domains is less than 5; (2) the absolute width or height is more than 100 pixels, and (3) the 0 value rate is more than 95% of binary image.

**Bar Segment Detection.** We first extract connected components from the filtered image by grouping adjacent pixels of similar color, (for binary image, the adjacent pixels of 1 belong to the same component), and then represent the objects in the figure as a bounding box. The small components (e.g., characters) are filtered according to the area ratio of the connected domains firstly.

As a baseline, we first propose a relatively simple hand-coded bar segment detection method. Such bar segments typically have several bars with the same width distributed uniformly on the x-axis, which is the most distinct feature of the bar chart. For this reason, a projection method is used to detect such bar segments. All the columns are subsequently vertically projected to the x-axis when the area value of the connected domain reached the threshold. The bar charts are detected through defining rules for the ratio of columns that are completely blank and the ratio of the sum of equal columns.

There is an alternative method on using machine learning approaches for image classification. In this paper, we choose a simple CNN model to be the training model for bar charts detecting. With the analysis of the intermediate results (bar segments) generated by the hand-coded method, we can easy get the dataset of CNN. The training set consists of 12000 positive samples and 4000 negative samples, while the test set consists of 3000 positive samples and 1500 negative samples. Experiment of different parameters is conducted and the best model is chosen with the learning rate 1 and iteration times 800.

**In-image text recognition.** Because the text in figures is of great significance, a text extraction method is used to extract the important text in the figure. Not only should we extract the text from the image, but also the position of the text. Text localization detects different text regions in images. In this part, we focus on x-label, y-label, and legends. All above information are obtained respectively for the necessity of constructing complete corresponding relation. To do this, we use a detection procedure with simple rules. The coordinate axes are used to partition the region. Then the direction of the character is distinguished through combination of the character spacing and character size, and the character region, which is vertical and slanted, is rotated. For optical character recognition (OCR), the open-source tool OCROPY (https://github.com/MissCristal/ocropy) is used. Sometimes the deletions, insertions, and substitutions of letter or number tokens appeared in the extracted text information from bar chart can be found from the related figure caption text. Spell correction of the extracted in-image text is realized by computing Levenshtein Distance between the extracted in-image text and the corresponding figure caption text.

**Panel Segmentation.** The last step consists of segmenting the figure into panels using the information extracted from the previous two modules. The sub bar charts detected in the bar segment detection step are found in most case that it is a part of a whole figure in which some parts are not bar charts. Although in our study we just extract information from the image, the information from the article text can be potential additional information to the image. Hence, the order of the sub charts is an important information for further study.

A rectangle segmentation algorithm is used to extract the potential sub bar graphs contained within the figures. After bar segment detection and in-image text recognition processes, the detected bar segments are extended to produce a larger rectangle containing important information such as x-label, y-label, and legends of a sub bar chart. If the area cover the whole figure, the segmentation process is not able to continue because it have only one sub chart. Otherwise, the multiple sub bar charts are generated by reducing the enlarged rectangle based on the in-image text detected. The sub charts that are not sub bar graph were filtered out. On the basis of our observation, sub charts are usually arranged from left to right and from top to bottom. Therefore, we name the sub charts according to the panel position.
Quantitative Information Extraction. The quantitative information in bar charts is obtained through recognition of the x-axis and the height of each bar. Twenty uniformly distributed rows per graph are traversed. If the row crossed the bar, it is traversed up and down to find the critical point and the row index is recorded. The x-axis is here defined as the row at which the row index is recorded the greatest number of times. If the x-axis is in the bottom of the figure, then all of the bars are on top of the x-axis. All the graduated lines including the x-axis are removed to separate all the bars. Then the bars are filled in to make them all solid. Then all of the columns are traversed up to down. The first row index for which the pixel is 0; otherwise, 1 was set as the value, and it is denoted as vector U. Then vector U is traversed the height of a bar is defined if there are no fewer than 5 consecutive identical values. If the x-axis is in the center of the figure, and the bars are located on both sides of the x-axis. All the columns are traversed from both sides, and steps similar to those outlined above are used.

Results
To test our approach, we created a gold standard corpus of images. We randomly selected a sample of 300 open-access articles folds from PubMed Central that consist of jpg image. Altogether, these 300 articles contain 1769 figures, an average of about 6 figures per article. For these figures, we manually annotated the number of image containing bar charts and the number of bar charts, resulting in 534 (32.2%) images containing at least one bar chart and in total 1659 sub bar chart. The analysis demonstrated that bar charts are indeed a very important image type in the biomedical literature. Among the figures contain at least one bar chart, 87 (16.4%) thereof contain more than 5 bar charts. This corpus was selected and annotated after we finished implementing the algorithm presented here. While we develop the method, another corpus is downloaded from PubMed Central. This two corpus have no overlap.

As shown in the top part of Table 1, the hand-coded algorithm is able to detect the bar charts with a high precision of 95.47% and a recall 59.08%. The CNN classifiers based on image features alone have lower precision than the hand-coded algorithm, but higher recall. Overall, the CNN classifier combined with hand-coded algorithm performs better than the hand-coded classifier and CNN classifier alone with F-measures of 88.66%. To clarify the cause of the result, we analyze the result of the combined method. Most of the false positive cases are axis diagram (e.g., line chart), and most of the true negative cases are bar chart with no distinct bars (e.g., very low bars).

Table 1. Evaluation of our approach.

<table>
<thead>
<tr>
<th>Task</th>
<th>Method</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bar segment detection</td>
<td>Hand-coded</td>
<td>0.9547</td>
<td>0.5908</td>
<td>0.7299</td>
</tr>
<tr>
<td></td>
<td>CNN</td>
<td>0.8945</td>
<td>0.8545</td>
<td>0.8691</td>
</tr>
<tr>
<td>Panel segmentation</td>
<td>Hand-coded</td>
<td>0.9770</td>
<td>0.8115</td>
<td>0.8866</td>
</tr>
<tr>
<td></td>
<td>Hand-coded + CNN</td>
<td>0.9811</td>
<td>0.5273</td>
<td>0.6859</td>
</tr>
</tbody>
</table>

The results of the bar panel segmentation algorithm are shown in the middle part of Table 1. We define the complete panel segmentation as to segment the panel with the x-label, y-label and the legends (if the bar chart has legends). Our method correctly produce 98.11% of the panel segmentation at a recall of 52.73%, which leads to an F-measure of 68.59%. In order to determine the different reasons of the low recall, we analyze the figure which is detected as the bar segments but not obtained the complete bar chart panel. (10.31%) Figures identified by our system share at least one data (e.g., x-label), which leads to incomplete panel. Other factors such as the sub charts are very close in a big image, the legends of the bar chart are not below the x-axis or is described in the caption could also make the panel incomplete.

Results of the extracted information are shown in the bottom part of Table 1. If information extracted from the bar charts can fill all the columns of the corresponding table, it is defined as a correct extraction. Almost 39.81% of the bar panels are extracted. 82.40% thereof are correct. Low quality of images is the most important reason of incorrect information extraction. The incorrect cases could be split into two classes of roughly the same size: the wrong number of categories in the group data, the wrong number of bars. The primary cause for the first class of incorrect cases was the fact that different forms of x-label (such as slanted x-label) make it hard to segment the labels. The causes for the second class are the facts that some images contain large numbers of very thin bars that are difficult to detect and the interferences such as the graduated lines, the slash bars of some bar chart lead to more bar numbers than exactly being contained. Usually, the number of detected bars was greater than the number of real bars, owing to the margin of error occurring in lines.
Table 2. The results of running the pipeline on the open access subset of PubMed Central.

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Total articles</td>
<td>14596</td>
</tr>
<tr>
<td>Processed articles</td>
<td>11973</td>
</tr>
<tr>
<td>Total figures from processed articles</td>
<td>80378</td>
</tr>
<tr>
<td>Processed figures</td>
<td>61238</td>
</tr>
<tr>
<td>Detected bar charts</td>
<td>44537</td>
</tr>
</tbody>
</table>

Table 2 shows the results of running the pipeline on the subset of PubMed Central. We start with about 14596 articles folds, which is accessed from the data interface (OA web service) in PubMed and searched by the keyword combination of “soybean,” “gene” and “expression.” About 18% article folds are discarded by the reason of containing no article with XML format or no JPG image. Remaining articles contain around 80378 figures. In order to reduce additional computational cost, no-information figures are filtered in the image preprocessing steps. We ended up with more than 61238 figures, in which about 44537 bar charts are detected.

Discussion

In this paper, we have developed a comprehensive system for automatically extracting information from bar charts. Our result confirm that our bar chart detection method can achieve a high accuracy, allowing us to segment the bar panel with a high precision. While the bar charts are so frequent, the low recall of the panel segmentation is not a severe problem at this point. To extract the relation of the categorical data in perspective, the conditions (x-labels and legends) should match to the bars exactly. Although the recall is low, about 40% is still in a reasonable range.

Since we invested effort in extracting information from bar charts, we give our opinion on the use of such information. Our method can help to construct dataset provided direct reference and evidence for the researchers. For example, the statistical analysis of PCR and phenotypic data are often expressed by means of bar chart. PCR is the gold standard, used to validate the results of RNA-seq.20 PCR data are more accurate and repeatable than high-throughput RNA-seq data. However, PCR data are always reported in academic papers, which have seldom been collected from the specialized databases. A combination of text mining method and our bar chart mining method is able to construct the database not only containing information from articles, but also the quantitative information of the experimental results. Illustration of phenotypic data also account for a large proportion in the bar charts, which is currently a bottleneck of genome-wide association study (GWAS) and molecular breeding. Currently, most public genotype-phenotype databases did not include the corresponding data regarding soybean organisms.21 For this reason, one of the practical values of the current dataset is that it contain correlation information for genotype and phenotype, which could be extracted from bar charts in publications.

It seems reasonable to assume that these results can provide necessary information to biomedical application. We plan to investigate this in future work. The results obtained from bar chart processing pipeline indicate that it is feasible to extract relations from bar charts, but it is clear that this procedure is far from perfect. Aiming to the problem of low image resolution, the automatic analysis of vector diagram seems to be an efficient way to extract such relations from existing publications in the future.

Conclusion

In this article, we present the most common image — bar charts — to be automatically assessed in a reliable way. Our results show that the hand-coded algorithm and the CNN method we proposed can detect the bar segment at a high accuracy. We also depict that relation and quantitative information can be extracted from the bar charts with satisfactory precision. In order to demonstrate and exemplify the algorithm’s ability and advantage, the literatures related to soybean gene expression are selected and tested. Based on these results, we believe that our proposed bar chart mining method is a viable and promising approach to provide more power to gather relations such as gene regulation. All in all, we have demonstrated a novel methodology that detects bar charts for our use case. Further research will be to apply the methodology to a larger set of citations with more diverse content.

References

11. Kim D, Yu H. Figure text extraction in biomedical literature. PloS one. 2011;6(1):e15338.
Evaluating the Effectiveness of Auditing Rules for Electronic Health Record Systems

Monica Hedda, BE1, Bradley A. Malin, PhD1, Chao Yan, MS1, Daniel Fabbri, PhD1

1Vanderbilt University, Nashville, TN

Abstract

Healthcare organizations (HCOs) often deploy rule-based auditing systems to detect insider threats to sensitive patient health information in electronic health record (EHR) systems. These rule-based systems define behavior deemed to be high-risk a priori (e.g., family member, co-worker access). While such rules seem logical, there has been little scientific investigation into the effectiveness of these auditing rules in identifying inappropriate behavior. Thus, in this paper, we introduce an approach to evaluate the effectiveness of individual high-risk rules and rank them according to their potential risk. We investigate the rate of high-risk access patterns and minimum rate of high-risk accesses that can be explained with appropriate clinical reasons in a large EHR system. An analysis of 8M accesses from one-week of data shows that specific high-risk flags occur more frequently than theoretically expected and the rate at which accesses can be explained away with five simple reasons is 16 - 43%.

Introduction

Electronic health record (EHR) systems can improve the quality of patient care, safety and education, while reducing costs and enabling research. To encourage the adoption and use of EHR systems by healthcare providers, the US government passed the Health Information Technology for Economic and Clinical Health (HITECH) Act of 2009 and established incentives for healthcare providers that demonstrate meaningful use of EHR system to provide better patient care. As a result, integration of EHR systems into healthcare organizations (HCOs) has continually increased. However, the increased accessibility of protected health information (PHI) in EHR systems leads to a greater potential for misuse and abuse by the authorized users. Such events can result in penalties levied by federal and state regulators.

An EHR is fundamentally a collaborative information system, which, traditionally, is protected through proactive strategies, such as fine-grained access control technologies. Such technology is often integrated into EHR systems; however, the dynamics of patient care, in combination with the difficulty in predicting who needs access to a patient's medical record when, make it challenging to deploy such fine-grained control schemas without triggering a substantial quantity of false alerts and slowing care workflows. Despite acknowledging the potential for insider threats, HCOs typically do not instantiate fine-grained controls. This implicitly suggests that HCOs deem the losses associated with impacts on workflow and care to be greater than those brought about by employees who misuse or abuse their privileges.

Still, HCOs do not neglect insider threats entirely. In lieu of fine-grained proactive protections, HCOs tend to rely upon retrospective mechanisms, such as auditing and investigation. In the United States, the Security Rule of the Health Information Portability and Accountability of 1996 (HIPAA) requires that all HCOs maintain audit logs, analyze them for inappropriate use and report misuse. Hospitals maintain audit log of all accesses to PHI and the audit log is often reviewed by administrative officers to detect inappropriate access. However, the sheer volume of accesses documented by large HCOs makes manual review infeasible. The number of access transactions is often over one million per day, while officers have only one or two people at their disposal (often allocating only a portion of their time) to run investigations. As a result, many HCOs prioritize their investigations by monitoring patient records deemed to be very important persons (VIPs) or upon patient complaints. In the latter scenario, compliance officers investigate the accesses to patient records after a complaint has been registered.

More recently, there has been a push to (semi-)automate the auditing process. However, there are many challenges an HCO faces to do so. For instance, the information often required to determine if an access is inappropriate is not stored in the audit log. As a consequence, HCOs have deployed rule-based methods to capture high-risk behavior and promote them to compliance officers for review. Figure 1 shows the process associated with such a traditional rules-based auditing system. Unfortunately, rule-based flagging systems can result in high false positives. For example, a typical rule is to flag when an employee accesses an EHR of a patient with the same last name. Yet, for individuals with a common name, clearly this rule will trigger an excessive amount of alerts.
Given the state of affairs, we set out to assess the validity of rules for auditing accesses made in EHR systems. The goal of this study is to test, through simulation and theoretical analyses, if these flags occur at a higher rate than expected, and therefore serve as a valid means to detect inappropriate behavior. Our contributions in this work are as follows:

1. **Investigate the difference between observed and expected high-risk accesses.** We introduce an approach to investigate the difference between the observed and expected rate of high-risk accesses in EHR systems for typical expert-specified rules. If a rule holds merit, we anticipate that the observed rate of high-risk accesses will be higher than the expected rate of high-risk accesses to the EHR system. Using one week of data from Vanderbilt University Medical Center (VUMC), we show that there are many rules for which this difference is statistically significant.

2. **Select and prioritize rules based on deviation between observed and expected.** We introduce an approach for selection and prioritization of the high-risk rules. This approach is based on the magnitude of the deviation between the observed and expected frequency of high-risk accesses for each rule.

3. **Prioritize flagged high-risk accesses for investigation.** To improve the manageability of a manual review process in resource constrained environments, explanation-based filtering\textsuperscript{10} can be utilized to prioritize the flagged accesses for manual review. Note that rule-based flagging and explanation-based filtering are complementary approaches to detect inappropriate behavior. While rules capture the high-risk behavior, explanations reduce the set of accesses that need to be investigated to a set of un-explained accesses. We find synergy between these two auditing approaches and introduce an explanation-based mechanism to prioritize high-risk accesses flagged by rules for manual investigation. We show that many, though not all, of the high-risk accesses can be explained away with clinically justifiable reasons.

![Figure 1. The process by which an HCO investigates accesses to EHRs deemed to be of high risk.](image)

**Background**

Many HCOs use modern EHR systems equipped with automatic log collection to maintain the audit log. Each time an employee accesses a patient’s record, the automatic log collection process adds a transaction to the audit log with the details of the access including user, patient, action performed (e.g., viewed reports) and time of access.

Various auditing strategies have been proposed to detect inappropriate insider accesses in EHRs. Boxwala et al. introduced an approach to automate the process of detecting suspicious access to EHRs using statistical and supervised machine-learning techniques\textsuperscript{9}. In this technique, audit logs and data from EHRs were applied to construct features to learn predictive models and rank suspicious, as well as non-suspicious, accesses. Training data for such models were provided by HCO privacy officials. Recognizing that not all suspicious accesses are affiliated with a specific pattern, a variety of frameworks have been developed to detect anomalous accesses based on deviations from expected behavior\textsuperscript{14, 15}.

Still, all of these methods assume that all accesses should be considered for predictive purposes, which is potentially problematic because there is substantial class imbalance. Specifically, the rate at which inappropriate accesses actually transpire is likely to be orders of magnitude smaller than the rate at which appropriate accesses are committed. Thus, to mitigate skew in the prior probabilities for such statistical models, as well as manual audits, the notion of an explanation-based auditing system (EBAS) was proposed. EBAS works by filtering out accesses to the
EHR according to explanations generated automatically from the data by a mining algorithm\textsuperscript{16,18}. While all of these auditing strategies offer certain benefits over the simple rule-based auditing system, currently approaches based on the latter are in common use by HCOs.

**Methodology**

We hypothesize that a high-risk audit rule holds merit when the observed frequency at which it fires is higher than what would occur due to routine daily behavior. To test this hypothesis, we compare the observed frequency of high-risk accesses in a large EHR audit log with what one might expect to observe at random. We apply a goodness of fit test to determine if there is a significant difference between the observed and expected frequencies. We further examine the observed high-risk accesses flagged by each rule to determine the minimum rate in that these accesses can be explained with clinical reasons.

**Data overview**

The data investigated in this study is drawn from the VUMC EHR system. Table 1 depicts the data investigated in this study. These data are an integration of EHR audit log, employee personal information, patient personal information (e.g., names, dates of birth, and residential addresses) with information about the department for which the employee is affiliated (e.g., the Anesthesiology department).

<table>
<thead>
<tr>
<th>Total Accesses</th>
<th>Repeat Accesses</th>
<th>Self-Accesses</th>
<th>Unique Non-Self Accesses ((L_{np}))</th>
<th>Unique Employees ((E))</th>
<th>Unique Patients ((P))</th>
<th>Unique Departments</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.5M</td>
<td>6.9M</td>
<td>21K</td>
<td>710K</td>
<td>13K</td>
<td>152K</td>
<td>2.1K</td>
</tr>
</tbody>
</table>

We designate an access as a **Self-access** when the employee has accessed his/her own record. We assume this occurs when the first name, last name and date of birth of the employee and patient in the access are the same. We designate an access as a **Repeat access** when the employee accesses the record of the same patient earlier in the week. All of the accesses except the first access are considered as **Repeat accesses**.

**Types of high-risk behavior**

While there are many types of high-risk behavior, we selected the following types for our experiments through background analysis. Specifically, we investigate five high-risk rules in this study:

1. **HCO Co-Worker**: The EHR user and patient are both employees of the VUMC.
2. **Department Co-Worker**: The EHR user and patient work in the same VUMC department.
3. **Last Name**: The EHR user and patient have the same last name.
4. **Geographic Proximity**: The EHR user lives within 0.25 miles of the patient.
5. **Residential Street**: The EHR user lives on the same street as the patient.

In addition, we added one rule to ascertain if the results of our experiments are merely an artifact of the data or if they are indicative of suspicious behavior:

6. **First Name**: The EHR user and patient have the same first name.

**Method overview**

In this section we provide an overview of the method to test our hypothesis and to determine the minimum rate of high-risk accesses explained with a clinical reason, as depicted in Figure 2.

The steps in this method are defined broadly as follows:

1. Determine the observed frequency of the high-risk accesses.
2. Determine the expected frequency of high-risk accesses by:
   2.1. Using simulations with random samples of users and patients.
   2.2. Using simulations with permutations of users and patients.
2.3. Using a theoretical formulation.

3. Compare the observed frequency of high-risk accesses to the expected frequency of high-risk accesses, and determine the significance of the deviation between the observed and expected frequency.

4. Use the explanation-based method to identify the observed high-risk accesses that can be explained with clinical reasons, and determine the minimum rate that observed high-risk accesses can be explained.

Figure 2. The steps to compute the observed and expected frequencies of high-risk accesses, and the minimum rates that high-risk accesses can be explained.

Figure 3 depicts the method to obtain the observed accesses by sampling the audit log and simulation of the expected accesses using permutation and random sampling methods. We explain these methods in detail in the following sections.
**Observed.** We obtain the observed frequency of high-risk accesses empirically from the set of unique employee-patient pairs \( L_{EP} \), where \( E \) is the list of employees (or users) and \( P \) is the list of patients in the employee-patient access pairs \( L_{EP} \). These pairs are obtained from the 710,000 unique accesses in the audit log and are devoid of any self-access. We select a random sample \( S \) of 100,000 pairs from \( L_{EP} \), each of which is assessed for the high-risk criteria. We count the occurrence of the high-risk accesses across the sample and calculate the frequency of the high-risk accesses in the sample as \(|\text{high-risk accesses}| / |S|\).

**Expected.** To simulate accesses and obtain an expected frequency distribution of high-risk behaviors we apply both permutation and random sampling methods. We use two distinct methods to confirm these simulation methods do not result in selection bias and that the sample selected by our methods are representative of the population. We compare the results of the simulations to verify if the results lead to the same conclusion.

**Expected: Permutation.** In this approach, we construct simulated accesses by shuffling the data points in \( S \). We use two types of permutation methods to simulate accesses and verify that results of both the methods lead to the same conclusion. 1) **Permute Patients:** This method shuffles the list of patients while holding the list of employees in sample \( S \) constant. 2) **Permute Employees:** This method shuffles the list of employees while fixing the list of patients in sample \( S \).

**Expected: Random Sampling.** We obtain the employee list \( E \) and the patient list \( P \) from the set of employee-patient pairs \( L_{EP} \). Next, we select a random sample of 100,000 employees \( S_E \) and 100,000 patients \( S_P \) (without replacement) from \( E \) and \( P \), respectively. We then construct simulated accesses by randomly matching the records in \( S_E \) and \( S_P \).

For each simulation, we calculate the frequency of high-risk accesses in sample \( S \).

**Expected: Theoretical Formulation.** The expected frequency of high-risk accesses is computed empirically, using the probabilities of high-risk accesses occurring among the employees and patients in sample \( S \). We determine the expected frequency for five of the six rules presented above.

The expected frequency of the high-risk accesses using probabilities is computed as \( \sum_{i=1}^{n} P_{E_i}P_{P_{t_i}} \), where, \( P_{E_i} = |\text{Employee with attribute value } x| / |S_E| \), \( P_{P_{t_i}} = |\text{Patient with attribute value } x| / |S_P| \), attributes: [last name, first name, residential street name, work department name] and \( S_E \) and \( S_P \) are the lists of employees and patients in \( S \), respectively.

---

**Figure 3.** An overview of the process for sampling the observed accesses and simulation of the expected accesses.
Experimental Evaluation. We run 10 experiments each for the randomization and permutation methods to compute the observed and expected frequencies of high-risk accesses. We compute the ratio of the mean observed frequency to mean expected frequency for each of the high-risk rules. We also compute the percentage of observed and expected high-risk accesses for each high-risk type to determine the rate of observed and expected high-risk accesses.

Minimum rate of high-risk accesses explained

While there are many operational and clinical reasons that can explain the reason for accesses in an EHR, we select primary treatment, payment and healthcare operations (TPO)\textsuperscript{17} to ascertain the extent to which high-risk accesses can be explained. We specifically focus on explanations in the form of 1) scheduled appointments, 2) ordered lab results, 3) ordered medications, 4) admission, discharge, and transfer events, and 5) clinical documentation. A high-risk access can have multiple explanations (e.g., patient had a scheduled appointment with the accessing employee, and patient also had a lab order with the accessing employee). We use the explanation-based approach to prioritize the observed high-risk accesses for further investigation by administrative officers, with unexplained accesses considered as high priority for the investigation. The explained accesses can be ranked using the type and number of explanations available for the access.

Since we do not exhaust the list of possible reasons, we compute the minimum rate at which high-risk accesses can be explained for each high-risk rule. Additional plausible explanations for the access exist (e.g., user performed surgery on the patient) and could be invoked to raise the rate.

Goodness of fit chi-square test

We apply a \( \chi^2 \) test to determine the goodness of fit between the observed (empirical) and expected (simulated) number of occurrences of high-risk accesses. This test is designed to ascertain if there is a significant difference, such that these deviations are likely not the result of chance alone. The measure of goodness of fit is \( \chi^2 = \sum_{i=1}^{n}(O_i - E_i)^2 / E_i \), where \( O_i \) and \( E_i \) are the observed and expected high-risk event frequencies of type \( i \), respectively.

We test this value against a \( \chi^2 \) distribution with 1 degree of freedom. This is because there are two categories: 1) High-risk accesses, 2) Non-high-risk accesses. We perform this test at the 0.01 significance level (i.e., we accept the alternative hypothesis when the value result is below this level).

Results

In this section, we summarize the deviation of the observed from the expected high-risk access rates obtained by four methods 1) Permute Patients, 2) Permute Employees, 3) Random Sampling and 4) Theoretical formulation, for each high-risk access rule. We begin by presenting the rate of observed and expected high-risk accesses. Next, we summarize the minimum explanation rate for the high-risk accesses. Finally, we report the statistical significance of the deviation between the observed and expected high-risk access rates.

Observed versus expected frequencies of high-risk access

Table 2 summarizes the observed to expected frequency ratios for the various high-risk access rules. It was found that the ratio of observed to expected frequencies varies from 0.99 to 4.33 for the high-risk behavior rules. The observed frequency of the high-risk accesses is higher than the expected frequency for all the high-risk rules except for the HCO Co-Worker rule, which is approximately 1 (at 0.99).

Table 2. Observed to expected frequency ratio for the high-risk access rules.
As expected, the ratio of observed to expected frequencies for the First Name high-risk class ranges from 1.12 to 1.17 for the four methods, suggesting there is no significant deviation between the observed and expected frequencies for this rule.

While the Geographic Proximity rule identifies if the patient and employee live within a fixed distance (0.25 miles), the Residential Street rule identifies if the patient and user live on the same street. Limiting the high-risk criteria to street name results in the higher ratio of observed to expected for the Residential Street rule than the ratio of observed to expected for Geographic Proximity rule.

Table 3 shows the percentage of observed and expected high-risk accesses for each high-risk type in a sample of 100,000 accesses. The average percentage of observed high-risk accesses ranged from 0.03% to 3.8%. Though the percentage of high-risk accesses for HCO Co-Worker is higher than other types (by more than 3%), the observed frequency of high-risk accesses does not deviate from the expected (see Table 2). This suggests that these accesses can be assigned the lowest priority for investigation. The percentage of observed and expected high-risk accesses for the rest of the high-risk rules is less than 1%, but given that millions of accesses are committed per week, this small percentage yields non-trivial numbers of high-risk accesses.

Table 3. The observed versus expected percentage for the high-risk access rules.

<table>
<thead>
<tr>
<th>High-Risk Rule</th>
<th>Observed</th>
<th>Expected</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Permute Patients</td>
</tr>
<tr>
<td>HCO Co-Worker</td>
<td>3.84%</td>
<td>3.84%</td>
</tr>
<tr>
<td>First Name</td>
<td>0.25%</td>
<td>0.21%</td>
</tr>
<tr>
<td>Last Name</td>
<td>0.14%</td>
<td>0.09%</td>
</tr>
<tr>
<td>Geographic Proximity</td>
<td>0.16%</td>
<td>0.07%</td>
</tr>
<tr>
<td>Residential Street</td>
<td>0.12%</td>
<td>0.03%</td>
</tr>
<tr>
<td>Department Co-Worker</td>
<td>0.04%</td>
<td>0.01%</td>
</tr>
</tbody>
</table>

Minimum rate of high-risk access explained away

Table 4 summarizes the average rate (over 10 experiments) at which the observed high-risk accesses can be explained with clinical reasons. Notably, the selected set of explanations accounted for less than 50% of the accesses.

Table 4. The rate at which high-risk alerts would be explained away.

<table>
<thead>
<tr>
<th>High-Risk Rule</th>
<th>Observed Accesses Explained Away</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>HCO Co-Worker</td>
<td>38.78%</td>
<td>0.67</td>
</tr>
<tr>
<td>First Name</td>
<td>35.59%</td>
<td>2.54</td>
</tr>
<tr>
<td>Last Name</td>
<td>21.43%</td>
<td>2.88</td>
</tr>
<tr>
<td>Geographic Proximity</td>
<td>24.79%</td>
<td>3.13</td>
</tr>
<tr>
<td>Residential Street</td>
<td>16.11%</td>
<td>4.88</td>
</tr>
<tr>
<td>Department Co-Worker</td>
<td>43.90%</td>
<td>9.32</td>
</tr>
</tbody>
</table>

Table 5 summarizes the distribution of the explanations per high-risk rule. The highest number of high-risk accesses is explained with the Clinical documentation explanation for all high-risk rules, with the percentage of accesses explained in the range of 15% to 43%. Scheduled Appointment explains 2% to 8% of the high-risk accesses. The other four explanations explain less than 5% of the high-risk accesses for all high-risk rules. The Clinical Documentation explanation shows high standard deviation (9.26) for the high-risk rule Department Co-Worker because of two outlier experiments with the highest and lowest number of explained accesses, respectively.

Hypothesis test

Table 6 shows the $\chi^2$ result for a sample $S$ of 100,000 unique accesses that are devoid of self-accesses. The expected number of accesses for this experiment was simulated through the permutation method (i.e., shuffling the list of patients and keeping list of employee fixed in the observed accesses). It should be noted that we did not include the Co-Worker rule in the $\chi^2$ test because the results showed that there was no difference between the observed and expected accesses for this high-risk class.
The result of the $\chi^2$ for high-risk rules Last Name, Geographic Proximity, Residential Street and Department Co-Worker indicated a probability < 0.0001. This is below the 0.01 significance level, such that we accept the alternative hypothesis for these high-risk rules (i.e., the difference between the observed and expected frequency of high-risk accesses for these rules is statistically significant). The result of the $\chi^2$ for First Name indicated a probability of 0.0113, which is above the 0.01 significance level, such that we reject the alternative hypothesis (i.e., there is no significant difference between observed and expected frequencies of high-risk accesses). This is notable because it suggests that our control rule is functioning correctly.

### Table 5. Distribution of explanations per high-risk rule (STD DEV = Standard dev).

<table>
<thead>
<tr>
<th>High-Risk Rule</th>
<th>% Observed Accesses Explained Away</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Scheduled Appointment</td>
</tr>
<tr>
<td></td>
<td>%</td>
</tr>
<tr>
<td>HCO Co-Worker</td>
<td>7.50</td>
</tr>
<tr>
<td>First Name</td>
<td>8.76</td>
</tr>
<tr>
<td>Last Name</td>
<td>2.23</td>
</tr>
<tr>
<td>Geographic Proximity</td>
<td>6.33</td>
</tr>
<tr>
<td>Residential Street</td>
<td>2.99</td>
</tr>
<tr>
<td>Department Co-Worker</td>
<td>8.26</td>
</tr>
</tbody>
</table>

### Table 6. Results of the $\chi^2$ test for goodness of fit between the observed and expected with one degree of freedom per experiment.

<table>
<thead>
<tr>
<th>High-Risk Rule</th>
<th>Observed</th>
<th>Expected</th>
<th>$\chi^2$</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>First Name</td>
<td>245</td>
<td>208</td>
<td>6.42</td>
<td>0.0113</td>
</tr>
<tr>
<td>Non High-Risk</td>
<td>99755</td>
<td>99792</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Last Name</td>
<td>140</td>
<td>91</td>
<td>25.87</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Non High-Risk</td>
<td>99860</td>
<td>99909</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Geographic Proximity</td>
<td>166</td>
<td>66</td>
<td>150.1</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Non High-Risk</td>
<td>99834</td>
<td>99934</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Residential Street</td>
<td>115</td>
<td>28</td>
<td>267.29</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Non High-Risk</td>
<td>99885</td>
<td>99972</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Department Co-Worker</td>
<td>36</td>
<td>11</td>
<td>54.58</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Non High-Risk</td>
<td>99964</td>
<td>99989</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Discussion

This study examined the extent to which high-risk EHR access rules are plausible in practice. Our empirical investigation illustrates that the observed rate at which high-risk rules are triggered is higher, at a statistically significant level, than what one would expect at random for several typical classes of high-risk behavior. This significant deviation suggests that there may be systematic EHR user behavior that requires further investigation, implying those rules may hold merit. Still, not all rules deviate to the same degree. In this respect, we further believe that the magnitude of the deviation of the observed frequency of high-risk accesses from their expected frequency obtained from each high-risk class may be a plausible measure to assist in the prioritization of auditing rules in emerging game theoretic frameworks.18
Geographical Proximity and Residential Street rules are designed to capture the same high-risk behavior (i.e. a user accessing records of a patient living in close geographic vicinity of the user). However, these two rules yield different results in terms of number of accesses flagged and the deviation of observed frequency from expected frequency of flagged accesses. Notably, the user and patient in 15% of the accesses flagged by the Residential Street rule do not live within 0.25 miles of each other, and the Geographical Proximity rule missed 37% of the Residential Street rule accesses. Also the length of the streets in the city varies from 0.4 miles to over 10 miles leading to a non-uniform application of the geographic vicinity criteria. This result indicates that rule definitions play an important role in effectively capturing high-risk behavior.

Despite their potential, high-risk access rules often have a high false positive rate. This makes them prohibitively expensive for HCOs to systematically investigate, which is a concern given the limited budgets available to privacy officers. However, we show that high-risk access rules can be complemented through an explanation-based model, such that many accesses can be explained away by valid TPO reasons (16% to 44% depending on the high-risk rule at minimum). We believe this is notable because it suggests that high-risk rules and explanations are not correlated. Nonetheless, we believe that, in this setting, the explanations can be used to prioritize high-risk accesses for manual investigation. The unexplained accesses can be considered high priority for investigation, while the explained accesses can be ranked using the type and number of explanations available for the access.

There are, however, several limitations of this study that we wish to highlight for future investigations. First, it should be recognized that a rule-based auditing system is inherently limited by its reliance on predefined rules, which themselves are often based on domain expertise. More broadly there are many possible reasons for inappropriate access. As a result, the access coverage (i.e., proportion of accesses effectively monitored by high-risk rules) is low. Second, rule-based flagging of high-risk accesses is dependent on the correctness and completeness of EHR data. Incorrect EHR data (e.g., the wrong patient’s last name is entered into the EHR system) leads to gaps in identifying a potential high-risk access. Similarly, an incomplete address or a P.O. Box would lead to gaps in flagging a potentially high-risk access according to the Residential Proximity or Residential Street rule. Third, an explanation-based system relies solely on the data stored in the database to generate explanation for an access. Missing information (or non-documented relationships) may result in few unexplained appropriate accesses. For example, EHR systems maintain records of patient appointments with doctors, but they do not explicitly record the relationship between the doctor and the nurse working together at the appointment. Thus, the system cannot readily explain the access of patient’s record by the nurse working with the doctor, though the access in this case is appropriate. Other research has posited enhancing explanations with additional data learned from diagnosis information. Fourth, this study does not test if a flagged access is in fact an inappropriate access. A flagged access needs to be investigated manually by a privacy officer to determine if it is a true inappropriate access. However, this manual investigation is beyond the scope of this study. Fifth, in this study we only consider simple high-risk rules. In a future investigation we plan to study more complex and nested high-risk rules. Sixth, this study suggests that different high-risk rules yield different results, but does not investigate the reasons for the differences.

Finally, we note that this investigation focused on data from only a one-week period from a single medical center. As such, it will be necessary to validate these findings with data from a broader time period and other healthcare organizations.

**Conclusions**

This paper examined the rate of high-risk access rules in the electronic health record of a large healthcare organization. Specifically, we compared the observed and expected rates to ascertain the extent to which such rules are potentially useful in practice. The primary finding of this investigation was that such rules appear to detect behaviors that are statistically significantly different than what would transpire under random activities. There are many reasons why such deviation might transpire, but our investigation shows that such rules should not be dismissed.

**Acknowledgements**

This research was sponsored, in part, by grants 1526014 and 1536871 from the National Science Foundation and R01LM010207 from the National Institutes of Health.

**References**


Mining Electronic Health Records to Extract Patient-Centered Outcomes Following Prostate Cancer Treatment

Tina Hernandez-Boussard, PhD, Panagiotis D. Kourdis, PhD, Tina Seto, MS, Michelle Ferrari, RN, Douglas W. Blayney, MD, Daniel Rubin, MD, James D. Brooks, MD
Stanford University, School of Medicine, Stanford, CA

Abstract

The clinical, granular data in electronic health record (EHR) systems provide opportunities to improve patient care using informatics retrieval methods. However, it is well known that many methodological obstacles exist in accessing data within EHRs. In particular, clinical notes routinely stored in EHR are composed from narrative, highly unstructured and heterogeneous biomedical text. This inherent complexity hinders the ability to perform automated large-scale medical knowledge extraction tasks without the use of computational linguistics methods. The aim of this work was to develop and validate a Natural Language Processing (NLP) pipeline to detect important patient-centered outcomes (PCOs) as interpreted and documented by clinicians in their dictated notes for male patients receiving treatment for localized prostate cancer at an academic medical center.

Introduction

Prostate cancer is the most common malignancy in men with an estimated 21% prevalence among new cancer cases in males for 2016.(1) Given the excellent survival rates, patients undergoing clinical treatment for prostate cancer often focus on patient-centered outcomes of care to guide treatment choices, such as rates of urinary incontinence (UI) or irritative voiding symptoms and erectile/sexual dysfunction (ED).(2, 3) However, the measurement of these outcomes at a population level has been hindered due to the lack of availability of these outcomes in generalizable, large-scale study cohorts.

Under the US healthcare reform, there is increased focus on quality of health care delivery.(4) Increasingly, standardized quality metrics have been developed and proposed to measure key components of care across the full continuum of care delivery, including patient-centered outcomes.(5) Benchmarking and reporting such quality metrics from multiple health care providers offers a unique opportunity to improve clinical practice and ensure that patients are receiving high-quality care and treatment options that correspond to their personal values.

Granular clinical information important for quality assessment is routinely collected within electronic health records (EHRs).(6) Due to their significant recent adoption,(7) healthcare workers have identified remarkable benefits and significant challenges in using these data to approach a learning health-care system.(8) Particularly challenging is the fact that most information in EHRs is stored as unstructured free text.(9) However, many quality measures, including patient-centered outcomes, are captured in EHRs only as free text.(10)

We have designed and developed the infrastructure that can leverage routinely collected information from EHRs to efficiently and accurately assess clinicians’ documentation of important patient-centered outcomes following treatment for prostate cancer. Using urinary incontinence and erectile dysfunction as the example, we developed a Natural Language Processing (NLP) pipeline using the Java-based open source software GATE (General Architecture for Text Engineering) to parse strings containing clinical notes stored in the EHRs. This work incorporated electronic phenotypes for each PCOs that can improve its generalizability across systems.

Methods

Data Source

Our pipeline identified patients within a large academic EHR-system using ICD-9/10 and CPT codes. The healthcare system provides inpatient, outpatient and primary care services and has a fully functional Epic system installed since 2008. (Epic Systems Corporation, Verona Wisconsin) To improve cohort identification and validation, the EHR records were linked to the state California Cancer Registry, which includes detailed information on patients’ histology, pathology, disease progression and survival, as well as treatments received outside of our academic health care system. An Oracle relational database was internally deployed to organize the internal and external structured and unstructured data elements. The patient cohort includes demographics, healthcare encounters, diagnoses/problem lists, clinical reports (narratives & impressions), encounter notes/documents, lab &
diagnostics results, medications (down to the ingredient level), treatment plans, procedures, billing summaries, patient history, patient surveys, and follow-up/survival information. To develop vocabularies related to urinary incontinence, irritative voiding symptoms and erectile dysfunction, a minimum of 100 charts selected at random were reviewed and terminology extracted. We performed manual chart review to estimate the positive and negative predicted values of the workflow. We will randomly select 200 records. A urology research nurse manually validated these reports to create a gold standard.

**Patient-Centered Outcome Phenotypes**

Phenotypic algorithms were developed for identifying and extracting both cases and controls of UI and ED assessment from EHRs. Input categories include ICD-9/ICD-10 codes, billing codes, medications and vocabularies matched with existing ontologies from the National Center for Biomedical Ontology (NCBO) and Unified Medical Language System (UMLS) concepts.(11) To improve accuracy, vocabularies were also manually curated with EHR terms found during manual chart review. The phenotypes of each of the three PCOs are available through the publicly accessible repository PheKB (https://phekb.org) a knowledgebase for EHR-based phenotypes.(12)

**NLP Extraction of PCOs**

Our NLP pipeline analyzes the clinical narrative text of EHRs using GATE software.(13) GATE provides several customizable processing resources that perform specific NLP processing tasks, i.e. tokenizers, sentence splitters, gazetteers which annotate documents based on look-up lists of keywords, parsers etc. Below, we describe the building components of our GATE-based NLP application for extracting PCOs from clinical notes, which comprises (1) an ANNIE module to detect PCO mentions in narrative texts, (2) a ConText module to determine the semantic context of the PCO mentions, and (3) A JAPE module to annotate the PCO mentions in the text.

We first used the English tokeniser from GATE's ANNIE plugin component that splits the text into simple constituent tokens such as numbers, punctuation and words of different types. Next, sentence splitting was performed using the RegEx sentence splitter, which is suitable when faced with irregular inputs, such as those regularly encountered in the contents of EHR clinical notes.

Following tokenization and sentence splitting, a Hash Gazetteer (part of the ANNI E) was run on the clinical note to find occurrences of PCO mentions. We developed three types of Gazetteer list files containing the various PCO-related keywords in text that indicate PCO mentions. In addition to detecting the PCO mentions, our pipeline detects the context of the PCO mention using the ConText algorithm.(14) The ConText algorithm determined whether the PCO mentioned in the clinical report was negated, hypothetical, historical, or experienced by someone other than the patient. PCO mentions were Type I if the ConText algorithm did not negate the term but the text indicated a negated PCO. PCO mentions were Type II if the ConText algorithm did not negate the term, but the text indicated an “affirmed” PCO. Finally, PCO mentions were Type III if the ConText algorithm negated them and the text indicated a negated PCO. (Table 1)

PCOs were classified as follows: “Affirmed”, meaning the patient had the symptom; “Negated”, the patient did not have the symptom; or “Discussed Risk”, the clinician documented the discussion regarding risks of PCOs by treatment with the patient.

<table>
<thead>
<tr>
<th>Type</th>
<th>Description</th>
<th>Example sentences</th>
<th>Annotation Value</th>
<th>ConText Negated</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>No urinary incontinence</td>
<td>His urinary control is good.</td>
<td>Negative</td>
<td>No</td>
</tr>
<tr>
<td>II</td>
<td>Urinary incontinence</td>
<td>He reports urinary leakage.</td>
<td>Affirmed</td>
<td>No</td>
</tr>
<tr>
<td>III</td>
<td>No urinary incontinence</td>
<td>No urinary complaints.</td>
<td>Negative</td>
<td>Yes</td>
</tr>
</tbody>
</table>

A gazetteer list file was also used to include keywords related to discussion of post-operative risk. When PCO terms appear with these terms in a sentence, the corresponding sentence is considered as a discussion of post-operative risk with the patient, rather than a PCO per se. Additionally, a Gazetteer list is used to exclude terms that are used in an alternative context (e.g. the term leakage related to urinary incontinence vs. leakage around the foley catheter). Finally, the regular expression pattern-matching engine called Java Annotation Patterns Engine (JAPE) was used to
create annotations in the text of the clinical notes based on the patterns in the text detected by the Gazetteer files and ConText algorithm.

**Evaluation**

The gold standard was used to evaluate our pipeline. The 200 records in the gold standard were processed and recorded all mentions of PCOs and classified each according to the annotation value. These data were compared with the annotations in the gold standard and precision and recall metrics were calculated. The university’s Internal Review Board approved the study.

**Results**

We identified 7,109 male individuals who received treatment for prostate cancer from 2008 to 2016 from a single, large, academic medical center. Patients’ demographics are displayed in Table 2. Patients had a mean age at time of diagnosis of 65.1 (SD 8.9). The majority of patients had localized disease, with 65% Stage II and 83% with a Gleason \(\leq 7\).

Table 2. Patient Demographics for Prostate Cancer Patients in the EHR Database who received treatment between 2008-2016

<table>
<thead>
<tr>
<th>Variable</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>7,109</td>
</tr>
<tr>
<td>Age. Mean (SD)</td>
<td>65.19 (0.11)</td>
</tr>
<tr>
<td>Ethnicity, %</td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>69.68</td>
</tr>
<tr>
<td>Hispanic</td>
<td>2.15</td>
</tr>
<tr>
<td>Black</td>
<td>4.03</td>
</tr>
<tr>
<td>Asian</td>
<td>9.95</td>
</tr>
<tr>
<td>Other</td>
<td>14.18</td>
</tr>
<tr>
<td>Year of Diagnosis, %</td>
<td></td>
</tr>
<tr>
<td>2005</td>
<td>9.20</td>
</tr>
<tr>
<td>2006</td>
<td>11.11</td>
</tr>
<tr>
<td>2007</td>
<td>11.45</td>
</tr>
<tr>
<td>2008</td>
<td>10.92</td>
</tr>
<tr>
<td>2009</td>
<td>10.23</td>
</tr>
<tr>
<td>2010</td>
<td>9.44</td>
</tr>
<tr>
<td>2011</td>
<td>9.36</td>
</tr>
<tr>
<td>2012</td>
<td>7.98</td>
</tr>
<tr>
<td>2013</td>
<td>6.96</td>
</tr>
<tr>
<td>2014</td>
<td>7.05</td>
</tr>
<tr>
<td>2015</td>
<td>6.30</td>
</tr>
<tr>
<td>BMI, mean (SD)</td>
<td>18.87 (0.16)</td>
</tr>
<tr>
<td>Stage, %</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0.44</td>
</tr>
<tr>
<td>1</td>
<td>10.82</td>
</tr>
<tr>
<td>2</td>
<td>64.65</td>
</tr>
<tr>
<td>3</td>
<td>11.76</td>
</tr>
<tr>
<td>4</td>
<td>8.55</td>
</tr>
<tr>
<td>Unknown</td>
<td>3.77</td>
</tr>
<tr>
<td>Gleason, %</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.01</td>
</tr>
<tr>
<td>6</td>
<td>6.99</td>
</tr>
<tr>
<td>7</td>
<td>6.89</td>
</tr>
<tr>
<td>8</td>
<td>1.70</td>
</tr>
<tr>
<td>9</td>
<td>1.11</td>
</tr>
<tr>
<td>10</td>
<td>1.83</td>
</tr>
</tbody>
</table>
Table 3 demonstrates the number of PCOs extracted from the different types of clinical notes, either history & physical or progress. The majority of PCO information extracted from the NLP pipeline came from the progress notes for both urinary incontinence and erectile dysfunction as compared to the information found in the history & physical note set.

**Table 3. Number Of Notes with an Identified PCO, Stratified By Type Of Mention And Note Location**

<table>
<thead>
<tr>
<th>Note type</th>
<th>Urinary Incontinence</th>
<th>Erectile/Sexual Dysfunction</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>History &amp; Physical (n = 10093)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Affirmed</td>
<td>362</td>
<td>570</td>
</tr>
<tr>
<td>Affirmed-History</td>
<td>75</td>
<td>37</td>
</tr>
<tr>
<td>Negated</td>
<td>1019</td>
<td>758</td>
</tr>
<tr>
<td>Negated-History</td>
<td>65</td>
<td>8</td>
</tr>
<tr>
<td>Discuss Risk</td>
<td>1380</td>
<td>1267</td>
</tr>
<tr>
<td><strong>Progress (n =155274)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Affirmed</td>
<td>7183</td>
<td>6359</td>
</tr>
<tr>
<td>Affirmed-History</td>
<td>1156</td>
<td>0</td>
</tr>
<tr>
<td>Negated</td>
<td>6734</td>
<td>5770</td>
</tr>
<tr>
<td>Negated -History</td>
<td>1122</td>
<td>293</td>
</tr>
<tr>
<td>Discuss Risk</td>
<td>3587</td>
<td>3410</td>
</tr>
</tbody>
</table>

The patient-level assessments for the two PCOs in the EHR notes for urinary incontinence and erectile dysfunction within our study cohort are displayed in Table 4. The results indicate that ED was assessed in 77.5% of the patients while UI in 59.4% prior to the primary treatment and 29% and 33% immediately following treatment.

**Table 4. Rates Of PCO Assessment Pre- And Post-Treatment, Stratified By Affirmed And Negated Mentions**

<table>
<thead>
<tr>
<th>Time Period</th>
<th>PCO</th>
<th>Affirmed</th>
<th>Negated</th>
<th>Assessment Rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Pre-Treatment</strong></td>
<td>UI</td>
<td>289</td>
<td>1517</td>
<td>59.4</td>
</tr>
<tr>
<td></td>
<td>ED</td>
<td>1409</td>
<td>947</td>
<td>77.5</td>
</tr>
<tr>
<td><strong>0-3 Months Post-Treatment</strong></td>
<td></td>
<td>581</td>
<td>450</td>
<td>32.5</td>
</tr>
<tr>
<td></td>
<td>ED</td>
<td>690</td>
<td>218</td>
<td>28.6</td>
</tr>
</tbody>
</table>

We estimated that patients had a pre-treatment rate of UI to be 16%, with varying severity and pre-treatment rate of 30% for ED (Figure 1). The rates for UI remained at approximately 40-50% 1-2 years after surgery and the rates for ED were between 45-70% following treatment.
The F-measure accuracy scores for the UI annotations against an unseen gold standard manually annotated by our board-certified Urology nurse was 87% affirmed, 96% negated, and 91% discuss risk. For the ED annotations the F-measure was 85%, 92%, and 90%, respectively. (Table 5)

<table>
<thead>
<tr>
<th>Patient-Centered Outcome</th>
<th>Affirmed</th>
<th>Negated</th>
<th>Discuss Risk</th>
</tr>
</thead>
<tbody>
<tr>
<td>Urinary Incontinence</td>
<td>Precision</td>
<td>0.8667</td>
<td>0.9444</td>
</tr>
<tr>
<td></td>
<td>Recall</td>
<td>0.8667</td>
<td>0.9714</td>
</tr>
<tr>
<td></td>
<td>F1-score</td>
<td>0.8667</td>
<td>0.9577</td>
</tr>
<tr>
<td>Erectile/Sexual Dysfunction</td>
<td>Precision</td>
<td>0.8567</td>
<td>0.9312</td>
</tr>
<tr>
<td></td>
<td>Recall</td>
<td>0.8412</td>
<td>0.9011</td>
</tr>
<tr>
<td></td>
<td>F1-score</td>
<td>0.8489</td>
<td>0.9159</td>
</tr>
</tbody>
</table>

Discussion

Under healthcare reform, stakeholders emphasize the importance of quality and patient-centered healthcare. Accordingly, government officials seek quality metrics that include meaningful outcomes, including those that go beyond mortality and recurrence. However, efforts to establish guidelines for prostate cancer treatment have been difficult to formulate due to insufficient evidence regarding relative benefits and risks of the different treatment options, particularly important patient-centered outcomes such as urinary and erectile dysfunction.(2, 15) Here we found clinicians regularly recorded this information in the unstructured text of EHRs. The current work expands on existing models by focusing on ontology-based dictionaries to annotate free text associated with patient-centered outcomes. We have demonstrated our ability to efficiently and accurately use our methods on EHRs for quality assessment of patient-centered outcomes.

The rates of urinary incontinence and erectile dysfunction identified through our pipeline were highly concordant with rates from surveys that are reported in the literature.(16-18) This highlights the ability of NLP algorithms to leverage routinely collected information from EHRs and efficiently and accurately assess clinicians’ documentation.
of patient-centered outcomes. Our study suggests that patient-centered outcomes, particularly for prostate cancer patients, are important and monitored not only for the patient but also for the clinician. Given the lack of population-based information on these important outcomes, NLP tools can clearly advance research and evidence in this area.

Our NLP pipeline shows promising results for the task of advancing patient-centered outcomes research (PCOR). To date, most quality metric analyses have used administrative datasets. While these datasets are easy to use and readily available, they are generated for billing purposes and lack important clinical details. Paper records may have more detail, but cannot be algorithmically assessed. Prospective studies are not readily available and often contain ascertainment bias. These issues have limited the efficiency and effectiveness of efforts for quality improvement, organizational learning, and comparative effectiveness research, particularly in the area of PCOR. Using our NLP pipeline, we can obtain population rates on PCOs efficiently and with high certainty. Such work can significantly advance the field and depth of PCOR.

There are limitations to our study. First, our algorithms have been developed and tested in a single academic center. However, the clinical terms used in our algorithms are disseminated with a national repository (pheKB.org) and multiple clinicians across different healthcare settings have vetted the clinical terms. We will be testing our algorithms in another healthcare system to ensure their generalizability. A second limitation is that our system only reports what the clinician documents and does not capture patient-reported outcomes. Many studies have highlighted that clinicians’ documentation can be sparse and incomplete. However, our previous work indicated that these patient-centered outcomes are more prevalent in the clinical text than elsewhere in the EHRs. Finally, our algorithms have been developed using American English and vocabularies and language rules would need to be developed for usage where another language is used in the EHR.

A significant benefit of our NLP pipeline is that it leverages multiple sources of data to identify patients and outcomes, including registry data, quality of life surveys, and other information from clinical trials within our institute. Our dataset is updated regularly and it will likely continue to perform well, including the assessment of these outcomes immediately after the introduction of new treatments and technologies.

Conclusions

We developed an NLP pipeline for detecting clinical mentions of patient-centered outcomes in prostate cancer patients. The current performance of the system appears sufficient to be used for population-based health management and to enhance evidence needed to help patients identify treatment pathways that reflect their healthcare values. Given the importance of these events under the healthcare reform, wide deployment of fully computerized algorithms that can reliably capture PCOs will have numerous applications for the healthcare industry. These approaches are the basis of a learning healthcare system and target in fostering healthcare quality through information technology and data resource utilization.

References

Abstract

There is growing interest in applying machine learning methods to Electronic Medical Records (EMR). Across different institutions, however, EMR quality can vary widely. This work investigated the impact of this disparity on the performance of three advanced machine learning algorithms: logistic regression, multilayer perceptron, and recurrent neural network. The EMR disparity was emulated using different permutations of the EMR collected at Children's Hospital Los Angeles (CHLA) Pediatric Intensive Care Unit (PICU) and Cardiothoracic Intensive Care Unit (CTICU). The algorithms were trained using patients from the PICU to predict in-ICU mortality for patients on a held out set of PICU and CTICU patients. The disparate patient populations between the PICU and CTICU provide an estimate of generalization errors across different ICUs. We quantified and evaluated the generalization of these algorithms on varying EMR size, input types, and fidelity of data.

Introduction

Electronic Medical Records (EMR) are currently adopted by approximately 84% of hospitals in the United States\(^1\). With the recent success of machine learning, there are numerous research efforts to extract medically relevant and actionable information from troves of EMR using machine learning algorithms\(^2\). However, due to varying data entry protocols and bespoke implementations of EMR systems, data quality can vary widely across institutions\(^3\). This work investigated the dependence of advanced machine learning algorithms on the varying EMR quality.

EMR data discrepancies were emulated using permutations of EMR collected from Children’s Hospital Los Angeles (CHLA) Pediatric Intensive Care Unit (PICU) and Cardiothoracic Intensive Care Unit (CTICU). Permutations studied included varying the number of patient encounters available for training; which variables were included and excluded as model inputs; and varying the fidelity of drug information. The effects of the permutations were analyzed on established machine learning algorithms: logistic regression, multilayer perceptron, and recurrent neural network. The analysis of these algorithms illustrate the impact of data quality on varying levels of model complexity.

The task used to evaluate these algorithms was prediction of in-ICU mortality: whether the patient survived their ICU encounter. Mortality was chosen because it provides an objective measure of a model’s ability to extract information from EMR\(^4\). Predictions were generated for each algorithm after 12 hours of observation and used to create Area Under the Receiver Operating Curves (AUROC).

EMR Source

This study used anonymized EMR collected in the Pediatric Intensive Care Unit (PICU) and Cardiothoracic Intensive Care Unit (CTICU) at Children’s Hospital Los Angeles (CHLA) between 2002 and 2016. The database contained 21,881 ICU encounters (defined as the time between a patient’s admission to and discharge from the ICU). Of these ICU encounters, 16,706 were from the PICU (12,093 patients with 4.85% mortality) and 5,175 were from the CTICU (3,088 patients with 3.32% mortality). Each encounter contained irregularly charted measurements of the patient’s physiology (e.g. heart rate, respiratory rate), laboratory test results (e.g. glucose, creatinine), and treatments (e.g. intubation, epinephrine) administered throughout their ICU stay. Additionally, each encounter was accompanied by static information describing demographics (e.g. age, sex, and race), diagnosis, and disposition at the end of the ICU encounter (whether they survived).

Pre-Processing EMR for Machine Learning

The collected EMR was a list of observations and treatments (in long format) charted by the clinical team, time-stamped and uniquely identified to each patient. The following sub-sections describe pre-processing techniques that restructured the EMR for machine learning purposes.
Data Curation: In collaboration with physicians, the dataset was curated to remove erroneous observations and aggregate variables of similar groups. For example, minimum and maximum values were defined for each observation such that measurements which were incompatible with life (e.g. heart rate of 1200 beats per minute) were corrected. Additionally, different observations of similar variables such as invasive and non-invasive blood pressures were combined into a single variable when medically appropriate.

Data Pivot: To accommodate the structure required for machine learning algorithms, the long format EMR was pivoted to wide format. In other words, the EMR was reshaped such that the timestamps were the rows, observations were the columns, and values populated the cells. The result was a sparse, irregularly sampled matrix for each patient, henceforth referred to as a patient-matrix. Figure 1 illustrates an example of this process.

Data Standardization: Data standardization was applied to facilitate training of the machine learning models while maintaining medical interpretability of the variables. Physiologic observations (labs and vitals) were converted to z-scores using the means and standard deviations computed from the training set (described later). This process was done to provide numerical stability when training the models (via stochastic gradient descent). Treatments administered to the patients such as drugs and interventions were normalized to values between [0, 1] using upper limits defined in collaboration with clinicians. Values of 0 indicated the presence of no drugs while values between (0, 1) indicated percentage of dosage from the maximum dose in the training data.

Data Imputation: To control the scope of this study, the imputation method was kept simple to align with the data standardization methods and maintain medical interpretability. Physiologic observations were forward-filled following an initial measurement until the next available measurement. This choice was based on the clinical insight that measurements are more frequently taken during times of hemodynamic instability and less frequently when the patient appears stable. If a physiologic variable had no recorded measurements for an entire encounter, that variable was set to zero. Because of the prior standardization of physiologic variables, this imputation was equivalent to imputing the population mean derived from the training set. For drugs and interventions, missing values were imputed with zero. Since these variables were normalized to [0, 1], a zero entry indicated the absence of treatment. Figure 2 depicts the imputation process.

Figure 1: Illustration of pivot operation (and curation) applied to long format EMR to create a patient-matrix. Note that for a particular patient, a column of the pivoted data may have a blank entry in each row as in PEEP above.

Figure 2: Illustration of imputing a patient-matrix. Data standardization is not applied here to show realistic values.
Table 1: Percentage of demographic groups across the datasets. The numbers in parenthesis are the number of encounters of each dataset.

<table>
<thead>
<tr>
<th>Mortality</th>
<th>Train % (8,404)</th>
<th>Valid % (4,122)</th>
<th>PICU Test % (4,176)</th>
<th>CTICU Test % (5,175)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Survived</td>
<td>95.34</td>
<td>95.03</td>
<td>94.88</td>
<td>96.67</td>
</tr>
<tr>
<td>Died</td>
<td>4.66</td>
<td>4.97</td>
<td>5.12</td>
<td>3.33</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Age Group</th>
<th>Mortality</th>
<th>Train % (8,404)</th>
<th>Valid % (4,122)</th>
<th>PICU Test % (4,176)</th>
<th>CTICU Test % (5,175)</th>
</tr>
</thead>
<tbody>
<tr>
<td>[0, 1)</td>
<td></td>
<td>17.32</td>
<td>16.79</td>
<td>16.62</td>
<td>57.58</td>
</tr>
<tr>
<td>[1, 5)</td>
<td></td>
<td>23.41</td>
<td>25.23</td>
<td>23.42</td>
<td>20.37</td>
</tr>
<tr>
<td>[5, 10)</td>
<td></td>
<td>17.65</td>
<td>17.78</td>
<td>17.58</td>
<td>12.73</td>
</tr>
<tr>
<td>[10, 18)</td>
<td></td>
<td>34.28</td>
<td>32.97</td>
<td>36.04</td>
<td>7.77</td>
</tr>
<tr>
<td>18+</td>
<td></td>
<td>7.35</td>
<td>7.23</td>
<td>6.35</td>
<td>1.55</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Race / Ethnicity</th>
<th>Mortality</th>
<th>Train % (8,404)</th>
<th>Valid % (4,122)</th>
<th>PICU Test % (4,176)</th>
<th>CTICU Test % (5,175)</th>
</tr>
</thead>
<tbody>
<tr>
<td>African American</td>
<td></td>
<td>8.06</td>
<td>7.67</td>
<td>8.48</td>
<td>6.03</td>
</tr>
<tr>
<td>Asian</td>
<td></td>
<td>6.62</td>
<td>6.48</td>
<td>6.15</td>
<td>9.45</td>
</tr>
<tr>
<td>Caucasian</td>
<td></td>
<td>19.17</td>
<td>18.49</td>
<td>18.63</td>
<td>18.16</td>
</tr>
<tr>
<td>Hispanic</td>
<td></td>
<td>51.91</td>
<td>52.26</td>
<td>51.32</td>
<td>49.82</td>
</tr>
<tr>
<td>Other</td>
<td></td>
<td>14.22</td>
<td>15.11</td>
<td>15.42</td>
<td>16.54</td>
</tr>
</tbody>
</table>

Train, Validation, Test Partitioning: The data were partitioned into train, validation, and test sets with the demographics summarized in Table 1. The training set consisted of 50% percent of the PICU patients (8,404 encounters, 4.66% mortality), selected at random. Half of the remaining PICU patients (4,122 encounters, 4.97% mortality) constituted the validation set used to tune the algorithm’s hyper-parameters. Model performance was evaluated on two test sets: the last 25% of the patients from the PICU (4,176 encounters, 5.12% mortality); and all of the patients in the CTICU (5,175 encounters, 3.32% mortality). These test sets were not used during any model training and development. Since the CTICU specializes in congenital heart defects and generally treat younger populations than the PICU (see age group distributions in Table 1), the CTICU was used as a surrogate to represent an institution with disparate patient population and treatments. Scores on the PICU test set show the algorithms’ generalization to similar ICUs while scores on the CTICU test set estimate performance across different ICUs.

Data Permutations

Permutations of CHLA EMR were generated to emulate the data disparity found across different EMR systems. The following subsections describe three types of permutations: training set fraction (varying data size), input types (varying input variables), and drug encoding (varying data fidelity).

Training Set Fraction: A common concern with applying machine learning techniques such as neural networks is the amount of data required for training the models to perform well.6,7 This question is particularly relevant because the number of patient records can vary widely across hospitals. Two primary factors contribute to the disparity: 1) the size of the institution (number of critical care beds available) and; 2) the institution’s date of EMR adoption. To assess the dependence of algorithm performance on training data size, the algorithms were trained on successively decreasing sizes – 100, 90, 75, 50, 25, 10% – of the full training dataset described earlier. The resulting model from each of these trainings was then evaluated on the same complete PICU and CTICU test sets. The validation set remained fixed.

Input Types: The medical community has had many discussions on the validity of incorporating treatments and interventions into patient acuity scores.8,9 Since treatments typically are not applied randomly, their application is informative of patient state. However, the ‘insight’ or value of such information is complicated by non-standard practices across clinicians when administering treatments.10 Conversely, the same physiologic measurement can have different interpretations that depend on the treatments being received. For instance a heart rate may appear normal in isolation, but the same heart rate in the presence of vasopressors may indicate a higher risk of mortality. This permutation study parses the dependence of the algorithms on internal and external inputs. Internal refers to the observations describing
the patient’s status (labs and vitals) and external refers to the inputs that describe the treatments given to the patient (drugs and interventions). Figure 3a illustrates the splitting of the inputs from baseline (combined) to internal and external inputs.

**Drug Encoding:** The fidelity of the captured information may vary from institution to institution. For instance, data may be limited and contain only an indication of whether or not a particular drug was administered, but not its exact quantity. Moreover, treatment protocols vary from institution to institution. For example, one institution may preferentially provide albumin infusions while another may favor saline for fluid therapy. This permutation experiment emulates the variance in data fidelity using two drug encoding schemes: 1) decimating drug values to binary indicators; and 2) indexing the drugs based on biomedical ontologies by encoding them to the National Library of Medicine’s Medical Subject Headings (MeSH). Figure 3b illustrates the baseline input’s encoding to these two schemes.

**Figure 3:** The complete patient-matrix contains information about the patient’s internals (vitals, labs) and externals (drug administrations and interventions). a) Illustrates parsing of internals and externals from the complete patient-matrix. b) Shows the encoding of the drug subset in patient-matrix to the two permutation schemes: binary drug indicator and MeSH.

**Methods**

Three representation learning algorithms were assessed: logistic regression (LR), multilayer perceptron (MLP), and recurrent neural network (RNN). These algorithms were chosen because they illustrate the effects of data quality as a function of model complexity and have shown success in a variety of medical applications. The LR and MLP have been used in detection of prostate cancer, detection of Acute Respiratory Distress Syndrome (ARDS), and estimation of childhood asthma risk. The RNN has recently shown success in medical tasks such as predicting critical decompensation of patients on the floor, early detection of heart failure, and predicting ICU mortality. The following sub-sections describe more specifically the variant of each algorithm used. Figure 4 shows the architecture of the three models.

**Logistic Regression:** The logistic regression is one of the simplest machine learning models. It assumes that the input \((x)\) and output data \((y)\) are related by the mapping \(y = \sigma(W \cdot x)\), where \(\sigma(x) = 1 / (1 + e^{-x})\). The matrix of weights, \(W\), is learned from the data. For our task, the input to the LR is the pre-processed EMR derived patient-matrix, and the output is the patient’s mortality response, i.e. disposition (alive or not) at the end of ICU encounter. Although the input is a patient-matrix, the LR operates on individual time-slices (a vector) of the patient matrices.

**Multilayer Perceptron:** The MLP, also called a feed forward neural network, is a classic deep learning method which models the mapping from inputs to output with \(y = f^* (x)\), where \(f^*\) is a series of different functions called layers. The construction is loosely inspired by biological neural networks. Each node (hidden unit) in a layer is associated with a weight, \(w_{i,j}\), that connects to every node in the subsequent layer, analogous to the communication between neurons. The MLP architecture used here consisted of a 2-layer MLP, each with hidden units of 256 (chosen through grid search of validation performance). Note that the MLP also operates on time-slices of the patient matrix, and that the LR described above is equivalent to a 0-layer MLP when trained similarly using stochastic gradient descent.
Figure 4: Model architectures for a) logistic regression, b) multilayer perceptron, and c) recurrent neural network. Each model is composed of layers that operate on the patient’s input sequences (multi-colored vectors) to predict the patient’s in-ICU mortality. The numbers in each layer (purple dense layer and blue LSTM layer) are the number of hidden units of that layer.

Recurrent Neural Network: The RNN refers to the family of neural networks specialized for processing sequential data\textsuperscript{18,19}. The RNN can be regarded as an extension of the MLP that contains connections which feed each layer’s outputs back into itself. This recurrent structure makes RNNs ideal for processing EMR because the hidden states and feedback loop can elegantly integrate and incorporate past information about the patient with newly acquired measurements\textsuperscript{17}. This characteristic also makes it a fitting choice for modeling the dynamic disease progression encountered in a critical care setting. The particular RNN variant used is the Long Short Term Memory (LSTM)\textsuperscript{20}. The model is composed of a 2-layer LSTM, each with 397 hidden units (equal to the number of variables in the input patient-matrices). The output layer is a dense layer applied to every time-step of the output from the previous LSTM layer. Although the RNN ingests only a single time-slice (vector) from the patient matrix at each time step, its feedback loop effectively gives the RNN a look into the previous time-slices also. This is a contrast to the LR and MLP.

The above three models were trained using mini-batch sizes of 128 and optimized using RMSprop\textsuperscript{21}, a gradient descent optimizer that utilizes an adaptive learning rate, to minimize the binary cross-entropy between the model’s predicted in-ICU probability and the patient’s true mortality response. Each model’s layers were initialized using Glorot uniform, a method that samples weights from a uniform distribution with the variance scaled based on the number of parameters of the preceding and subsequent layers. After every epoch of training (an epoch is defined as one cycle through the training set), the models predicted the in-ICU mortality for patients in the validation set and the binary cross-entropy was computed. If the binary cross-entropy did not decrease after 15 epochs, the model’s learning rates were decimated by 5. After 2 reductions, training is stopped and the weights associated with the best validation performance are used to predict the mortality of the patients in the two test sets. During training, dropout of 20\% was applied to the input sequences (across time) as a data augmentation parameter.
Results

Model training for each data permutation study was iterated five times, with each iteration corresponding to a different initialization point (for the model weights) and different validation stopping point. The resulting model from each iteration was then applied to the PICU and CTICU test sets, and the Area Under the Receiver Operating Curve (AUROC) was generated from the model's in-ICU mortality prediction after 12 hours of observation. The numbers reported in Tables 2–4 are the mean and standard deviations of the AUROCs from these iterations. Baseline (BL) refers to the processed-EMR containing all available training encounters and input variables. Baseline is the same across the different EMR permutation studies and occupies the first row of Tables 2–4 while being named accordingly (100%, Combined, NoEncoding).

Table 2 shows the performance of each algorithm as a function of the number of patients in the training set. Table 3 shows the performance when limiting the model inputs to either only internal (physiologic observations) or only external (treatments) variables. Table 4 shows the performance as a function of drug encoding strategy. Figure 5 offers a visual summary of Tables 2–4. As a point of reference, AUROCs were also computed from the predictions of two standard pediatric severity of illness scoring systems, PIM2 and PRISM3. PIM2’s AUROCs on the PICU and CTICU test sets were 0.868 and 0.774, while those of PRISM3-12 were 0.856 and 0.723, respectively.

<table>
<thead>
<tr>
<th>Training Set Fraction</th>
<th>LR</th>
<th>MLP</th>
<th>RNN</th>
<th>LR</th>
<th>MLP</th>
<th>RNN</th>
</tr>
</thead>
<tbody>
<tr>
<td>100% (BL)</td>
<td>0.907 +/- 0.001</td>
<td>0.909 +/- 0.001</td>
<td>0.921 +/- 0.003</td>
<td>0.803 +/- 0.002</td>
<td>0.808 +/- 0.001</td>
<td>0.801 +/- 0.008</td>
</tr>
<tr>
<td>75% (6306)</td>
<td>0.899 +/- 0.001</td>
<td>0.904 +/- 0.001</td>
<td>0.914 +/- 0.001</td>
<td>0.808 +/- 0.004</td>
<td>0.813 +/- 0.001</td>
<td>0.809 +/- 0.011</td>
</tr>
<tr>
<td>50% (4204)</td>
<td>0.894 +/- 0.002</td>
<td>0.902 +/- 0.001</td>
<td>0.905 +/- 0.002</td>
<td>0.795 +/- 0.002</td>
<td>0.807 +/- 0.002</td>
<td>0.791 +/- 0.003</td>
</tr>
<tr>
<td>25% (2102)</td>
<td>0.890 +/- 0.002</td>
<td>0.896 +/- 0.001</td>
<td>0.904 +/- 0.003</td>
<td>0.790 +/- 0.002</td>
<td>0.804 +/- 0.002</td>
<td>0.793 +/- 0.007</td>
</tr>
<tr>
<td>10% (840)</td>
<td>0.867 +/- 0.002</td>
<td>0.891 +/- 0.001</td>
<td>0.881 +/- 0.003</td>
<td>0.773 +/- 0.002</td>
<td>0.775 +/- 0.003</td>
<td>0.783 +/- 0.011</td>
</tr>
</tbody>
</table>

Table 3: In-ICU mortality AUC on the test sets as a function of input types used.

<table>
<thead>
<tr>
<th>Input Types</th>
<th>LR</th>
<th>MLP</th>
<th>RNN</th>
<th>LR</th>
<th>MLP</th>
<th>RNN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Combined (BL)</td>
<td>0.907 +/- 0.001</td>
<td>0.909 +/- 0.001</td>
<td>0.921 +/- 0.003</td>
<td>0.803 +/- 0.002</td>
<td>0.808 +/- 0.001</td>
<td>0.801 +/- 0.008</td>
</tr>
<tr>
<td>Internals</td>
<td>0.899 +/- 0.001</td>
<td>0.903 +/- 0.001</td>
<td>0.917 +/- 0.002</td>
<td>0.801 +/- 0.002</td>
<td>0.808 +/- 0.002</td>
<td>0.793 +/- 0.005</td>
</tr>
<tr>
<td>Externals</td>
<td>0.841 +/- 0.001</td>
<td>0.833 +/- 0.002</td>
<td>0.841 +/- 0.002</td>
<td>0.727 +/- 0.003</td>
<td>0.721 +/- 0.001</td>
<td>0.679 +/- 0.005</td>
</tr>
</tbody>
</table>

Table 4: In-ICU mortality AUC on the test sets as a function of the drug encoding used.

<table>
<thead>
<tr>
<th>Drug Encoding</th>
<th>LR</th>
<th>MLP</th>
<th>RNN</th>
<th>LR</th>
<th>MLP</th>
<th>RNN</th>
</tr>
</thead>
<tbody>
<tr>
<td>None (BL)</td>
<td>0.907 +/- 0.001</td>
<td>0.909 +/- 0.001</td>
<td>0.921 +/- 0.003</td>
<td>0.803 +/- 0.002</td>
<td>0.808 +/- 0.001</td>
<td>0.801 +/- 0.008</td>
</tr>
<tr>
<td>Binary</td>
<td>0.909 +/- 0.001</td>
<td>0.912 +/- 0.001</td>
<td>0.922 +/- 0.002</td>
<td>0.803 +/- 0.004</td>
<td>0.807 +/- 0.001</td>
<td>0.794 +/- 0.009</td>
</tr>
<tr>
<td>MeSH</td>
<td>0.906 +/- 0.001</td>
<td>0.910 +/- 0.001</td>
<td>0.922 +/- 0.001</td>
<td>0.782 +/- 0.001</td>
<td>0.785 +/- 0.002</td>
<td>0.752 +/- 0.007</td>
</tr>
</tbody>
</table>

Discussion

A large performance disparity was observed between the PICU and CTICU test sets across all data permutations and algorithms. The performance on CTICU data was approximately 10% lower than on the PICU data. This is not surprising since the models were trained on PICU encounters only, and the CTICU patient population and treatments can be very different from those in the PICU. On the PICU test set, the RNN models substantially and consistently outperformed their LR and MLP counterparts across the different permutation experiments, with the lone exception coming from the experiment using only 10% (840 encounters) of the training data for model development. On the CTICU test set, the MLP models consistently outperformed their RNN and LR counterparts; the lone exception was again observed in the 10% of training data experiment, where RNN outperformed MLP. The results indicate that...
the MLP learned more generalizable patterns than the RNN, despite the latter’s more advanced sequence-processing capabilities. The MLP proved more robust, relative to the RNN, in transferring across disparate ICUs. The RNN’s large number of parameters (∼15 times more trainable parameters) may have led to overfitting to the PICU data.

Table 2 and Figure 5a illustrate reduction in each model’s performance on both PICU and CTICU as the amount of training data was reduced. When trained using only 840 encounters (training fraction of 10%), the models perform reasonably well on the ∼4000 encounters in the PICU and CTICU test sets, with the weakest performing model (LR) scoring 0.867 and 0.773 on the PICU and CTICU, respectively.

Table 3 and Figure 5b show little performance degradation when external information was removed from the baseline inputs. This indicates that internal information contributed almost entirely to the each model’s prediction of the patient’s severity of illness. Although externals alone performed significantly lower than internals and baseline, the performance increase from internals to baseline indicates that the models extracted additional information from the external variables.

The results in Table 4 and Figure 5c show that model performance, regardless of algorithm, remained about the same when changing drug information from real-valued to binary indicators (absent or present). This similar performance may be due to standard quantities and protocols frequently being followed during drug administration. Consequently, knowledge of a clinical decision to apply a particular drug, regardless of its numerical value, may be sufficient to understanding the patient’s severity of illness.

Table 4 and Figure 5c further show that on the PICU set, the models using baseline drug encoding performed similarly to models using MeSH-encoded drug information. On the CTICU set, however, the degradation of RNN performance from its baseline model to the one using MeSH was 5%. The RNN’s significantly worse degradation on the CTICU (compared to slight degradation on the PICU test set) was also observed in the baseline-to-externals experiment (Table 2 and Figure 5b). This pattern may be due to the RNN overfitting the information gleaned from external variables. The LR and MLP models process a time-slice in isolation of the other time-slices in a patient-matrix, but the RNN builds a history of the applied treatments. This sequential processing of data likely made the RNN learn to weigh external information more. Since the CTICU contained a more disparate distribution in treatments, over-utilizing the
information learned from the PICU may have improved its performance on the PICU but hindered its generalization to the CTICU.

Conclusions

This work measured the effects of varying EMR quality on the performance of three advanced machine learning algorithms: logistic regression, multilayer perceptron, and recurrent neural network. Three sets of permutations of CHLA EMR were generated to emulate the EMR data disparity across institutions. The first measures the performance of the models as a function of training data size. The second measures the dependence of the models on varying input types. The third measures the effects of varying fidelity in drug data. The algorithms were analyzed by measuring their performance in predicting in-ICU mortality in two different ICUs.

Performance of all three models degraded with decreasing size of the training set. Even when trained on 10% of the available encounters (840), all models still performed comparably with two standard pediatric severity of illness scores, PIM2 and PRISM3-12. Additionally, the MLP generalized better than the RNN when they were trained on PICU but tested on CTICU data.

Future work includes measuring the effects of the machine learning models on other varying types of EMR. For example, measuring the difference in performance due to varying temporal fidelity by incorporating waveform measurements from bedside monitors. Moreover, the analysis can be further strengthened by analyzing the performance on other clinically relevant tasks such as predicting ICU readmission.

Acknowledgements

We are grateful to the Laura P. and Leland K. Whittier Foundation for funding this work. We are also thankful to Sareen Shah, M.D. for providing significant medical domain expertise; Alysia Flynn for providing assistance in aggregating data; and Jon Williams and Alec Gunny for technical discussion and comments.

References


Classifying Acute Ischemic Stroke Onset Time using Deep Imaging Features

King Chung Ho1,2, William Speier, PhD2, Suzie El-Saden, MD2, Corey W. Arnold, PhD1-3

1Department of Bioengineering; 2Medical Imaging Informatics; 3Department of Radiological Sciences,
University of California, Los Angeles, CA

Abstract

Models have been developed to predict stroke outcomes (e.g., mortality) in attempt to provide better guidance for stroke treatment. However, there is little work in developing classification models for the problem of unknown time-since-stroke (TSS), which determines a patient’s treatment eligibility based on a clinical defined cutoff time point (i.e., <4.5hrs). In this paper, we construct and compare machine learning methods to classify TSS<4.5hrs using magnetic resonance (MR) imaging features. We also propose a deep learning model to extract hidden representations from the MR perfusion-weighted images and demonstrate classification improvement by incorporating these additional imaging features. Finally, we discuss a strategy to visualize the learned features from the proposed deep learning model. The cross-validation results show that our best classifier achieved an area under the curve of 0.68, which improves significantly over current clinical methods (0.58), demonstrating the potential benefit of using advanced machine learning methods in TSS classification.

1 Introduction

Stroke is the primary cause of long-term disability and the fifth leading cause of death in the United States, with approximately 795,000 Americans experiencing a new or recurrent stroke each year [1]. Several treatments exist for stroke, including intravenous and intra-arterial tissue plasminogen activator (IV/IA tPA), and mechanical thrombectomy (clot retrieval). Guidelines support selecting tPA treatment administration only within a maximum of 4.5 hours from stroke symptom onset due to the increased risk of hemorrhage for longer times since stroke (TSS). However, about 30% of the population have unknown time since stroke (TSS), making these patients ineligible for treatment with tPA despite the fact that their strokes may have actually occurred within the treatment window [2].

Predictive models have been made in attempt to predict stroke patient outcomes (e.g., mortality) using clinical variables (e.g., age) and imaging features (e.g., lesion volume) [3–5]. Additional algorithms are under development that attempt to predict patient response to a specific treatment [6]. While much work has been done in predicting stroke patient outcome and treatment response, there is limited work in determining TSS. Studies are underway to investigate the use of a simple imaging feature, a mismatch pattern between magnetic resonance (MR) diffusion weighted imaging (DWI), on which stroke pathophysiology is immediately visible, and fluid attenuated inversion recovery (FLAIR) imaging, on which strokes are not visible for 3-4 hours [7–9], to estimate TSS. The mismatch pattern is known as “DWI-FLAIR mismatch.” While this method is the current state-of-the-art for determining eligibility for thrombolytic therapy in cases of unknown TSS, computing mismatch is a difficult task that requires extensive training and for which clinician agreement has been found to be only moderate, leading to less accurate performance [10–12]. Separately, limited work has been done in TSS classification using MR perfusion-weighted image (PWIs) for TSS classification, which may contain information that encodes TSS [13].

Machine learning models have been applied widely and can achieve good classification performance for problems in the healthcare domain because of their ability to learn and utilize patterns from data to make prediction [14]. Recent developments in deep learning [15] have drawn significant research interest because of the technique’s ability to automatically learn feature detectors specific to the data for classification and prediction tasks, achieving state-of-the-art performance in challenging problems (e.g., ImageNet [16], video classification [17], etc.). In this work, we hypothesize that machine learning models can be used to better classify TSS by learning latent representative features from MR images. We developed a deep learning algorithm based on an autoencoder architecture [18] to extract imaging features (i.e., deep features) from PWIs and evaluate the effectiveness of four machine learning classifiers with and without the deep features to classify TSS. We performed retrospective testing on images from stroke patients by censoring the known TSS and comparing performance to published results using DWI-FLAIR mismatch.

2 Related Work

“DWI-FLAIR mismatch” is defined as the presence of visible acute ischemic lesion on DWI with no traceable hyper-intensity in the corresponding region on FLAIR imaging (Figure 1) [7]. The work of using DWI-FLAIR mismatch
was first introduced by Thomalla et al. [10], in which they used the mismatch pattern to identify stroke patients with less than 3-hour stroke onset. The method achieved a high specificity of 0.93 and a high positive predictive value (PPV) of 0.94, with a low sensitivity of 0.48 and a low negative predictive value (NPV) of 0.43. Aoki et al. [19] and Petkova et al. [20] followed the same method and applied on their datasets. Both achieved a high sensitivity (0.83 and 0.90 respectively) and a high specificity (0.71 and 0.93 respectively), but Aoki et al. reported a moderate PPV of 0.64.

Work has also been done in using DWI-FLAIR mismatch to classify TSS<4.5hrs, which is the current clinical cutoff time for IV tPA treatment. Ebinger et al. [21] developed a mismatch model and it achieved a specificity of 0.79 and a sensitivity of 0.46. Later, a large multicenter study was done by Thomalla et al. [22] to assess the ability of DWI-FLAIR mismatch. The mismatch method achieved a specificity of 0.78 and a PPV of 0.83, with a sensitivity of 0.62 and a NPV of 0.54. The study interobserver agreement of acute ischemic lesion visibility on FLAIR imaging was moderate (kappa = 0.569). Emeriau et al. [23] also investigated the use of mismatch pattern and the model achieved a PPV of 0.88, but a sensitivity of 0.55, a specificity of 0.60, and an NPV of 0.19. The AUC of using mismatch patterns in the identification of TSS was 0.58. There are ongoing large multicenter clinical trials, such as the WAKE-UP trial in the European Union [8] and the MR WITNESS trial in the United States [24], to further investigate the use of DWI-FLAIR mismatch in guiding treatment decisions for patients with unknown TSS.

The above preliminary work using DWI-FLAIR mismatch demonstrates a potential opportunity for using image analysis to classify TSS. However, existing studies all suffer from the use of relatively simplistic features and models [10–12]. Furthermore, it has been proposed that DWI-FLAIR mismatch may be too stringent, and therefore miss individuals who could benefit from thrombolytic therapy [25]. In this work, we develop machine learning models to classify acute ischemic stroke patient TSS using MR imaging features. We proposed a deep learning model, which is based on an autoencoder architecture [18], to extract latent representative imaging features (deep features) from PWIs. We compared the performance of various models (stepwise multilinear regression, support vector machines, random forest, and gradient boosted regression tree) to classify TSS with and without the deep features, and determined the best model for classifying TSS. We also provided a visualization strategy to interpret the deep features, and correlate them to the input images.

3 Methods

3.1 Dataset

In a study approved by the UCLA institutional review board (IRB), clinical stroke data was transferred from our institution’s data repository into a REDCap [26] database. The database holds 1,059 acute stroke patients from 1992 to 2016 who have received at least one or more of the following revascularization treatments: IV tPA, IA tPA, or mechanical thrombectomy. The corresponding patient pre-treatment MR PWIs, apparent diffusion coefficient (ADC) images, DWIs and FLAIR images were obtained from the UCLA Medical Center picture archiving and communication system (PACS).

For this study, we define the following inclusion criteria: 1) patients must experience acute ischemic stroke due to middle cerebral artery (MCA) occlusion; 2) patients must have a recorded time for which the stroke symptoms are first observed; 3) patients must have a recorded time for which the first imaging is obtained before treatment; and 4) patients must have a complete imaging set of PWIs, DWIs, FLAIRs, and ADCs. Patients’ TSS was calculated by
subtracting the time at which the stroke symptoms were first observed from the time at which the first imaging was obtained. We followed the existing DWI-FLAIR TSS classification task [23] to binarize the TSS into two classes: positive (1; <4.5hrs) and negative (0; ≥4.5hrs). After applying the inclusion criteria, 105 patients were obtained (83 positive class; 22 negative class). The patient characteristics are summarized in Table 1. This cohort subset was used to build the models for TSS classification.

### 3.2 Image Preprocessing

Intra-patient registration of pre-treatment PWI, DWI, ADC and FLAIR images was performed with a six degree of freedom rigid transformation using FMRIB’s Linear Image Registration Tool (FLIRT) [27]. Gaussian filters were applied to remove spatial noise and a multi-atlas skull-stripping algorithm [28] was used to remove skulls. Different tissue type masks (e.g., cerebrospinal fluid (CSF), gray/white matter) were identified using Statistical Parametric Mapping 12 (SPM12) [29] and CSF was excluded from this analysis. The sparse perfusion deconvolution toolbox (SPD) [30] and the ASIST-Japan Perfusion mismatch analyzer (PMA) were used to perform perfusion parameter map generation and arterial input function (AIF) identification (see Section 3.3.2.2).

### 3.3 Feature Generation

#### 3.3.1 Baseline MR imaging feature

PWIs are spatio-temporal imaging data (4D) that show the flow of a gadolinium-based contrast bolus into and out of the brain over time. They contain concentration time curves (CTCs) for each brain voxel that describe the flow of the contrast (i.e. signal intensity change) over time. Perfusion parameter maps [31] can be derived from PWIs that describe the tissue perfusion characteristics, including cerebral blood volume (CBV), cerebral blood flow (CBF), mean transit time (MTT), time-to-peak (TTP), and time-to-maximum (Tmax). Briefly, CBV describes the total volume of flowing blood in a given volume of a voxel and CBF describes the rate of blood delivery to the brain tissue within a volume of a voxel. By the Central Volume Theorem, CBV and CBF can be used to derive MTT, which represents the average time it takes the contrast to travel through the tissue volume of a voxel. TTP is the time required for the CTC to reach its maximum, which approximates the time needed for the bolus to arrive at the voxel with delay caused by brain vessel structure. Tmax is the time point where the contrast residue function reaches its maximum, which approximates the true time needed for the bolus to arrive at the voxel.

Intensity features (e.g., DWI voxel intensity, CBF voxel value) are often generated for voxel-wise stroke tissue outcome prediction [32]. Yet, generating intensity features based on entire brain MR images may be less descriptive to the stroke pathophysiology and less predictive of TSS because often stroke occurs in only one cerebral hemisphere. Therefore, we generated the imaging features only within regions that have Tmax>6s [33], which capture both the dead tissue core and the salvageable tissue that can possibly be saved by treatments. Feature generation involves two steps: 1) perfusion parameter maps were calculated using the SPD toolbox [30], and the region of interest was defined by Tmax>6s; and 2) the average intensity value was calculated within the region of interest for each image (DWI, ADC, FLAIR, CBF, CBV, TTP, and MTT), resulting in a set of data with seven intensity features. All the features were then standardized independently to zero mean with a standard deviation of 1. These baseline imaging features were used in building the classifiers for TSS classification.

### Table 1. Acute Ischemic Stroke patient sub-cohort characteristics.

<table>
<thead>
<tr>
<th>Demographics</th>
<th>Patients (n = 105)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>69.6±21.0</td>
</tr>
<tr>
<td>Gender</td>
<td>43 males</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Clinical Presentation</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Time since stroke (continuous)</td>
<td>158±108 mins</td>
</tr>
<tr>
<td>NIHSS†</td>
<td>12.9±8.08</td>
</tr>
<tr>
<td>Atrial fibrillation</td>
<td>1 (28); 0 (77)</td>
</tr>
<tr>
<td>Hypertension</td>
<td>1 (62); 0 (43)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Prediction label</th>
<th>Time since stroke (binary)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Time since stroke (binary)</td>
</tr>
<tr>
<td></td>
<td>≤4.5hrs (83); ≥4.5hrs (22)</td>
</tr>
</tbody>
</table>

† NIHSS = NIH Stroke Scale International; scale: 0 (no stroke symptoms) - 42 (severe stroke)
3.3.2 Deep feature Generation

3.3.2.1 Deep Autoencoder (AE)

We hypothesized that a deep learning approach can automatically learn feature detectors to extract latent features from PWIs that can improve TSS classification. We therefore implemented a deep autoencoder (deep AE) that is based on a stacked autoencoder [18] to learn the hidden features from PWIs (Figure 2). Each PWI voxel CTC, with a size of 1 x t (t = time for perfusion imaging), is transformed by the deep AE into k new feature representations that can represent complex voxel perfusion characteristics. The learning of these features is automatic and it is achieved by the hierarchical feature detectors, which are sets of weights that are learned in training via backpropagation. The deep AE consists of an encoder and a decoder. The encoder consists of two components: 1) an input layer; and 2) fully-connected layers, in which input neurons are fully-connected to each previous layer’s output neuron. The encoder is connected to the decoder, which follows reversely the same layer patterns of the encoder. The encoder output (i.e., the middle layer output of the deep AE) is the k new feature representations that can be used for TSS classification (in this work, k = 4).

The proposed network is trained via an unsupervised learning procedure in which the decoder output is the reconstruction of the encoder input. The network is optimized to obtain weights, Θ, that minimize the binary cross entropy loss between the input, I, and the reconstructed output, I(Θ), across the samples with size n [34]:

\[
\text{arg min}_\theta \text{ loss} = \frac{1}{n} \sum_{i=1}^{n} \left[ (I_i \ast \log(I(\Theta)) + (1 - I_i) \ast \log(1 - I(\Theta)) \right]
\]  
(Eq. 1)

3.3.2.2 Training data generation

As previous work suggests [35], regional information corresponding to a voxel’s surroundings can improve classification in MR images. Therefore, a small region was included in each training voxel, leading to a size of 3 x 3 x t patch (width x height x time; the z-dimension is omitted; t = 64 in our dataset), where the center of the patch is the voxel of interest for the deep AE feature learning. Each training patch was also coupled with its corresponding arterial input function patch [36], which describes the contrast agent input to the tissue in a single voxel, to improve the learning of hidden features. Each training patch was unrolled into a 1D vector, leading to a size of 1152x1. The 1D data were used to train the deep AE. In total, 105,000 training data were generated by sampling randomly and equally from all the patient PWIs.

3.3.2.3 Deep AE Configuration and Implementation

We observed that standard batch gradient descent did not lead to a good convergence of the deep AE during training. We suspect that this may be due to an inappropriate learning rate (default: 0.01), which typically requires careful
tuning. Therefore, we optimized the deep AE using Adam, which computes adaptive learning rates during training and has demonstrated superior performance over other methods [37]. An early-stopping strategy was applied to improve the learning of deep AE weights and prevent overfitting, where the training would be terminated if the performance did not improve over five consecutive epochs (max number of training epochs: 50). The deep AE was implemented in Torch7 [34], and the training was done on two NVIDIA Titan X GPUs and an NVIDIA Tesla K40 GPU. We explored different architectures of the deep AE, including different numbers of encoder hidden layers (from 1-3) and different numbers of hidden units (factor of 2, 4, and 6). Ten-fold patient-based cross-validation was performed to determine the optimal architecture for the deep AE (with an input size of 1152×1 and k = 4 new feature representations). Once the deep AE was trained, we used it to learn four new AE feature maps from each patients’ PWIs by aggregating the deep AE encoder output of all voxels. Then, average intensity values from AE feature maps (denoted as deep features) were generated in the regions of interest following the same procedure as described in Section 3.2.1.

3.4 Machine Learning models for TSS classification

We constructed and compared the performance of four machine learning methods for TSS classification: stepwise multilinear regression (SMR), support vector machine (SVM), random forest (RF), and gradient boosted regression tree (GBRT). Briefly, SMR is a stepwise method for adding and removing features from a multilinear model based on their statistical significance (e.g., F-statistics) to improve model performance [38]. SVM is a supervised learning classification algorithm that constructs a hyperplane (or set of hyperplanes) in a higher dimensional space for classification [39]. RF is an ensemble learning method in which a multitude of decision trees are randomly constructed and the classification is based on the mode of the classes output by individual trees [40]. GBRT is an ensemble learning method similar to RF, in which a multitude of decision trees are randomly generated, yet these trees are added to the model in a stage-wise fashion based on their contribution to the objective function optimization [41].

Different machine learning methods may not perform equally on the same feature set. Also, different model hyperparameter (e.g., a SVM’s hyperparameter, C) contribute differently to the classification. Evaluating model performance without hyperparameter tuning may lead to decreased predictive power due to over-fitting, especially on small and imbalanced datasets. Therefore, we performed leave-one-patient-out validation for evaluation, with a nested cross-validation for tuning model hyperparameters, following the proposed method [42]. A feature selection method, stability selection [43], was also applied to select the optimal feature subset before cross-validation to determine the best features for modeling (except for SMR because it has built-in feature selection method). This feature selection method produces a fairer feature comparison by aggregating different feature selection results from random subsampling of data and feature subsets. An overview of steps is shown in Figure 3. The SVM and RF were developed using the Python Scikit-learn library [44]. The SMR and GBRT were developed using MATLAB and the XGBoost library [45] respectively.

Figure 3. An overview of steps to predict TSS (<4.5 hrs). The SPD toolbox was used to generate perfusion parameter maps (e.g., Tmax) from the PWIs. A region of interest mask was defined on Tmax>6s region. Then, the mask was applied to the perfusion maps and MR images (DWI, ADC, and FLAIR) to generate average intensity values. A total of seven baseline average intensity features were used to train the classifiers to predict TSS<4.5hrs. Classifier performances (with and without the addition of deep features) were compared.
4 Results and Discussion

4.1 Deep AE training

The optimal model architecture for the proposed deep AE is 1152-192-4-4-192-1152, with an average mean square error (MSE) of $0.675 \pm 0.246$ (average deep AEs MSE is $1.29 \pm 0.755$). The small MSE indicates the reconstruction of input signal is efficient with the encoder-decoder structure, and the encoder output is a compact representation for the input. Figure 4 shows the MSE along epoch of each fold for the optimal deep AE. All the models across folds converged within first ~10 epochs, with minimal weight adjustments in the following epochs as indicated by small changes in the MSE. Most of the models were stopped within 25 epochs (except the model in fold 9).

![MSE vs. epoch for each fold of the validation for the optimal deep AE (in logarithmic scale for better visualization). All the models across folds converged within ~10 epochs, with minimal adjustments in the following epochs as indicated by small MSE change.](image)

4.2 TSS classification

Leave-one-patient-out validation (see Section 3.3) was performed to evaluate each classifier. The model performance was measured via AUC and model bias via F1-score (Table 2). Youden’s index [46] was used to determine optimal receiver operating characteristic curve (ROC) cutoff points, which were used to calculate the F1-score, sensitivity, specificity, true predictive value (TPV), and negative predictive value (NPV) (Table 3).

**Table 2.** The AUC and F1-score of different classifiers on predicting TSS with baseline imaging features and with/without deep features. B (baseline features), B+AE (baseline and deep features).

<table>
<thead>
<tr>
<th>Models</th>
<th>AUC</th>
<th>B</th>
<th>B+AE</th>
<th>F1-score</th>
<th>B</th>
<th>B+AE</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) SMR</td>
<td>0.570</td>
<td>0.683</td>
<td>0.608</td>
<td>0.765</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(2) SVM</td>
<td>0.470</td>
<td>0.640</td>
<td>0.632</td>
<td>0.859</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(3) RF</td>
<td>0.529</td>
<td>0.651</td>
<td>0.847</td>
<td>0.818</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(4) GBRT</td>
<td>0.526</td>
<td>0.623</td>
<td>0.862</td>
<td>0.681</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 3.** The sensitivity, specificity, TPV, and NPV of different classifiers on predicting TSS with baseline imaging features and with/without deep features. B (baseline features), B+AE (baseline and deep features).

<table>
<thead>
<tr>
<th>Models</th>
<th>sensitivity</th>
<th>specificity</th>
<th>TPV</th>
<th>NPV</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) SMR</td>
<td>0.458</td>
<td>0.687</td>
<td>0.818</td>
<td>0.905</td>
</tr>
<tr>
<td>(2) SVM</td>
<td>0.506</td>
<td>0.880</td>
<td>0.636</td>
<td>0.364</td>
</tr>
<tr>
<td>(3) RF</td>
<td>0.857</td>
<td>0.750</td>
<td>0.333</td>
<td>0.667</td>
</tr>
<tr>
<td>(4) GBRT</td>
<td>0.893</td>
<td>0.560</td>
<td>0.286</td>
<td>0.667</td>
</tr>
</tbody>
</table>
With the additional deep features, all classifiers (SMR, SVM, RF, and GBRT) showed improvement in AUC. The increase (~10%) of AUC after adding the deep features demonstrated the usefulness of these deep features and their association with the TSS. Figure 5 shows the ROCs of the classifiers trained with the baseline features and the deep features, with a reference ROC based on the DWI-FLAIR mismatch method is shown for comparison [23]. All the classifiers generally performed better with the addition of the deep features. These classifiers performed better than AUC=0.60 (as compared to reference AUC of 0.58), demonstrating the ability of using imaging features with machine learning models to classify TSS. Among all the classifiers, the SMR trained with baseline imaging features and the deep features performed the best, with an AUC of 0.683. Comparing to the reference mismatch method, SMR achieved higher sensitivity (0.69 vs 0.55) and NPV (0.33 vs 0.19) while maintaining similar specificity (0.59 vs 0.60) and TPV (0.86 vs 0.88). Therefore, SMR was determined to be the most suitable classifier for TSS classification.

We observed that the models trained with only the baseline imaging features had low performance. The best model was the SMR with an AUC of 0.570, and other models had lower AUCs (<0.55). This may be due to the insufficient baseline features for classifier construction (i.e., only mean intensity features across MR images and perfusion maps were used). In future work, feature generation techniques, such as descriptive statistics, will be investigated to generate more features for TSS classification. Although these models were less predictive, models trained with additional deep features showed significant improvement in TSS classification. This supports our hypothesis that PWIs contain information encoding TSS, and that the proposed deep AE extracted hidden features in PWIs are predictive of TSS.

4.3 Deep Feature Visualization

To understand what the deep AE learned to extract in the encoder output layer, we applied the visualization technique, top-m selection [47], on the encoder output layer and correlated the results to the four deep AE activation maps (denoted as ae1, ae2, ae3, and ae4). Briefly, every input CTC will cause a hidden neuron unit to output a value (i.e., activation). Some input CTCs will give a higher activation while other inputs will give a lower activation. High activation values indicate the presence of relevant features in the input (e.g., wide and high peak) that can “excite” the hidden units [47]. For each hidden unit of the optimal deep AE encode layer, we performed top-m selection to obtain the top m signals ($x^*$) from the training data that cause the most activations:

$$x^* = \{x_j\}, j \in \text{first m of Sort}(h_i(\theta, x)),$$

(Eq. 2)

where $\text{Sort}(\cdot)$ is a descending operation, and $h_i(\theta, x)$ is the activation of the $i$th hidden unit. Figure 6 shows the four deep AE activation maps and their corresponding average curves of the top-50 (i.e., $m=50$) input CTCs. Our results show that different hidden units appear to capture different type of input signals. For example, the ae1 map shows higher activations (brighter) in the acute stroke region (left brain; high Tmax), and the corresponding average top-50 CTC (blue curve) has delayed and low concentration, which matches the visualization (that is, the hidden unit detects...
the stroke-affected CTCs). In contrast, the ae3 map shows low activations (darker) in the acute stroke region and the corresponding average top-50 CTC (red curve) has early sharp and high peak. We also calculated the Pearson correlation coefficient between the deep features generated from these activation maps and TSS. The ae3 deep feature showed statistically significant correlation with TSS (p-value<0.05). These visualization and correlation results demonstrated that the learned features from the optimal deep AE contained information that was predictive of TSS.

Understanding deep learning representations is challenging because it requires making sense of non-linear computations performed over many network weights [48]. Our visualization result is the first step to attempt to understand what the deep AE is learning. This is important for using deep learning model in medical image analysis because deep learning is often a “black-box” approach that yields superior, but hard-to-interpret results. However, the visualization result in this work is not conclusive. Further research is required to understand what the deep network is learning. One next step we plan to pursue is to apply different visualization techniques (e.g., deconvolution [49]) on the learned networks and perform statistical tests to draw correlations between them to an observation (e.g., small TSS).

5 Conclusion and Future work

In this paper, we showed that SMR, SVM, RF, and GBRT models were able to classify TSS, with SMR achieving the highest AUC. We proposed a deep AE architecture to extract representative features from PWIs and showed that adding deep features boosted the classifiers’ performance, showing the potential application of deep learning feature extraction techniques in TSS classification. In addition, we utilized a visualization method to interpret the features learned in the deep AE and discussed the possible research opportunity in understanding deep learning models for medical images.

We now discuss a few limitations and areas of future work. First, there are roughly 1,100 patients available in our dataset, but the majority of them were missing one or more MR images and were therefore not included in this analysis. Our next step will be looking into multimodal and denoising deep learning frameworks that are capable of handling missing data. Second, we will explore different feature generation techniques (e.g., descriptive and histogram statistics) to enlarge the feature set for training the classifiers. Third, we plan to explore several visualization techniques, such as deconvolution [49] and gated backpropagation [50] in order to understand the deep AE’s features and to draw both visual and statistical correlations to TSS classification. Finally, TSS<4.5hrs is the current clinical cutoff time for IV tPA treatment, yet this may not be an absolute time point in which a stroke patient can benefit from treatment because of changing brain pathophysiology [51]. We therefore plan to investigate and extend our models to more classes (e.g., a ±0.5hr boundary), rather than just TSS<4.5hrs/TSS≥4.5hrs.

6 Acknowledgements

This research was supported by National Institutes of Health (NIH) Grant R01 NS076534, UCLA Radiology Department Exploratory Research Grant 16-0003, and an NVIDIA Academic Hardware Grant.
References


Use of the Multidimensional Health Locus of Control to Predict Information-Seeking Behaviors and Health-Related Needs in Pregnant Women and Caregivers

Lauren E. Holroyd, BA1, Shilo Anders, PhD2, Jamie R. Robinson, MD1,2, Gretchen Purcell Jackson, MD, PhD2

1Vanderbilt University School of Medicine, Nashville, Tennessee; 2Vanderbilt University Medical Center, Nashville, Tennessee

Abstract

Pregnancy produces important health-related needs, and expectant families have turned to technologies to meet them. The ability to predict needs and technology preferences might aid in connecting families with resources. This study examined the relationships among Multidimensional Health Locus of Control (MHLC) scores, information-seeking behaviors, and health-related needs in 71 pregnant women and 29 caregivers. Internal MHLC scores were positively correlated with information-seeking behaviors, including website and patient portal use. Higher Chance scores were associated with decreased portal or pregnancy website use (p=0.002), with the exception of FitPregnancy.com (p=0.02). MHLC scores were not significantly correlated with number of health-related needs or whether needs were met. Individuals with needs about disease management had higher Powerful Others scores (p=0.01); those with questions about tests had lower Powerful Others scores (p=0.008). MHLC scores might be used to identify individuals less likely to seek information and to predict need types.

Introduction

Improving maternal and infant health are key priorities of many national and international women’s health initiatives1,2,3. There are nearly 6.5 million pregnancies and 4 million births each year in the United States4, and the Agency for Healthcare Research and Quality’s Healthcare Cost and Utilization Project reported up to 94% of births are complicated by maternal or fetal conditions5. Pregnancy is a time in which women not only have many questions pertaining to their own health but also distinct health concerns about their unborn children. In the last two decades, infant and maternal mortality rates have decreased, but increases in maternal age, obesity, opioid abuse, use of assisted reproductive technologies, and chronic health conditions have resulted in a doubling of severe maternal morbidity6. Infant mortality and prematurity, with their associated lifelong health problems, remain important problems in minority populations7. Studies of first time parents have shown the transition to parenthood, including the timeframe from pregnancy to postpartum, to be one of the most stressful periods of adult development. During this time, there are increased social and emotional needs for both mothers and fathers8. All pregnancies, even uncomplicated, healthy pregnancies, may raise significant health-related questions for all individuals involved in the care of the mother or child. The role and concerns of extended family and additional caregivers are becoming increasingly important as 27% of children live with a single parent and 10% live with a grandparent9.

There has been little research about the specific information needs of pregnant women and their caregivers, whether these needs are met, and what populations are at risk for unmet needs. Studies that have been completed are mostly small and restricted usually focused medical questions related to genetic risks and prenatal diagnoses10,11,12,13. The organization and educational content of prenatal visits have not adapted to the increasingly complex population of pregnant women and their family structures. Many women have turned to technologies such as smartphone apps and the Internet to address their unmet needs during pregnancy14,15,16. An easy method connect expectant families with appropriate health information technologies might ultimately lead to improved maternal and fetal outcomes.

Prior research has shown that healthcare consumers have strong technology usage preferences, which are not reliably predicted by demographic characteristics17. Tools to predict health-related needs and technology preferences would be useful in tailoring or recommending resources to specific populations. The Multidimensional Health Locus of Control (MHLC) is a potential target. The MHLC is a survey, based on social learning theory and designed to measure beliefs surrounding sources of health-related behaviors. The survey contains 18-items with 6-items for each of its three subscales: Internal, Chance and Powerful Others. Each item is scored on a 6-point, Likert-type, scale from 1 (Strongly Disagree) to 6 (Strongly Agree). Scores for each subscale reflect the total score for the corresponding 6 items, with subscales having a minimum score of 6 and maximum score of 3618.
The Internal subscale measures the strength of the belief that the respondent’s health is the result of their own behaviors. Higher scores on the Internal subscale indicate a more Internal Health Locus of Control. Lower scores on the Internal subscale indicate a more External Health Locus of Control or stronger belief that the respondent’s health is determined by factors aside from themselves. The Powerful Others subscale measures the degree to which the respondent believes health is controlled by individuals seen as “powerful others”, such as clinicians or a higher power. Higher scores on the Powerful Others subscale indicate a stronger belief in external control of health by such influential forces. The Chance subscale indicates the degree to which the respondent believes that health is out of his or her control. Higher scores on both the Powerful Others and Chance subscales represent a more External Health Locus of Control while lower scores on the Powerful Others and Chance subscales indicate a more Internal Health Locus of Control.

The MHLC has been used in numerous studies, primarily in the nursing and health psychology realm, to predict health behaviors and health status. In particular, the scale has been used to predict categories of patients who are likely to adhere to medication regimens or engage in preventative care. Observational studies examining medication adherence have consistently demonstrated External Health Locus of Control, higher scores on Powerful Others and Chance subscales, is predictive of medication non-adherence. Studies examining the relationship between MHLC and preventative or healthy lifestyle behaviors, in contrast, have produced conflicting results. Studies of pregnant women have shown that MHLC scores do not change significantly following delivery and that development of post-partum depression is significantly correlated with MHLC scores. The MHLC has been suggested as a potential screening tool for post-partum depression.

Additionally, several studies have provided evidence that MHLC can predict health-related needs and information-seeking behaviors. The relationship between Health Locus of Control and information-seeking behaviors was examined initially in 88 undergraduate students using a predecessor to the MHLC. This study showed that individuals with high Internal Health Locus of Control scores sought out more health-related information in the form of pamphlets. A small sociological study published in 2004 examined the relationship between MHLC and the desire for information in 81 mothers of children with disabilities. This study found that mothers with low scores on Chance and Internal subscales (i.e., those individuals with a low External Health Locus of Control) may be motivated to initiate an extended search for a cure to their child’s disability. However, few contemporary studies have examined the relationship between MHLC and information-seeking behaviors with new technologies such social media, the Internet, and smartphone based applications.

This study examined the relationships among MHLC scores, health-related needs, and information-seeking behaviors of pregnant women and their caregivers. We hypothesized that higher Internal Health Locus of Control scores would be predictive of increased information-seeking behaviors and that individuals with higher Internal Health Locus of Control scores would have more numerous and different categories of health-related needs than those with higher External Health Locus of Control scores.

**Methods**

**Overview**

This study employed data from a comprehensive study of health-related needs and information management practices of pregnant women and their caregivers. We describe the study population and data sets available from the larger study, and the specific methods of analyses employed for this project.

**Study Population**

Pregnant women and caregivers were recruited for the study from the Junior League Fetal Center (FCV) and Expect with Me (EWM) group prenatal program at Vanderbilt University Medical Center (VUMC). The FCV is an interdisciplinary advanced maternity care setting, which provides expert care in the diagnosis and treatment of fetal anomalies or pregnancy complications in a single location. EWM is a group prenatal care program, which provides traditional prenatal care, education, and peer support during pregnancy. Groups of 8-12 women of similar gestational ages meet 10 times throughout pregnancy. The group prenatal care model has been demonstrated to reduce preterm births and healthcare costs during and after pregnancy, and it is aligned with the Institute of Medicine’s Six Domains of Quality. These recruitment sites were selected to include participants experiencing both complicated and relatively normal pregnancies.

English- or Spanish-speaking adults, age >=18 years, with pregnancies less than 36 weeks gestational age, and addresses within 100 miles of VUMC were eligible for the study. Pregnant women enrolled in the study were allowed
to invite up to three caregivers to participate in the study. Caregivers were defined as any individual who would have significant involvement in the care of the mother or baby during and after pregnancy. Written informed consent was obtained from all participants. The research protocol was approved by the VUMC Institutional Review Board.

**Measures**

Enrolled participants completed multiple questionnaires, including a socio-demographics survey, technology usage assessment, and the MHLC; a semi-structured interview was completed during a research visit to VUMC. Participants were compensated with a $25 gift card for their time, and travel to VUMC was reimbursed.

The socio-demographics questionnaire included items to assess age, race, ethnicity, home location, and whether the participant had other children. Gestational age of the pregnancy was calculated based on the interview date and expected delivery date. Relationships between pregnant participants and caregivers were recorded. A technology usage questionnaire asked participants if they 1) had access to or had used Vanderbilt’s online patient portal, My Health at Vanderbilt (MHAV) and 2) had used online resources to obtain information about pregnancy, and if so, which resources (i.e., BabyCenter.com, TheBump.com, FitPregnancy.com, WhatToExpect.com, Pregnancy.com, Text4Baby.org, or other). Participants were able to indicate using multiple websites as sources of information.

Participants also completed a semi-structured interview with one to two members of the research team. Participants answered questions related to prior pregnancies, their current pregnancy, information-seeking behaviors, health-related needs, and biggest concerns. All interviews were audio-recorded, transcribed, and de-identified prior to analysis.

---

**Figure 1. Consumer Health-Related Needs Taxonomy**

Health-related needs were manually extracted from interview transcripts and classified by semantic type using a validated Consumer Health-Related Needs Taxonomy. Major categories from this taxonomy are depicted in Figure 1. Previous studies have utilized this taxonomy to categorize consumer health questions from patient journals and patient portal messages and have validated the taxonomy with inter-rater reliability of its application. The taxonomy divides health-related needs into five semantic categories: informational, medical, logistical, social, and other. Informational needs are questions that require clinical knowledge, such as prognosis of a particular diagnosis. Medical needs request medical care, such as the request for a medication change after a side effect. Logistical needs are questions regarding practical information, such as what facilities offer a particular treatment. The social category involves emotional or social needs. Lastly, the other category includes needs not currently included in the taxonomy schema or needs that span multiple categories, such as how to be a good parent. Identified needs were classified independently by at least
two research team members, and consensus was achieved by discussion for discrepancies. The results of this needs analysis and inter-rater reliability of the taxonomy application for this task are reported elsewhere.

In our analysis, we report descriptive statistics as appropriate on all measures, including scores on the MHLC. Regression analyses examined the relationships among MHLC scores, information-seeking behaviors, and health-related needs for all participants. In subsequent regression models, we controlled for whether the current pregnancy represented the participant’s first child, differences between the two models are reported in the results section text below. Significance values and corresponding test statistics (e.g. Pearson coefficients or logistic regression) are reported. All analyses were completed in SPSS version 24 (IBM, 2016).

Results

Study Population and Demographics

One hundred participants were enrolled in the study, and five were excluded from this analysis due to incomplete MHLC data. Thus, 67 pregnant women and 28 caregivers were included in the analysis. Table 1 shows participant characteristics. This population included 21 men and 74 women with a mean age of 29.3 years and average gestational age of the pregnancy of 30.8 weeks. Sixty-eight (71.5%) of pregnant women and their caregivers were white and 90 (94.7%) were English speakers. Seventy-nine (83.2%) pregnant women and their caregivers were recruited from FCV and 49 participants (52.5%) reported having other children.

Table 1. Demographic characteristics of participants.

<table>
<thead>
<tr>
<th></th>
<th>All Participants (n= 95), n (%)</th>
<th>Pregnant Women (n= 67), n (%)</th>
<th>Caregivers (n=28), n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>21 (22.1%)</td>
<td>0</td>
<td>21 (75.0%)</td>
</tr>
<tr>
<td>Female</td>
<td>74 (77.9%)</td>
<td>67 (100%)</td>
<td>7 (25.0%)</td>
</tr>
<tr>
<td>Race</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>68 (71.5%)</td>
<td>48 (71.6%)</td>
<td>20 (71.4%)</td>
</tr>
<tr>
<td>Black</td>
<td>18 (18.9%)</td>
<td>12 (17.9%)</td>
<td>6 (21.4%)</td>
</tr>
<tr>
<td>Native Hawaiian or Pacific Islander</td>
<td>1 (1.0%)</td>
<td>1 (1.5%)</td>
<td>0</td>
</tr>
<tr>
<td>American Indian/Alaskan</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Other</td>
<td>8 (8.4%)</td>
<td>6 (9.0%)</td>
<td>2 (7.1%)</td>
</tr>
<tr>
<td>Ethnicity</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-Hispanic</td>
<td>82 (86.3%)</td>
<td>57 (85.1%)</td>
<td>25 (89.3%)</td>
</tr>
<tr>
<td>Hispanic</td>
<td>13 (13.7%)</td>
<td>10 (14.9%)</td>
<td>3 (10.7%)</td>
</tr>
<tr>
<td>Language Spoken</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>English</td>
<td>90 (94.7%)</td>
<td>63 (94.0%)</td>
<td>27 (96.4%)</td>
</tr>
<tr>
<td>Spanish</td>
<td>5 (5.3%)</td>
<td>4 (6.0%)</td>
<td>1 (3.6%)</td>
</tr>
<tr>
<td>Recruitment Location</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fetal Center at Vanderbilt</td>
<td>79 (83.2%)</td>
<td>54 (80.6%)</td>
<td>25 (89.3%)</td>
</tr>
<tr>
<td>Expect with Me Group Prenatal Care</td>
<td>16 (16.8%)</td>
<td>13 (19.4%)</td>
<td>3 (10.7%)</td>
</tr>
<tr>
<td>Has Other Children?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>46 (48.4%)</td>
<td>34 (50.7%)</td>
<td>12 (42.9%)</td>
</tr>
<tr>
<td>Yes</td>
<td>49 (52.6%)</td>
<td>33 (49.3%)</td>
<td>16 (57.1%)</td>
</tr>
<tr>
<td>Age, years</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean (SD)</td>
<td>29.3 (8.6)</td>
<td>27.6 (6.2)</td>
<td>33.3 (11.6)</td>
</tr>
<tr>
<td>Estimated Gestational Age, weeks</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean (SD)</td>
<td>30.8 (4.0)</td>
<td>30.7 (4.1)</td>
<td>31.0 (3.9)</td>
</tr>
</tbody>
</table>

Multidimensional Health Locus of Control Scores

Figure 2 reports the mean MHLC scores calculated for all participants, pregnant women, and caregivers. Average scores and standard deviation (SD) for participants were Internal - 26.9 (4.2), Chance - 15.9 (5.0), and Powerful Others - 17.9 (4.5). Pregnant women’s scores were: I - 26.4 (4.0), Chance - 17.7 (4.7) and Powerful Others - 15.6 (4.5). Caregivers on average had higher scores on all three subscales: Internal - 27.9 (4.7), Chance - 16.8 (5.8), Powerful Others - 18.2 (4.6). Scores on the Internal subscale were found to be independent of scores on the Chance and Powerful Others subscales. Scores on the Chance and Powerful Others subscales were significantly positively

905
correlated with each other (p=0.004). Internal scores were negatively correlated with Chance scores and positively correlated with Powerful Others scores but these relationships were not significant (p=0.49 and p=0.47, respectively).

Table 2 summarizes the information-seeking behaviors of the participants. Of the 95 participants, fifty-four (54.7%) reported using the MHAV patient portal. Fifty (52.6%) reported using MHAV and websites to obtain information about their pregnancy. Thirty-five participants (39.8%) reported using only one of either websites or MHAV, and 10 individuals (10.5%) reported using neither MHAV nor websites to obtain information about their pregnancy.

In pregnant women and caregivers who reported using the Internet, the average number of websites participants accessed to obtain pregnancy related information was 2.0. BabyCenter.com was the most frequently accessed website by both pregnant women and caregivers. Forty-five participants (47.4%) reported accessing “other” websites to obtain pregnancy information. “Other” websites reported by participants included search engines (e.g., Google.com, WebMD.com, Yahoo.com), hospital specific websites (e.g., MayoClinic.com, ClevelandClinic.com), and pregnancy focused websites (e.g., EverydayFamily.com, JustMommies.com). Pregnant women were more likely than caregivers to access any of the specified resources to obtain information regarding pregnancy.

Table 3 shows the relationships between information-seeking behaviors and MHLC scores. Higher Chance scores were found to be associated with fewer information seeking behaviors (p=0.002). Neither Internal nor Powerful Others
scores were predictive of number of resources sought. These relationships remained constant when controlling for whether the current pregnancy was the participant’s first child. Regression analyses looking at specific websites accessed by pregnant women and their caregivers revealed that lower Chance scores were predictive of MHAV use, regardless of whether this was the participant’s first child (p=0.007, 0.02). Higher Chance scores were significantly associated with accessing FitPregnancy (p=0.02). Access of the other specified websites was not associated with MHLC scores.

Table 3. Relationships between MHLC scores and information-seeking behaviors.

<table>
<thead>
<tr>
<th></th>
<th>Internal Score Statistic (p-value)</th>
<th>Chance Score Statistic (p-value)</th>
<th>Powerful Others Score Statistic (p-value)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total # of Information-Seeking Behaviors</td>
<td>-0.04 (0.4)</td>
<td>-0.3 (0.002)*</td>
<td>-0.8 (0.2)</td>
</tr>
<tr>
<td># of Websites</td>
<td>0.1 (0.3)</td>
<td>0.6 (0.6)</td>
<td>-0.08 (0.5)</td>
</tr>
<tr>
<td>Specific Resources</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MHAV</td>
<td>-0.9 (0.4)</td>
<td>-2.8 (0.007)*</td>
<td>0.3 (0.7)</td>
</tr>
<tr>
<td>BabyCenter.com</td>
<td>0.1 (0.9)</td>
<td>-1.5 (0.1)</td>
<td>-1.2 (0.2)</td>
</tr>
<tr>
<td>TheBump.com</td>
<td>-0.5 (0.7)</td>
<td>0.6 (0.6)</td>
<td>-1.3 (0.2)</td>
</tr>
<tr>
<td>FitPregnancy.com</td>
<td>-0.7 (0.5)</td>
<td>2.4 (0.02)*</td>
<td>0.3 (0.8)</td>
</tr>
<tr>
<td>WhatToExpect.com</td>
<td>1.8 (0.08)</td>
<td>0.2 (0.9)</td>
<td>0.2 (0.8)</td>
</tr>
<tr>
<td>Pregnancy.com</td>
<td>1.2 (0.2)</td>
<td>1.3 (0.2)</td>
<td>1.7 (0.1)</td>
</tr>
<tr>
<td>Text4Baby.org</td>
<td>-0.4 (0.9)</td>
<td>-0.7 (0.5)</td>
<td>-0.7 (0.5)</td>
</tr>
<tr>
<td>Other</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* signifies p≤0.05

Health-Related Needs

A total of 1014 health-related needs were reported by 95 participants. On average, participants reported 10.7 health-related needs, 11.8 for pregnant women and 7.9 for caregivers. Of these needs, 672 (66.3%) were classified as informational, with 305 being about problems (i.e., diseases or conditions). Informational needs regarding interventions were more often reported by pregnant women compared to caregivers (21.3% vs. 10.4%) while informational needs regarding management of problems were more often reported by caregivers than pregnant women (14.8 vs. 16.7%). Of the remaining health-related needs identified, 87 (8.6%) were classified as medical, 162 (16.0%) logistical, 90 (8.9%) social, and 3 (0.3%) other.

Of the 1014 needs identified, half (50.5%) were partially met while approximately one-fourth (26.8%) were unmet. Caregivers reported fewer met needs than pregnant women (16.2% vs 24.5%). Additional details regarding health-related needs of pregnant women and caregivers can be seen in Table 4.

Table 4. Health-related needs of pregnant women and their caregivers

<table>
<thead>
<tr>
<th></th>
<th>All Participants (n=95)</th>
<th>Pregnant Women (n=67)</th>
<th>Caregivers (n=28)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Needs (n=1014), n (%)</td>
<td>Needs (n=792), n (%)</td>
<td>Needs (n=222), n (%)</td>
</tr>
<tr>
<td>Informational</td>
<td>672 (66.3%)</td>
<td>530 (66.9%)</td>
<td>142 (64.0%)</td>
</tr>
<tr>
<td>Normal</td>
<td>1 (0.1%)</td>
<td>1 (0.1%)</td>
<td>0</td>
</tr>
<tr>
<td>Problem</td>
<td>305 (30.1%)</td>
<td>225 (28.4%)</td>
<td>80 (36.0%)</td>
</tr>
<tr>
<td>Management</td>
<td>154 (15.2%)</td>
<td>117 (14.8%)</td>
<td>37 (16.7%)</td>
</tr>
<tr>
<td>Tests</td>
<td>20 (2.0%)</td>
<td>18 (2.3%)</td>
<td>2 (0.9%)</td>
</tr>
<tr>
<td>Interventions</td>
<td>192 (18.9%)</td>
<td>169 (21.3%)</td>
<td>23 (10.4%)</td>
</tr>
<tr>
<td>Medical</td>
<td>87 (8.6%)</td>
<td>72 (9.1%)</td>
<td>15 (6.8%)</td>
</tr>
<tr>
<td>Logistical</td>
<td>162 (16.0%)</td>
<td>116 (14.6%)</td>
<td>46 (20.7%)</td>
</tr>
<tr>
<td>Social</td>
<td>90 (8.9%)</td>
<td>72 (9.1%)</td>
<td>18 (8.1%)</td>
</tr>
<tr>
<td>Other</td>
<td>3 (0.3%)</td>
<td>2 (0.2%)</td>
<td>1 (0.5%)</td>
</tr>
<tr>
<td>Status</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unmet</td>
<td>272 (26.8%)</td>
<td>181 (22.9%)</td>
<td>91 (41.0%)</td>
</tr>
<tr>
<td>Partially Met</td>
<td>512 (50.5%)</td>
<td>417 (52.7%)</td>
<td>95 (42.8%)</td>
</tr>
<tr>
<td>Met</td>
<td>230 (22.7%)</td>
<td>194 (24.5%)</td>
<td>36 (16.2%)</td>
</tr>
</tbody>
</table>

Relationships between MHLC scores and the total number of health-related needs are shown in Table 5. MHLC scores were not significantly correlated with the total number of health-related needs. Internal score was positively correlated...
with number of logistical needs, and this relationship approached significance (p=0.06). In modeling relationships between MHLC scores and subcategories of health-related needs, the Powerful Others score was associated with management and test categories of informational needs. Informational management needs were positively correlated with Powerful Others score (p=0.01) and informational test needs were negatively correlated with Powerful Others score (p=0.008). When controlling for whether this pregnancy was the first child for the participants, Powerful Others score were not associated with informational management needs. There was no significant relationship between MHLC scores and the whether health-related needs were met.

Table 5. Relationships between MHLC scores and health-related needs.

<table>
<thead>
<tr>
<th>Category of Health-Related Need</th>
<th>Internal Score Statistic (p-value)</th>
<th>Chance Score Statistic (p-value)</th>
<th>Powerful Others Score Statistic (p-value)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total # of Health-Related Needs</td>
<td>0.1 (0.2)</td>
<td>-0.3 (0.8)</td>
<td>-0.006 (0.9)</td>
</tr>
<tr>
<td>Informational</td>
<td>-0.2 (0.8)</td>
<td>-0.4 (0.7)</td>
<td>-0.2 (0.8)</td>
</tr>
<tr>
<td>Normal</td>
<td>-0.7 (0.5)</td>
<td>-0.01 (1.0)</td>
<td>-0.8 (0.4)</td>
</tr>
<tr>
<td>Problem</td>
<td>0.7 (0.5)</td>
<td>0.6 (0.6)</td>
<td>-1.0 (0.3)</td>
</tr>
<tr>
<td>Management</td>
<td>-0.3 (0.8)</td>
<td>-0.03 (1.0)</td>
<td>2.5 (0.01)*</td>
</tr>
<tr>
<td>Tests</td>
<td>-0.4 (0.7)</td>
<td>-0.5(0.6)</td>
<td>-2.7 (0.008)*</td>
</tr>
<tr>
<td>Interventions</td>
<td>-0.6 (0.5)</td>
<td>-1.0 (0.3)</td>
<td>-0.4 (0.7)</td>
</tr>
<tr>
<td>Medical</td>
<td>0.5 (0.6)</td>
<td>-1.1 (0.3)</td>
<td>-1.5 (0.1)</td>
</tr>
<tr>
<td>Logistical</td>
<td>1.9 (0.06)*</td>
<td>-0.09 (0.9)</td>
<td>1.2 (0.2)</td>
</tr>
<tr>
<td>Social</td>
<td>-0.9 (0.4)</td>
<td>1.2 (0.3)</td>
<td>1.5 (0.1)</td>
</tr>
<tr>
<td>Other</td>
<td>-0.1 (0.9)</td>
<td>-0.08 (0.9)</td>
<td>1.1 (0.3)</td>
</tr>
<tr>
<td>Status</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Met</td>
<td>-0.05 (0.6)</td>
<td>-0.09 (0.4)</td>
<td>-0.1(0.3)</td>
</tr>
<tr>
<td>Partially Met</td>
<td>0.1 (0.2)</td>
<td>-0.03 (0.8)</td>
<td>0.1 (0.4)</td>
</tr>
<tr>
<td>Unmet</td>
<td>0.1 (0.2)</td>
<td>0.06 (0.6)</td>
<td>-0.04 (0.7)</td>
</tr>
</tbody>
</table>

* signifies p≤0.05, ^ p≤0.06

Discussion

This study explored the distribution of MHLC scores in both pregnant women and their caregivers and the relationships among MHLC scores, health-related needs, and information seeking-behaviors in this population. We found that pregnant women and caregivers had MHLC scores consistent with published norms. Wallston and Wallston reported normative averages for each MHLC subscale following survey administration to college students in multiple studies: Internal - 26.6; Powerful Others - 17.87; Chance - 16.72. These initial norms have been validated by contemporary data sets and in pregnant women. A 2013 survey of 230 pregnant Iranian women showed means and standard deviations for Internal - 25.48 to 27.2, (SD: 3.39-3.42), Powerful Others - 26.73 to 22.83 (SD: 4.42-4.79), and Chance - 19.03 to 19.2 (SD: 5.71-5.90). These studies show that scores on Chance and Powerful Others subscales are similar in magnitude, while Internal scores are consistently higher, which is congruent with our findings. These studies demonstrate that individuals often have a more Internal than External Health Locus of Control.

Relationships among the three subscales were unique to our population. Prior studies have shown Internal and Powerful Others subscales were independent of one another while Internal and Chance subscales were significantly, negatively correlated. In our population, Powerful Others and Chance scores were significantly, positively correlated; however, there was no significant association between Internal scores and Chance or Powerful Others scores. Differences in these relationships could be population specific. Prior studies have performed analyses on healthy, undergraduate students and not adult patients currently experiencing the stress and unpredictable nature of pregnancy. Our study participants might have been more likely to view health as a result of multiple sources.

In examining information-seeking behaviors, the majority of pregnant women and caregivers used both MHAV and websites to obtain information about pregnancy. Most individuals who reported using the Internet as an information source visited multiple websites related to pregnancy. This finding confirms prior research showing that health information technologies have become common resources for pregnant families and further supports initiatives and additional research into what technologies, including mobile phone applications, websites, or interactive technologies, would be most effective during pregnancy. Health information technology design research should focus on preferences for technology usage and the unique health-related needs of pregnant women and their caregivers. In this study of
pregnant women and caregivers, our findings were not consistent with prior work that found Internal Health Locus of Control to be positively correlated with information-seeking behaviors. Previous studies examining these relationships were completed using either healthy volunteers or mothers of children with disabilities. In both studies, the participants were not themselves experiencing a health condition. In contrast, our sample was composed of a majority of pregnant women and caregivers going through complicated pregnancies. The correlation of Internal scores with information seeking behaviors then might be stronger in individuals who are healthy and weaker in those intimately connected to the illness, such as a mother pregnant with a fetal anomaly.

Our study demonstrated that feeling like health is a result of fate or luck (i.e., higher Chance scores) was associated with reduced use of resources to obtain health-related information. In particular, individuals with higher Chance scores were less likely to report using MHAV, Vanderbilt’s patient portal. This was true regardless of whether the current pregnancy was the participant’s first child or not. Utilization of patient portals or Internet resources is an active process that provides patients the opportunity to take ownership of their health and participate actively in their care. It would follow that individuals who feel their actions have no impact on their health status would be not be as likely to engage in such activities. However, active participation in care has been shown improve health outcomes. The MHLC may offer an alternative way to identify individuals not likely to actively seek information to maintain or improve their health. Future research should focus on the best strategies to engage these individuals in health-related activities. Participants with higher Chance scores were likely to access FitPregnancy.com, a website which compiles short articles on various pregnancy topics, more than other websites. This finding may offer design guidance for developers of health information technologies to engage a population that may not be inclined take ownership over their healthcare. The brevity of articles and entertainment-news style of this site may have been less intimidating than the more formal style of sites such as BabyCenter.com or WhatToExpect.com and more approachable than services that require sign up, such as Text4Baby.org.

We found that individuals with health-related needs regarding the clinical management of pregnancy problems were significantly more likely to feel that their health was the result of the actions of others (i.e., had higher Powerful Others scores). This significant relationship remained even after account for whether the participant had other children. The most common clinical management questions included inquiries into the options available to treat a particular diagnosis or to manage conditions of pregnancy in general. Additionally, participants without other children who expressed questions regarding clinical tests were significantly less likely to feel their health was in the hands of others (i.e., had lower Powerful Others scores). This finding suggests that pregnant women or caregivers with low Powerful Others scores may seek a better understanding of their pregnancy-related testing because they feel expert healthcare providers are fully in control of the maternal or fetal outcomes. This difference between participants with children and those expecting their first child could potentially be explained by increased trust in prenatal testing after having a successful pregnancy.

This study has important limitations. This research was completed at a single, large, academic medical center using a patient population where a majority of pregnancies were complicated by either maternal or fetal diagnoses. Further, the nature and structure of care provided at FCV and EWM group prenatal care resulted in participants having more frequent interactions with the healthcare system during pregnancy than traditional prenatal care models. Our findings may not be applicable to uncomplicated pregnancies or individuals receiving traditional prenatal care. Additionally, while larger than most studies examining health-related needs in pregnancy, this study is still relatively small. The study design is also cross-sectional in nature, and MHLC scores, information-seeking behaviors, and health-related needs may change over the course of pregnancy. We are currently enrolling participants in a longitudinal study using a similar research protocol to assess changes in MHLC, information-seeking behaviors, and health-related needs over time during and after pregnancy. This latter study will provide a larger cohort and multiple time points at which to examine these variables and their relationships.

Conclusions

This study is one of the first to measure MHLC scores in pregnant women and caregivers and to determine relationships between the MHLC scores, information-seeking behaviors, and health-related needs in the perinatal setting. MHLC scores in pregnant women and caregivers were consistent with published norms, but their relationships differed from prior studies. Powerful Others and Chance scores were significantly and positively correlated, but there were no significant correlations between Internal scores and Chance or Powerful Others scores. Similar to other studies, Internal MHLC scores were positively correlated with information-seeking behaviors in pregnant women and caregivers. Expectant mothers and caregivers with higher Chance scores were less likely to access pregnancy websites or a patient portal, regardless of whether they had other children. The exception to this
trend was that individuals with higher Chance scores were significantly more likely to access FitPregnancy.com, an entertainment-style site with short articles. Pregnant women and caregivers who feel their health outcomes are due to chance may prefer the less formal and extensive resources. MHLC scores were not significantly correlated with the total number of health-related needs expressed or whether needs were met. Pregnant women and caregivers with informational questions about management had higher Powerful Others scores, and those with informational questions about tests had lower Powerful Others scores. These findings may guide designers of health information technologies in tailoring content for pregnant women and caregivers who are likely to express certain types of informational needs. Our ongoing research is exploring relationships among MHLC scores, information-seeking behaviors, and health-related needs over time.

Acknowledgements

This research was funded the Agency for Health Research and Quality, Grant R01 HS021496. This research was supported in part by the Surgical Outcomes Center for Kids at the Monroe Carell Jr. Children's Hospital at Vanderbilt and through the Section for Surgical Sciences at VUMC. Study data were collected and managed using REDCap electronic data capture tools, funded by grant UL1 TR000445 from NCATS/NIH and hosted at VUMC™. The authors are indebted to the many clinical staff, research coordinators, and trainees who participated in recruitment, coordination, and data collection for this study including Laurie Novak, Christopher Simpson, Jamie Ice, Amy Rudin, Patty Jones, Tracy Perry, Kelly Bennett, Mary Anne Carroll, Deborah Wage, Anne Pruet, Martha Dudek, Caitlin Gabarits, Jill Simmons, Marian Dorst, Sai Chennupati, Ebene Ingram, Mary Masterman, Hannah Huth, Daniel Scharf, Hannah Aaron, Alissa Valentine, Sharon Davis, and Heather Hankins.

References

1. Temmerman Marleen, Khosla Rajat, Laski Laura, Mathews Zoe, Say Lale. Women’s health priorities and interventions BMJ 2015; 351 :h4147
Accuracy and Completeness of Clinical Coding Using ICD-10 for Ambulatory Visits

Jan Horsky, PhD1,4, Elizabeth A. Drucker, MD, JD2,3,4, Harley Z. Ramelson, MD, MPH,1,4,5
1Brigham & Women’s Hospital, Div. of General Internal Medicine, Boston, MA; 2Newton-Wellesley Hospital, Dept. of Radiology, Newton, MA; 3Massachusetts General Hospital, Dept. of Radiology, Boston, MA; 4Harvard Medical School, Boston, MA; 5Partners HealthCare, Inc., Boston, MA

Abstract

This study describes a simulation of diagnostic coding using an EHR. Twenty-three ambulatory clinicians were asked to enter appropriate codes for six standardized scenarios with two different EHRs. Their interactions with the query interface were analyzed for patterns and variations in search strategies and the resulting sets of entered codes for accuracy and completeness. Just over a half of entered codes were appropriate for a given scenario and about a quarter were omitted. Crohn’s disease and diabetes scenarios had the highest rate of inappropriate coding and code variation. The omission rate was higher for secondary than for primary visit diagnoses. Codes for immunization, dialysis dependence and nicotine dependence were the most often omitted. We also found a high rate of variation in the search terms used to query the EHR for the same diagnoses. Changes to the training of clinicians and improved design of EHR query modules may lower the rate of inappropriate and omitted codes.

Introduction

The almost seventy thousand codes that comprise the 10th revision of the International Statistical Classification of Diseases and Related Health Problems (ICD-10-CM) are far more detailed than those in the preceding version that clinicians in the United States were working with since the late 1970s.1 This new level of complexity is expected to not only facilitate documenting and reporting causes of mortality and morbidity but to also extend the ability to identify and manage clinical processes with information technology by identifying changes in medication management and by monitoring data for health maintenance and preventive care purposes.2 Highly granular and more accurate data are also indispensable for rapidly expanding secondary uses such as detecting healthcare fraud, developing patient safety criteria, setting healthcare policy and developing public health initiatives, improving clinical performance and, crucially, allowing large-scale analyses for medical research. Codes are also essential in clinical care for phenotyping and predictive modeling of patient state.3 The transition also has significant implications on reimbursement from health care insurers. Diagnostic codes may be used to determine the severity of illness of a provider’s patient population and affect payment rates with newly adopted payment models. Codes that previously could not differentiate between several types of diabetes, for example, are now refined to capture important distinctions but require clinicians to add to their documentation causal underlying conditions or whether the disease was induced by drugs.4 A more detailed description of laterality and location in the patient’s body is also a newly added specification. The previous emphasis on organs and disease that prioritized physician-oriented content is expanded to also cover human responses to disease that are necessary for advanced nursing and long-term care.5

Many electronic health record (EHR) systems integrate clinical documentation and billing information and provide cross-mapping between primarily care-oriented and reimbursement- or report-oriented data. Problem lists, for example, need to conform to standardized vocabularies based on ICD-10 or SNOMED codes for the CMS EHR Incentive Program known as Meaningful Use.6 EHRs may employ their own proprietary reference terminologies that allow to search for and display diagnostic and other concepts in forms that clinicians find customary and meaningful while maintaining mapped connections in the background to more or less granular codes intended for reporting, financial or automated decision-support purposes. Typically, a clinician adding a coded term to a problem list or a fully-qualified ICD-10 code to a billing record starts with typing one or more words, an abbreviation or a string of characters into a free-text query field. A record search engine within the EHR then returns results based on their relevance to the search string and ranks them in a list according to differentiating logic for complete word or partial word matching. For example, a search initiated by typing “esrd” may return many diagnostic codes related to end-stage renal disease. Further search through the results may be necessary, either by reading the list or by repeating the search with different terms, to find the exact code appropriate for the intended purpose. More sophisticated systems...
provide automated assistance with this refinement by using “wizards” or other support interventions to help locate the target diagnosis quickly.

The almost four-fold increase in the number of diagnoses in the current coding system presents a formidable challenge to computer engineers and designers to develop algorithms and human interfaces that can, in the same short time available to clinicians and medical coders in routine practice, query, compare and select the best descriptive diagnostic or other code in the vastly expanded field. A recent survey of perspectives that coders and physicians had about the practical usefulness of ICD-10 showed that most agreed on the need for computer-assisted coding. If the query process is not effective, however, clinicians may find themselves facing a choice between accurate and “close enough” coding when time constraints preclude further refinement of the process. This learned behavior would directly contravene the goal of improved and precise documentation of clinical care made possible with the ICD-10 system.

This study was intended to describe search behavior of clinicians using tools available in large EHR systems who were entering ICD10 codes. Our objectives were to observe interactive behavior that may contribute to incomplete or inaccurate coding and to analyze variations in coded diagnoses for standardized clinical scenarios. Findings of systematic errors or difficulties in completing the coding task may help inform or revise training that clinicians currently receive and to provide evidence and insight for improvement to electronic coding and indicate a need to revise the coding system itself.

Methods

The study was designed as a simulation of a clinical documentation task where clinicians used standardized case scenarios to enter diagnostic codes into the EHR. We asked 23 physicians to read short vignettes describing a variety of ambulatory visits and then enter relevant ICD-10 codes into a mock patient record. Seventeen participants completed two sets of three scenarios, using each set for a different EHR; six completed only one set, using EHR 1, due to technical reasons. In total, there were 40 completed sets: 23 on EHR 1 (12 Sets A and 11 Sets B), and 17 on EHR 2 (10 Sets A and 7 Sets B). The order of set completion (A vs. B) alternated to minimize possible learning bias.

EHR 1 was a commercial and EHR 2 an internally-developed clinical information system. Both required an initial entry of a search term into a free-text field that returned a list of ICD-10 codes with descriptions. If the target term was in the results they could simply select it or further refine the list by entering a different search string. Decision support interventions were available on both systems and were either triggered automatically for a subset of diagnoses in EHR 1, with an option to disregard, or were designed as a part of the entry process on EHR 2. Participants choosing to use decision support on EHR 1 could click on modification terms in pre-determined sets and an algorithm would refine the results accordingly to a single ICD10 code. Initial search on EHR 2 returned a list filtered by patient parameters such as age and gender that could be also modified by selecting answers to term-specific questions (e.g., Laterality? Left, Right, etc.) Both systems used algorithms and branching logic during the guided-search phase to refine result lists and suggest fully specified billable codes. For example, if “otitis media” was the initially entered search term, the support intervention would show subsets of optional terms for laterality, chronicity and recurrence. The visual presentation of these terms, their number and content were different for each system.

Practicing ambulatory clinicians (22 physicians, 1 physician assistant) were recruited through internal email advertising as a sample of convenience. Twenty-one (91%) had ten or more years of professional experience, sixteen (65%) as primary care providers and seven as specialists. Ten (44%) used EHR 1 for 6 months or more, twenty-one (91%) daily in practice. All were proficient in using EHR 2 as it served as the primary ambulatory record system prior to an institution-wide transition to the commercial system. The ICD10 requirement went into full effect in the hospital during the transition period and clinicians have therefore been entering codes with both EHRs for approximately the same time. Both systems had proprietary interface terminology and participants did not use any other sources such as ICD10 on the web.

Participants completed the task individually in the presence of an experimenter on a single workstation that was connected to both EHR systems. They were instructed to read each scenario and then find relevant codes according to their own clinical judgment, and to use their preferred strategy in order to best simulate authentic behavior. The number of codes expected for each scenario was not explicitly stated, only that more than one may be necessary. Interactions of clinicians with the two systems were recorded into an audiovisual media file using Morae screen-capture software running in the background. The recordings of full screens and verbal comments were later analyzed.
Scenarios and accuracy-rating criteria for diagnostic codes

Study scenarios were taken verbatim from interactive case studies made available by the Centers for Medicare & Medicaid Services (CMS) on their Road to Ten website. Since the CMS recommendations for the correct ICD-10 codes were only one example of proper coding, we developed rating criteria for the appropriateness of codes in order to accommodate other correct coding options. Two physicians (HZR, EAD) independently reviewed codes entered by the participants and rated whether they were appropriate in the context of each scenario. They reached consensus on disagreements through a series of discussions. Ratings were based on two criteria: clinical accuracy and completeness. Although the rating was binary (appropriate or not), the reviewers acknowledged the potential range of responses due to variations in clinical judgment and in the complexities of the structure and intent of the ICD-10 coding system. A code rated “appropriate” could be clinically accurate but incomplete in a way that did not alter the diagnosis or miss a clinically crucial information. For example, indicating that allergic rhinitis was seasonal but not explicitly including the causative agent of pollen, or documenting tonsillitis or pharyngitis for a patient with an inflamed pharynx with tonsillar exudate, would be still considered “appropriate”. However, codes without clinically significant information would not. This would include omitting streptococcus as the etiology of tonsillitis in a case where the rapid strep test was known to be positive, or not specifying large intestine in a case of Crohn’s disease complicated by colonic abscess.

We identified one diagnosis in each scenario that was primary and reflected the main reason for visit. All others were categorized as secondary but were still required for complete documentation. They are identified in Table 1. Three of the six scenarios included an itemized clinical assessment in the vignette, as in the example below:

The patient is a three-year-old male brought in by his mother. He has had a low-grade fever of 100.5 for 3 days. He is complaining that his ears hurt with difficulty sleeping. He also has a non-productive cough that started yesterday. Examination of both ears reveals significant redness and fluid in both middle ears with no apparent involvement of the eardrum or tympanic membrane. Further examination of the patient’s breathing and other manifestations indicates an upper respiratory infection. The patient’s parents are chronic heavy smokers and the child is exposed to second-hand smoke in the home environment.

**History:** The patient has recurrent episodes of middle ear infections.

**Assessment:**
- Acute recurrent serous otitis media
- Upper respiratory infection (presumed of viral origin)
- Chronic secondary smoke exposure

In this example, otitis media was the primary diagnosis and the other two diagnoses were secondary diagnoses. Although the participants were given this listing they were not instructed on how many codes to enter. Three scenarios did not include a clinical assessment (as in the example above) and only contained the narrative. The full wording of all scenarios and examples of coding are available on the Road to Ten website maintained by the CMS.

Analysis

Analyses consisted of enumerating the entered codes and evaluating whether they were appropriate or not for the given scenario and whether any codes expected for complete coding of all implicit diagnoses were missing. We also analyzed free-text search terms entered in the query field for variation and the number of results returned by the algorithm. Group comparisons included type of scenario (with vs. without included assessment), type of diagnosis (primary vs. secondary) and electronic health record system (EHR 1 vs. EHR 2). Chi-square or Fisher’s Exact test were used where appropriate to compute group differences and statistical significance with SAS 9.4 software.

Results

We have analyzed three aspects of the diagnostic coding process: a) the appropriateness and completeness of codes for each scenario, b) variation and patterns of similarity of free-text search terms the participants used to initiate queries and c) the number of returned search results.

Accuracy and completeness of coding

The unit of observation was a single ICD-10 code in the context of one standardized scenario. The number of diagnostic codes required for a fully qualified (complete) description varied between 2 and 4 per scenario. We therefore expected 356 codes to be entered if all participants coded all scenarios completely (22 participants x 8 in Set
A plus 18 participants x 10 in Set B) Diagnoses for each scenario and the distribution of appropriate and omitted entries are shown in Table 1. Scenarios that included an explicit assessment are identified.

Just over a half (56%) of all entered diagnostic codes were rated as appropriate and about one quarter were omitted. Diagnoses with the highest accuracy rates, 90% or above, were essential hypertension, acute upper respiratory infection and streptococcal tonsillitis. Two diagnoses, type 1 diabetes mellitus with diabetic polyneuropathy and Crohn’s disease of large intestine with abscess had the greatest proportion of inappropriate codes entered than appropriate (52% and 58%, respectively). The most often omitted codes were for immunization not given (omission rate 91%), dependence on renal dialysis (78% omitted), and personal history of nicotine dependence that was omitted by two thirds of participants.

We also compared separately scenarios where an assessment was included with the narrative as it could serve as a cue to both the specificity of the diagnosis as well as the number of diagnoses that required coding. We computed their respective rates of accuracy and completeness. Results are shown in Table 2.

Table 1. Accuracy and completeness of ICD-10 coding for each scenario

<table>
<thead>
<tr>
<th>Scenario Sets and Diagnoses</th>
<th>Assessment Given in the Scenario</th>
<th>Primary Dx</th>
<th>Appropriate</th>
<th>Omitted</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Yes</td>
<td>%</td>
</tr>
<tr>
<td><strong>Set A</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Scenario 1</em></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Streptococcal tonsillitis</td>
<td>Primary</td>
<td>20</td>
<td>91%</td>
<td>2</td>
</tr>
<tr>
<td>Immunization</td>
<td></td>
<td>0</td>
<td>2</td>
<td>9%</td>
</tr>
<tr>
<td><em>Scenario 2</em></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Moderate persistent asthma</td>
<td>Primary</td>
<td>17</td>
<td>77%</td>
<td>5</td>
</tr>
<tr>
<td>Seasonal allergic rhinitis</td>
<td></td>
<td>17</td>
<td>77%</td>
<td>2</td>
</tr>
<tr>
<td>Second-hand smoke exposure</td>
<td></td>
<td>14</td>
<td>61%</td>
<td>4</td>
</tr>
<tr>
<td><em>Scenario 3</em></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Crohn’s disease of large intestine with abscess</td>
<td>Primary</td>
<td>13</td>
<td>48%</td>
<td>14</td>
</tr>
<tr>
<td>Old myocardial infarction</td>
<td></td>
<td>11</td>
<td>48%</td>
<td>2</td>
</tr>
<tr>
<td>Personal history of nicotine dependence</td>
<td></td>
<td>6</td>
<td>27%</td>
<td>1</td>
</tr>
<tr>
<td><strong>Set B</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Scenario 1</em></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acute recurrent otitis media, bilateral</td>
<td>Primary</td>
<td>11</td>
<td>61%</td>
<td>7</td>
</tr>
<tr>
<td>Acute upper respiratory infection</td>
<td></td>
<td>17</td>
<td>94%</td>
<td>0</td>
</tr>
<tr>
<td>Exposure to environmental tobacco smoke</td>
<td></td>
<td>13</td>
<td>72%</td>
<td>3</td>
</tr>
<tr>
<td><em>Scenario 2</em></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Type 1 diabetes with diabetic kidney disease</td>
<td>Primary</td>
<td>9</td>
<td>50%</td>
<td>4</td>
</tr>
<tr>
<td>Type 1 diabetes with diabetic polyneuropathy</td>
<td></td>
<td>4</td>
<td>21%</td>
<td>11</td>
</tr>
<tr>
<td>Dependence on renal dialysis</td>
<td></td>
<td>4</td>
<td>22%</td>
<td>0</td>
</tr>
<tr>
<td>End stage renal disease</td>
<td></td>
<td>9</td>
<td>50%</td>
<td>0</td>
</tr>
<tr>
<td><em>Scenario 3</em></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Paroxysmal atrial fibrillation</td>
<td>Primary</td>
<td>14</td>
<td>78%</td>
<td>4</td>
</tr>
<tr>
<td>Essential (primary) hypertension</td>
<td></td>
<td>18</td>
<td>100%</td>
<td>0</td>
</tr>
<tr>
<td>Underdosing of other hypertensive drugs</td>
<td></td>
<td>8</td>
<td>40%</td>
<td>4</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>206</td>
<td>56%</td>
</tr>
</tbody>
</table>
Primary diagnoses were significantly less likely to be omitted than secondary diagnoses. Only 5 out of 125 expected primary diagnoses were not entered. These omissions were all for the same diagnosis: type 1 diabetes with diabetic kidney disease. The omission rate for secondary diagnoses was 38%, with only a half coded appropriately. There was no significant effect of the EHR used to enter the codes on the main outcome measures and the proportions in each category were similar. However, when we analyzed primary diagnoses separately we found that significantly more appropriate codes were entered with EHR 2 (79%) than with the commercial EHR 1 (59%). The proportion of omitted codes was equally low for both systems (4%). Scenarios that included clinical assessment with the narrative showed higher rates of appropriate codes (73% vs. 40%) and reduced rates of codes that were omitted by three quarters (41% to 11%).

We next assessed the frequency and distribution of ICD-10 codes for every diagnosis in the scenarios (Table 3). The distribution pattern was typically a majority of entries in one or two appropriate codes and a few other codes with frequencies of one or two. For example, out of the 17 codes entered for paroxysmal atrial fibrillation, 13 (76%) were the correct I48.0 codes, two were I48.2 and two I48.9. Seven diagnoses had between 2 and 4 codes, five between 5 and 7 and two had nine and ten different codes, respectively. Two scenarios showed markedly different distribution: type 1 diabetes with diabetic polyneuropathy and Crohn’s disease of large intestine with abscess. In the diabetes case, only 4 out of the 15 codes entered (27%) were appropriate and the rest was distributed over 8 other diagnoses, each used by just one participant (by two in one case). For Crohn’s disease, 13 (48%) codes were the appropriate K50.114 and 8 other codes were also entered. There were three diagnoses with a complete agreement in coding among all participants: end-stage renal disease (N18.6), essential (primary) hypertension (I10) and acute upper respiratory infection (J06.9).

**Search terms and selection of codes**

We found no systematic differences in the type of search terms used, such as preference for different abbreviations, character sequences or full query terms that appeared to be EHR-specific. Table 4 shows the combined results for both systems, in descending order of use frequency. Terms with frequencies of 1 are grouped together and their cumulative proportion is shown. Relatively few terms were common to the queries of multiple clinicians. The proportion of common terms was typically 20%-40%, rarely exceeding 50% for each diagnosis while those that were unique and used only once comprised the largest group, often 40%-50% of all terms. For example, unique search strings such as ‘secondary’, ‘smoker’, ‘passive smoke’ or ‘parent smoke’ made up 40% of all search terms for the diagnostic code Z77.22 “Contact with and (suspected) exposure to environmental tobacco smoke”. The term ‘smoke’ by itself was used by only 6 participants (21%). Terms that appeared repeatedly in the search queries of multiple clinicians were either single-word disease

<table>
<thead>
<tr>
<th>Group Comparisons</th>
<th>ICD-10 Codes</th>
<th>Appropriate N %</th>
<th>Not Appropriate N %</th>
<th>Omitted N %</th>
<th>Exact χ² Test Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Scenario Diagnosis</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>&lt; .0001</td>
</tr>
<tr>
<td>Primary</td>
<td>125</td>
<td>34%</td>
<td>84</td>
<td>67%</td>
<td>5</td>
</tr>
<tr>
<td>Secondary</td>
<td>241</td>
<td>66%</td>
<td>121</td>
<td>50%</td>
<td>29</td>
</tr>
<tr>
<td><strong>All Diagnoses</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(ns)</td>
</tr>
<tr>
<td>EHR 1</td>
<td>212</td>
<td>58%</td>
<td>115</td>
<td>54%</td>
<td>44</td>
</tr>
<tr>
<td>EHR 2</td>
<td>154</td>
<td>42%</td>
<td>90</td>
<td>58%</td>
<td>21</td>
</tr>
<tr>
<td><strong>Primary Dx Only</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>&lt; .003</td>
</tr>
<tr>
<td>EHR 1</td>
<td>73</td>
<td>58%</td>
<td>43</td>
<td>59%</td>
<td>27</td>
</tr>
<tr>
<td>EHR 2</td>
<td>52</td>
<td>42%</td>
<td>41</td>
<td>79%</td>
<td>9</td>
</tr>
<tr>
<td><strong>Scenario Assessment</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>&lt; .0001</td>
</tr>
<tr>
<td>Given</td>
<td>177</td>
<td>48%</td>
<td>129</td>
<td>73%</td>
<td>29</td>
</tr>
<tr>
<td>Not given</td>
<td>189</td>
<td>52%</td>
<td>76</td>
<td>40%</td>
<td>36</td>
</tr>
</tbody>
</table>

Table 2. Accuracy and completeness by diagnosis type, scenario modality and EHR

Table 3. Accuracy and completeness by diagnosis type, scenario modality and EHR

<table>
<thead>
<tr>
<th>Group Comparisons</th>
<th>ICD-10 Codes</th>
<th>Appropriate N %</th>
<th>Not Appropriate N %</th>
<th>Omitted N %</th>
<th>Exact χ² Test Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Scenario Diagnosis</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>&lt; .0001</td>
</tr>
<tr>
<td>Primary</td>
<td>125</td>
<td>34%</td>
<td>84</td>
<td>67%</td>
<td>5</td>
</tr>
<tr>
<td>Secondary</td>
<td>241</td>
<td>66%</td>
<td>121</td>
<td>50%</td>
<td>29</td>
</tr>
<tr>
<td><strong>All Diagnoses</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(ns)</td>
</tr>
<tr>
<td>EHR 1</td>
<td>212</td>
<td>58%</td>
<td>115</td>
<td>54%</td>
<td>44</td>
</tr>
<tr>
<td>EHR 2</td>
<td>154</td>
<td>42%</td>
<td>90</td>
<td>58%</td>
<td>21</td>
</tr>
<tr>
<td><strong>Primary Dx Only</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>&lt; .003</td>
</tr>
<tr>
<td>EHR 1</td>
<td>73</td>
<td>58%</td>
<td>43</td>
<td>59%</td>
<td>27</td>
</tr>
<tr>
<td>EHR 2</td>
<td>52</td>
<td>42%</td>
<td>41</td>
<td>79%</td>
<td>9</td>
</tr>
<tr>
<td><strong>Scenario Assessment</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>&lt; .0001</td>
</tr>
<tr>
<td>Given</td>
<td>177</td>
<td>48%</td>
<td>129</td>
<td>73%</td>
<td>29</td>
</tr>
<tr>
<td>Not given</td>
<td>189</td>
<td>52%</td>
<td>76</td>
<td>40%</td>
<td>36</td>
</tr>
<tr>
<td>ICD-10</td>
<td>Scenario Set A</td>
<td>N</td>
<td>ICD-10</td>
<td>Scenario Set B</td>
<td>N</td>
</tr>
<tr>
<td>--------</td>
<td>----------------</td>
<td>---</td>
<td>--------</td>
<td>----------------</td>
<td>---</td>
</tr>
<tr>
<td><strong>Immunization</strong></td>
<td></td>
<td></td>
<td><strong>Acute recurrent serous otitis media, bilateral</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Z23</td>
<td>Encounter for immunization</td>
<td>2</td>
<td>H65.0</td>
<td>Acute serous otitis media</td>
<td>1</td>
</tr>
<tr>
<td>Streptococcal tonsillitis</td>
<td></td>
<td></td>
<td>H65.01</td>
<td>Acute serous otitis media, right ear</td>
<td>1</td>
</tr>
<tr>
<td>J02.*</td>
<td>Streptococcal pharyngitis</td>
<td>16</td>
<td>H65.02</td>
<td>Acute serous otitis media, left ear</td>
<td>1</td>
</tr>
<tr>
<td>J03.*</td>
<td>Acute streptococcal tonsillitis, unspecified</td>
<td>4</td>
<td>H65.03</td>
<td>Acute serous otitis media, bilateral</td>
<td>1</td>
</tr>
<tr>
<td>J03.01</td>
<td>Acute recurrent streptococcal tonsillitis</td>
<td>1</td>
<td>H65.06*</td>
<td>Acute serous otitis media, recurrent, bilateral</td>
<td>11</td>
</tr>
<tr>
<td>J03.90</td>
<td>Acute tonsillitis, unspecified</td>
<td>1</td>
<td>H65.07</td>
<td>Acute serous otitis media, recurrent, uns. ear</td>
<td>1</td>
</tr>
<tr>
<td><strong>Moderate persistent asthma</strong></td>
<td></td>
<td></td>
<td>H66.003</td>
<td>Acute suppurative otitis media without spontaneous rupture of ear drum, bilateral</td>
<td>2</td>
</tr>
<tr>
<td>J45.20</td>
<td>Mild intermittent asthma, uncomplicated</td>
<td>1</td>
<td><strong>Acute upper respiratory infection</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>J45.31</td>
<td>Malignant asthma w. acute exacerbation</td>
<td>1</td>
<td>J06.9*</td>
<td>Acute upper respiratory infection, unspecified</td>
<td>17</td>
</tr>
<tr>
<td>J45.40*</td>
<td>Moderate persistent asthma, uncomplicated</td>
<td>8</td>
<td>Exposure to environmental tobacco smoke</td>
<td></td>
<td></td>
</tr>
<tr>
<td>J45.41*</td>
<td>Moderate persistent asthma w. acute exacerbation</td>
<td>9</td>
<td>Y26.XXX</td>
<td>Exposure to smoke, fire and flames, undetermined intent, initial encounter</td>
<td>2</td>
</tr>
<tr>
<td>J45.50</td>
<td>Severe persistent asthma, uncomplicated</td>
<td>1</td>
<td>Z77.22*</td>
<td>Contact with and (suspected) exposure to environmental tobacco smoke</td>
<td>13</td>
</tr>
<tr>
<td>J45.901</td>
<td>Unspecified asthma with acute exacerbation</td>
<td>2</td>
<td>Z77.9</td>
<td>Other contact with and (suspected) exposures hazardous to health</td>
<td>1</td>
</tr>
<tr>
<td><strong>Seasonal allergic rhinitis</strong></td>
<td></td>
<td></td>
<td><strong>Type 1 diabetes with diabetic chronic kidney disease</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>J30.1*</td>
<td>Allergic rhinitis due to pollen</td>
<td>1</td>
<td>E10.21*</td>
<td>Type 1 DM with diabetic nephropathy</td>
<td>3</td>
</tr>
<tr>
<td>J30.2*</td>
<td>Other seasonal allergic rhinitis</td>
<td>13</td>
<td>E10.22*</td>
<td>Type 1 DM with diabetic chronic kidney disease</td>
<td>6</td>
</tr>
<tr>
<td>J30.9*</td>
<td>Allergic rhinitis, unspecified</td>
<td>3</td>
<td>E10.29</td>
<td>Type 1 DM with diabetic other chronic kidney complication</td>
<td>1</td>
</tr>
<tr>
<td>J31.0</td>
<td>Chronic rhinitis</td>
<td>1</td>
<td>E10.9</td>
<td>Type 1 DM without complications</td>
<td>2</td>
</tr>
<tr>
<td>T78.40XA</td>
<td>Allergy, unspecified, initial encounter</td>
<td>1</td>
<td>E11.9</td>
<td>Type 2 DM without complications</td>
<td>1</td>
</tr>
<tr>
<td><strong>Second-hand smoke exposure</strong></td>
<td></td>
<td></td>
<td><strong>Type 1 diabetes with diabetic polyneuropathy</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P96.81</td>
<td>Exposure to (parental) (environmental) tobacco smoke in the perinatal period</td>
<td>1</td>
<td>E08.40</td>
<td>DM due to underlying condition with diabetic neuropathy, unspecified</td>
<td>1</td>
</tr>
<tr>
<td>T59.811A</td>
<td>Toxic effect of smoke, accidental (unintentional), initial encounter</td>
<td>1</td>
<td>E10.40*</td>
<td>Type 1 DM with diabetic neuropathy, unspecified</td>
<td>1</td>
</tr>
<tr>
<td>Z57.31</td>
<td>Occupational exposure to environmental tobacco smoke</td>
<td>1</td>
<td>E10.42*</td>
<td>Type 1 DM with diabetic polyneuropathy</td>
<td>3</td>
</tr>
<tr>
<td>Z77.22*</td>
<td>Contact with and (suspected) exposure to environmental tobacco smoke</td>
<td>14</td>
<td>E11.40</td>
<td>Type 2 DM with diabetic neuropathy, unspecified</td>
<td>3</td>
</tr>
<tr>
<td>Z77.29</td>
<td>Contact with and (suspected) exposure to other hazardous substances</td>
<td>1</td>
<td>E11.43</td>
<td>Type 2 DM with diabetic autonomic (poly)neuropathy</td>
<td>1</td>
</tr>
<tr>
<td><strong>Crohn’s disease of large intestine with abscess</strong></td>
<td></td>
<td></td>
<td>E13.42</td>
<td>Other specified DM with diabetic polyneuropathy</td>
<td>1</td>
</tr>
<tr>
<td>K50.10</td>
<td>Crohn's disease of large intestine without complications</td>
<td>1</td>
<td>E34.9</td>
<td>Endocrine disorder, unspecified</td>
<td>1</td>
</tr>
<tr>
<td>K50.114*</td>
<td>Crohn's disease of large intestine with abscess</td>
<td>13</td>
<td>G62.81</td>
<td>Critical illness polyneuropathy</td>
<td>1</td>
</tr>
<tr>
<td>K50.914</td>
<td>Crohn's disease, unspecified, with abscess</td>
<td>5</td>
<td>G62.89</td>
<td>Other specified polyneuropathies</td>
<td>1</td>
</tr>
<tr>
<td>K65.0</td>
<td>Generalized (acute) peritonitis</td>
<td>1</td>
<td>G62.9</td>
<td>Polyneuropathy, unspecified</td>
<td>2</td>
</tr>
<tr>
<td>L02.91</td>
<td>Cutaneous abscess, unspecified</td>
<td>1</td>
<td><strong>Dependence on renal dialysis</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R10.0</td>
<td>Acute abdomen</td>
<td>2</td>
<td>N18.6*</td>
<td>End stage renal disease</td>
<td>1</td>
</tr>
<tr>
<td>R10.81</td>
<td>Abdominal tenderness</td>
<td>1</td>
<td>Z99.2*</td>
<td>Dependence on renal dialysis</td>
<td>3</td>
</tr>
<tr>
<td>R10.84</td>
<td>Generalized abdominal pain</td>
<td>2</td>
<td><strong>End stage renal disease</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Z87.19</td>
<td>Personal history of other diseases of the digestive system</td>
<td>2</td>
<td>N18.6*</td>
<td>End stage renal disease</td>
<td>9</td>
</tr>
<tr>
<td><strong>Old myocardial infarction</strong></td>
<td></td>
<td></td>
<td><strong>Paroxysmal atrial fibrillation</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I21.3</td>
<td>ST elevation (STEMI) myocardial infarction of unspecified site</td>
<td>1</td>
<td>I48.0*</td>
<td>Paroxysmal atrial fibrillation</td>
<td>13</td>
</tr>
<tr>
<td>I25.10*</td>
<td>Atherosclerotic heart disease of native coronary artery without angina pectoris</td>
<td>8</td>
<td>I48.2</td>
<td>Chronic atrial fibrillation</td>
<td>2</td>
</tr>
<tr>
<td>I25.2*</td>
<td>Old myocardial infarction</td>
<td>3</td>
<td>I48.91</td>
<td>Unspecified atrial fibrillation</td>
<td>2</td>
</tr>
<tr>
<td>I25.810</td>
<td>Atherosclerosis of coronary artery bypass graft(s) without angina pectoris</td>
<td>1</td>
<td><strong>Essential (primary) hypertension</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Personal history of nicotine dependence</strong></td>
<td></td>
<td></td>
<td>N10*</td>
<td>Essential (primary) hypertension</td>
<td>18</td>
</tr>
<tr>
<td>F17.211*</td>
<td>Nicotine dependence, cigarettes, in remission</td>
<td>2</td>
<td><strong>Underdosing of other hypertensive drugs</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Z72.0</td>
<td>Tobacco use</td>
<td>1</td>
<td>Z91.128*</td>
<td>Patient's intention underdosing of medication regimen for other reason</td>
<td>1</td>
</tr>
<tr>
<td>Z87.891*</td>
<td>Personal history of nicotine dependence</td>
<td>3</td>
<td>Z91.14*</td>
<td>Patient's other noncompliance with medication regimen</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Z91.19</td>
<td>Patient's noncompliance with other medical treatment and regimen</td>
<td>4</td>
</tr>
</tbody>
</table>
Table 4. Frequency and proportion of entered search terms (for N=1, the proportion shown is cumulative).

<table>
<thead>
<tr>
<th>Search Terms for Set A</th>
<th>N</th>
<th>%</th>
<th>Search Terms for Set B</th>
<th>N</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Scenario 1</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>strep</td>
<td>7</td>
<td>32%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>strep pharyngitis</td>
<td>3</td>
<td>14%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>tonsillitis</td>
<td>2</td>
<td>9%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(N=10) acute pharyngitis, pharyn, pharyngi, pharyngit, sore throat, strep pha, strep phar, strep pharyng, strep throat, strep tonsilli</td>
<td>1</td>
<td>45%</td>
<td>(N=10) acute otitis media, acute recurrent serious otitis media, acute serious otitis media, ot med, otit med, otitis media, otitis med, recurrent serious otitis media, sercus otitis media, serious otitis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(N=8) flu imm, flu shot, immuniz, immunization, immunization influ, influenza vaccine requirement, influenza, influenza vaccine</td>
<td>1</td>
<td>100%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Scenario 2</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>asthma</td>
<td>18</td>
<td>78%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(N=5) moder pers asthma, moderate persistent, moderate persistent asthma, persistent asthma, rad</td>
<td>1</td>
<td>22%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>allergic rhinitis</td>
<td>6</td>
<td>25%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>allergic rh</td>
<td>4</td>
<td>17%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>seasonal allergic rhinitis</td>
<td>2</td>
<td>8%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(N=10) allergic rhinitis, all rh, aller, allergic, allergic rh seaso, allergy, rhinitis, rhinitis aller, seas alle rhi, seasonal allergic rh</td>
<td>1</td>
<td>42%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>smoke</td>
<td>6</td>
<td>17%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>smoke exposure</td>
<td>5</td>
<td>14%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>second hand smoke</td>
<td>4</td>
<td>11%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>smoke exp</td>
<td>3</td>
<td>9%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>second hand</td>
<td>2</td>
<td>6%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>second-hand smoke</td>
<td>2</td>
<td>6%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>tobacco</td>
<td>2</td>
<td>6%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(N=11) environmental smoke, sec smoke, second hand smo, second hand smoke exposure, second smoke, second-hand, secondhand, smoke expo, smoke secondhand, tobacco exposure, tobacco second</td>
<td>1</td>
<td>31%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Scenario 3</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>crohn</td>
<td>10</td>
<td>29%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>abdominal pain</td>
<td>3</td>
<td>9%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>crohns</td>
<td>3</td>
<td>9%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>abd pain</td>
<td>2</td>
<td>6%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>abscess</td>
<td>2</td>
<td>6%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(N=14) Crohn, abcess, abd abs, abdominal abscess, abscess abdomen, abscess abdominal, chronic, crohn diseases, crohn's, crohn's abscess, crohn's disease, crohns abs, crohns disease, h/o croh</td>
<td>1</td>
<td>41%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>cad</td>
<td>9</td>
<td>50%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mi</td>
<td>2</td>
<td>11%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>myocardial infarction</td>
<td>2</td>
<td>11%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(N=5) coronary artery, coronary artery disease, csd, h/o myocardial infarction, ischemic heart disease</td>
<td>1</td>
<td>22%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>tobacco</td>
<td>4</td>
<td>22%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>former smoker</td>
<td>2</td>
<td>11%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>former tobacco</td>
<td>2</td>
<td>11%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>smoker</td>
<td>2</td>
<td>11%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>smoking</td>
<td>2</td>
<td>11%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(N=6) ex smok, ex-smoker, h/o smoking, history of tobacco, nicotine, past smoker</td>
<td>1</td>
<td>34%</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
names such as asthma (used in 78% of all queries for ‘moderate persistent asthma’) or neuropathy (40%), or common abbreviations such as ‘cad’ (50%), ‘uri’ (59%), or ‘esrd’ (70%).

We also examined the interactive behavior of clinicians with respect to repeated formulation of search queries. The mean number of search terms, that is, instances when participants entered a new search string into a free-text field, is shown in Table 5.

<table>
<thead>
<tr>
<th>EHR</th>
<th>Queries entered (N)</th>
<th>Queries per diagnosis</th>
<th>Number of queries per diagnosis (%)</th>
<th>Results returned per query (μ)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>(μ)</td>
<td>Max (N)</td>
<td>1</td>
</tr>
<tr>
<td>EHR 1</td>
<td>157</td>
<td>1.38</td>
<td>6</td>
<td>77%</td>
</tr>
<tr>
<td>EHR 2</td>
<td>114</td>
<td>1.41</td>
<td>9</td>
<td>79%</td>
</tr>
</tbody>
</table>

Participants typically entered only one initial search term. Almost 80% selected a code after first search attempt. Only a small minority (less than 15%) needed to or were willing to change the query term and repeat the search cycle, and fewer still (under 10%) tried more times. Several clinicians repeated the search up to nine times but they acknowledged during the test session that they would not have gone to this length under the time constraints of actual practice. The average of the number of queries per diagnosis was almost equal for the two systems although there was a large difference in the number of returned results. The commercial EHR 1 displayed on average six times as many results as EHR 2.

**Discussion**

We found a large variability in the accuracy of ICD-10 diagnostic codes in a simulation of six clinical scenarios by clinicians using two EHRs. This varying level of diagnostic accuracy (the use of appropriate codes) seemed to increase when the code could be searched for with relatively ubiquitous and intuitive terms and abbreviations. For example, diagnoses with the rate of appropriate coding close to 80% or higher, such as streptococcal tonsillitis, asthma, rhinitis, upper respiratory infection or atrial fibrillation (Table 1) also had a high proportion of initial search queries repeatedly formulated with the same or similar terms by many clinicians (Table 4). Search terms such as ‘strep’, ‘pharyngitis’, ‘asthma’, ‘rhinitis’ or ‘uri’ likely produced a list of results that included the target code. Conversely, the largest variation in search terms was for diagnoses related to nicotine dependence or tobacco smoke exposure that could be expressed in many ways and do not lend themselves easily to one highly specific word. Terms with relatively low specificity are more difficult to map in reference terminologies to related codes and therefore clinicians were likely looking at different sets of results when completing the same search task with very different initial query terms.

The large number of results that search queries often returned had to be further refined and this extra series of steps was a common source of frustration for many participants. For example, some clinicians pointed out that they had to read through a dozen or more variations on gestational hypertension and preeclampsia included in the results for a male patient before they found the appropriate code for primary hypertension. This low efficiency of the query and entry process may lead over time to adopting a satisficing strategy – a decision-making heuristic in which the first option that seems to address most needs is selected rather than the most optimal choice. The tendency of clinicians to initiate only one or two different queries per diagnosis (Table 5) also seems to support the notion that there is a practical (likely time) limit on every query and that clinicians may need to tradeoff accuracy or completeness for efficiency.

The omission rate, while minimal for primary diagnoses, reached almost forty percent for secondary diagnoses. The non-coded conditions, if otherwise documented in notes, may not substantially affect the treatment of individual patients but would provide a distorted view of the patient population with underestimates of disease severity and comorbidity. For example, old myocardial infarction, dependence on dialysis and personal history of nicotine dependence were more often missing (up to almost 80%) than coded (Table 1). Leaving out such key clinical information in the diagnostic codes may have financial implications for the clinicians and may have research repercussions for those using such data for investigations.
Conclusion

We demonstrated significant deficiencies in documented diagnostic codes using clinical simulations of two different EHR systems. The promise of improved clinical documentation with ICD-10 may not be quickly realized due to three main factors. First, clinicians entering codes are not adequately trained to understand the requirements and nuances of the ICD-10 coding system. Additional training and specialized resources may be necessary at many institutions. Second, the design of query systems in EHRs may negatively affect the code selection process and lead clinicians to choose a less specific or less desirable code. Lastly, accurate understanding of the meaning of a particular code, especially those that indicate “unspecified,” is often possible only within the context of the code hierarchy. Clinicians therefore may need to have a good working knowledge of the ICD-10 structure to correctly identify what is unspecified and what may need to be specified. The core objective of decision support interventions is to assist them with this comparison by clearly contextualizing the returned candidate entries as the code set is too large and complex to be effectively learned in its entirety. The lack of appropriate coding may also affect reimbursement rates, especially with increased adoption of alternate payment models that rely on case-mix analysis to adjust financial reimbursement, as well as on potential utility of ICD-10 codes for research and other purposes.

ICD-10 will improve national healthcare initiatives such as Meaningful Use, value-based purchasing, payment reform and quality reporting. Without ICD-10 data, there will be serious gaps in the ability to extract important patient health information needed to support research and public health reporting, and move to a payment system based on quality and outcomes.14 These goals can be achieved if HIT better supports clinicians in making adequately informed choices by providing decision support, guidance and effective tools.

References

8 Cartagena FP, Schaeffer M, Rifai D, Doroshenko V, Goldberg HS. Leveraging the NLM map from SNOMED CT to ICD-10-CM to facilitate adoption of ICD-10-CM. J Am Med Inform Assoc. 2015 May;22:659-70.
9 Morae. 3.3 ed. Okemos, MI: TechSmith Corporation; 2012.
14 AHIMA. What are the benefits of ICD-10? 2015 [cited; Available from: http://www.ahima.org/topics/icd10/faqs
Evaluating and Improving an Outpatient Clinic Scheduling Template Using Secondary Electronic Health Record Data

Michelle R. Hribar, PhD\textsuperscript{1}, Sarah Read-Brown, BA\textsuperscript{2}, Leah Reznick, MD\textsuperscript{2}, Michael F. Chiang, MD, MA\textsuperscript{1,2}

\textsuperscript{1}Medical Informatics & Clinical Epidemiology, and \textsuperscript{2}Ophthalmology
Oregon Health & Science University, Portland, OR

Abstract

Improving the efficiency of outpatient clinics is challenging in the face of increased patient loads, decreased reimbursements and potential negative productivity impacts of using electronic health records (EHR). We modeled outpatient ophthalmology clinic workflow using discrete event simulation for testing new scheduling templates that decrease patient wait time and improve clinic efficiency. Despite challenges in implementing the new scheduling templates in one outpatient clinic, the new templates improved patient wait time and clinic session length when they were followed. Analyzing EHR data about these schedules and their adherence to the template provides insight into new policies that can better balance the competing priorities of filling the schedules, meeting patient demand and minimizing wait time.

Introduction

Physicians today are pressured to see more patients in less time for less reimbursement due to persistent concerns about the accessibility and cost of healthcare.\textsuperscript{1,2} Furthermore, clinicians are concerned that the adoption of electronic health records (EHRs) has negatively impacted their productivity.\textsuperscript{3–5} For example, at Oregon Health & Science University (OHSU), which completed a successful EHR implementation in 2006 that received national publicity, ophthalmologists currently see 3-5\% fewer patients than before EHR implementation and require >40\% additional time for each patient encounter.\textsuperscript{6}

Clinic inefficiencies result when patients arrive and clinic resources (staff, exam rooms, and providers) are not available to serve them. This mismatch of arrivals and availability can be caused by ad-hoc scheduling protocols that increase patient wait time.\textsuperscript{7} Previous work has demonstrated that EHR timing data can be used for building clinic simulation models for studying scheduling templates before implementing them in the clinic.\textsuperscript{8} These models, along with prior research, confirm that scheduling longer encounters with higher variability at the end of the day helps reduce wait time.\textsuperscript{9}

While simulation models can predict improvements with new scheduling templates, implementing them in clinic faces real-world challenges such as competing clinic priorities, and established scheduling practices. While efficient workflow is a priority in outpatient clinics, so is filling all appointment slots and serving urgent walk-in patients. Clinic staff become adept at navigating clinic schedules, ensuring that all urgent patients are scheduled and appointment slots are filled, including those that are last-minute cancellations. New scheduling templates disrupt these practices and priorities, which can make following new schedules challenging.

We implemented a new scheduling template in one outpatient ophthalmology clinic (LR) at OHSU starting in September 2016 after preliminary studies demonstrated its effectiveness.\textsuperscript{10} As expected, there were challenges in implementing the new template, but when it was followed, there were reductions in average patient wait times and clinic session lengths. The purpose of this paper is to analyze challenges in a real-world implementation of simulation model based workflow improvements. Analyzing data about adherence to the new template will provide insight into new policies that can better balance the competing priorities of filling the schedules, meeting patient demand, and minimizing wait time, ultimately improving the clinic workflow. We found that 1) complex templates that the simulation predicts as optimal can be challenging to follow, 2) secondary use of EHR data allowed for a thorough analysis of a complex scheduling template and suggested ways to simplify it, 3) when the template was followed, scheduling according to exam length significantly improves patient wait time and session length, especially when short exams are scheduled at the start of the clinic, 4) these insights have impacts for developing new scheduling policies and strategies, and 5) EHR timing data enables detailed evaluation and improvement of clinic workflow modifications.
Methods
This study was approved by the Institutional Review Board at Oregon Health & Science University (OHSU).

Study Environment
OHSU is a large academic medical center in Portland, Oregon. The ophthalmology department includes over 50 faculty providers, who perform over 90,000 annual outpatient examinations. The department provides primary eye care, and serves as a major referral center in Pacific Northwest and nationally. We studied one provider’s outpatient clinic in pediatric ophthalmology (LR).

Over several years, an institution-wide EHR system (EpicCare; Epic Systems, Madison, WI) was implemented throughout OHSU. This vendor develops software for mid-size and large medical practices, is a market share leader among large hospitals, and has implemented its EHR systems at over 200 hospital systems in the United States. In 2006, all ophthalmologists at OHSU began using this EHR. All ambulatory practice management, clinical documentation, order entry, medication prescribing, and billing tasks are performed using components of the EHR.

Clinic Workflow
Interviews and clinic observations determined the basic clinic workflow of the majority of patient appointments as shown in Figure 1. Patients check in and wait to be seen. An ancillary staff member (an ophthalmic technician or orthoptist) performs an initial exam in an exam room. At the end of this exam, the patient’s eyes may be dilated. If this is the case, the patient returns to the waiting room while waiting for the dilation to take effect. After the dilation occurs (approximately 25 minutes), the patient is returned to an exam room and waits for the physician exam. If the patient’s eyes were not dilated, the patient remains in the exam room and waits for the physician. After the physician’s exam, the patients check out and leave. Waiting occurs before each of the two exams if a staff member (2 total) or physician (1 total) is not available.

Simulation Model Using EHR Timestamp Data
To build models for studying the clinic’s efficiency, we first had to collect large amounts of timing data for each step of the clinic workflow. In prior studies, we identified sources of timestamp data within the EHR and verified them against timing data from in person observations. We used the clinical data warehouse and ophthalmology datamart for OHSU’s EHR (EpicCare; Epic Systems, Madison, WI). While these timestamps are specific to OHSU’s implementation in ophthalmology, comparable timestamps are available for other vendors, installations and specialties. We used this timing data in our models, and for measuring the average patient wait time before and after we changed the scheduling templates.

We used Arena simulation software to build a discrete event simulation model of a pediatric outpatient ophthalmology clinic’s (LR) workflow using the EHR timing data. Previous studies validated this model and used it to test different scheduling templates. Schedules with the longest appointments near the end of the clinic session minimized average patient wait time, but could also lengthen clinic sessions. As a compromise, scheduling long patients near the end of the clinic, but not at the very end, still reduces wait time without unduly lengthening the clinic.

New Schedule Template
Based on the results of the simulation studies and input from the clinic provider and staff, we created a new scheduling template to be used within the pediatric provider’s clinic. Patients’ appointments were scheduled according to their anticipated length and dilation status: shortest appointments were earliest, with the longest appointments near the end of the clinic to minimize patient wait time. We classified the short appointments as roughly the shortest quarter of
appointments and the long appointments as the longest quarter. The remaining appointments were designated as medium. Each clinic session was a half-day with patients scheduled in 15 minute blocks. Because of OHSU’s scheduling policies, all appointment blocks must be the same length. The morning clinic schedule is given in Figure 2; the afternoon schedule was similar, but without the empty block and one each fewer of short and medium blocks. Morning clinics ran from 8:00 am until about 12:00 noon with 15 total patients; afternoon clinics began at 12:45 pm or 1:00 pm and ended by 5:00 pm. To ensure that afternoon clinics ended on time, there were usually fewer patients scheduled (13). In both clinic sessions, we scheduled the long patients near the end of the clinic, with 3 non-long patients after them to avoid clinic sessions ending late. In the morning clinic, the staff wished to include an empty slot in the middle of the long patient appointment slots to allow for catchup. The first appointment slot was double booked with one dilated patient and one not. This allowed the provider to get started quickly with the non-dilated patient.

Previous studies identified which exams were expected to be shorts and longs, but also found that there was quite a bit of variability among appointment lengths that was not easy to predict (Spatar D, et al. IOVS 2016;57:ARVO E-Abstract 5566). Fortunately, the provider was very good at assessing the length of the exam. For this reason, the provider documented her prediction (short, medium, or long) for the next appointment in the followup notes for the encounter, which are viewed by the scheduler when scheduling the next appointment. New appointments are always long.

Finally, the provider and staff wished to try alternating dilated and non-dilated appointments for preventing multiple patients waiting for the provider at the same time. Appointment slots on the schedule were designated as dilated or non-dilated exams, which required schedulers anticipating which type of exam the patient required at time of scheduling. Again, the provider documented her preference in the notes and other information about the appointment helped with this prediction.

The new scheduling templates were put into place starting in April 2016 for appointments scheduled September 1 or after. The provider started documenting her exam length and dilation predictions in her followup notes at this time. Even with a lead time of 5 months, there were still appointments already scheduled after September 1 using the old templates; the clinic decided to leave them as scheduled, regardless of whether they followed the new scheduling template.

New Schedule Template Evaluation Metrics

To assess the performance of the new scheduling template, we used two metrics: patient wait time and clinic length. The patient wait time was calculated using EHR timestamps and data about the appointment from the OHSU datamart. Exam lengths were determined by EHR audit log timestamps recorded between the appointment’s check in and check out times. Wait times are calculated by subtracting the exam length and dilation length (25 minutes for dilated exams, 0 for non-dilated exams) from the entire time the patient was checked in for their appointment.

We found that we needed to adjust this wait time since many patients arrived significantly early for their appointments. Because OHSU provides care for a large geographic region throughout the Pacific Northwest, many patients arrive early because of concerns about traffic and/or with the hope of being examined earlier. Often the patients are seen before their appointment times, so it was not meaningful to count only the wait time that occurred after their scheduled appointment time. Instead we eliminated any wait time that occurred before the exam started that was also before their scheduled appointment time. Once the patient started the exam, all subsequent wait time was counted. Clinic length is calculated as the difference between the first timestamp recorded during an exam in the clinic and the last timestamp recorded for the last exam of the clinic.
New Schedule Template Evaluation

To evaluate the impact of the new schedule in October – December 2016, we first compared the average patient wait times and clinic lengths to baseline sessions for the same clinic from July 2015 – June 2016, when the old scheduling templates were used. We used a t-test to compare our two metrics and investigated differences between the two groups with respect to new patient appointments, adult patients, dilated patients, early and late patients, double booked appointments, unfilled appointment blocks and known long exams at the start of the clinic session.

To further analyze the new schedule, we compared the scheduled appointments with the scheduling template to determine when and how well the templates were followed. Again, we used data from the EHR to assess the adherence to the scheduling template: the followup notes from the previous exam, the date of the previous exam, the type of scheduled encounter, and the age of the patient. The followup notes, when available, provided the provider’s predictions for exam length and dilation. When the notes weren’t available, we used our previously determined rules for predicting exam length based on encounter type and patient age (new vs. return, adult vs. child, pre-op, post-op or followup). The date of the previous appointment determined the dilation status (most patients require dilation once annually). We coded each appointment block according to its predicted exam length and dilation along with the corresponding template block’s exam length and dilation. We used linear regression to determine the wait time and session length impacts of following the schedule according to exam length and dilation.

Results

New vs. Old Scheduling Templates

We compared the performance of the new scheduling templates (Oct – Dec 2016) to our baseline (July 2015 - June 2016), which used the old templates. For our two main metrics, average patient wait time and clinic session length,

<table>
<thead>
<tr>
<th></th>
<th>Old Template</th>
<th>New Template</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n (%) Mean ± SD</td>
<td>n (%) Mean ± SD</td>
</tr>
<tr>
<td>Total encounters</td>
<td>1,376 -</td>
<td>677 -</td>
</tr>
<tr>
<td>Patient wait time per encounter</td>
<td>22.1 ± 18.2</td>
<td>20.6 ± 18.6</td>
</tr>
<tr>
<td>Clinic sessions</td>
<td>121</td>
<td>54</td>
</tr>
<tr>
<td>Session length</td>
<td>2:21 ± 32.0</td>
<td>242.3 ± 35.8</td>
</tr>
<tr>
<td>Volume per clinic session</td>
<td>12.5 ± 1.9</td>
<td>12.5 ± 2.3</td>
</tr>
<tr>
<td>Dilated patients</td>
<td>739 (54%) 5.2 ± 2.5</td>
<td>383 (57%) 7.1 ± 2.0</td>
</tr>
<tr>
<td>Adult Patients (≥ 18yrs)</td>
<td>88 (6%) 3.9 ± 0.8</td>
<td>63 (9%) 1.2 ± 1.0</td>
</tr>
<tr>
<td>New patients</td>
<td>277 (20%) 2.3 ± 1.6</td>
<td>128 (19%) 2.4 ± 1.4</td>
</tr>
<tr>
<td>Late patients (&gt; 15 Min)</td>
<td>64 (5%) 3.5 ± 0.8</td>
<td>31 (5%) 0.6 ± 0.7</td>
</tr>
<tr>
<td>Early patients (&lt; 15 Min)</td>
<td>369 (27%) 3.2 ± 2.0</td>
<td>182 (27%) 3.4 ± 1.8</td>
</tr>
<tr>
<td>Unfilled blocks *</td>
<td>240 2.1 ± 1.6</td>
<td>83 1.5 ± 1.2</td>
</tr>
<tr>
<td>Double booked blocks†</td>
<td>160 1.5 ± 0.9</td>
<td>63 1.2 ± 0.9</td>
</tr>
<tr>
<td>Adult patients in first 4 blocks</td>
<td>30 3.3 ± 0.5</td>
<td>16 0.3 ± 0.6</td>
</tr>
<tr>
<td>New patients in first 4 blocks</td>
<td>108 1.0 ± 0.9</td>
<td>32 0.6 ± 0.7</td>
</tr>
</tbody>
</table>

* Unfilled blocks represent any appointments where a patient canceled or didn’t show, and/or the staff was unable to fill the block
† Blocks that had 2 patients scheduled and both showed.

**Table 1: Comparison of New and Old Scheduling Templates.** For both patient wait time and session length, the new schedule was not significantly different from baseline (p > 0.05 for both). All other clinic characteristics were similar for both the new schedule and baseline.
the new template was not significantly different from the old template as shown in Table 1. The means for wait time (20.6 minutes for new and 22.1 minutes for old) and the means for session length (242.3 minutes for new and 232.1 minutes for old) were not significantly different when compared using a t-test ($p = 0.09$ and $p = 0.08$, respectively).

We also measured other characteristics relating to the schedules for both the old and new template periods in Table 1. Both groups were similar with respect to the number of dilated, new, adult, early and late patients per session. They also had similar numbers of unfilled blocks and double-booked blocks. Finally, because the new template does not schedule long patient appointments at the start of the clinic session, we compared the number of adult and new patients (always long appointments) at the start of the session. Both were similar, with slightly more new patients at the start of the baseline clinics.

**Scheduling Errors**

Since the wait times, clinic lengths and clinic characteristics for the new schedule test period were similar to the old templates in the baseline, we analyzed the new schedule test period for adherence to the new schedule template.

**Figure 3** shows the impact of schedule adherence on average patient wait time and session length. For each clinic session, the percentage of correct blocks for exam length and dilation was plotted versus the average patient wait time and session length for that clinic session. As the percentage of correctly scheduled exam length blocks increased, the wait time decreased significantly: 3.7 minutes per 10% increase in correct exam length blocks ($p < 0.0001$). Similarly, the session length decreased by 7.7 minutes per 10% increase in correct exam length blocks ($p = 0.008$), according to the linear regression models shown in the figure. Correctly scheduling the dilation blocks did not significantly decrease the wait time or session length.

To understand how the clinic adhered to the template, we analyzed the scheduling errors. **Table 2** outlines the different types of appointment blocks and the errors made in scheduling them; 240/677 (36%) appointment blocks were not scheduled correctly with respect to exam length and 270/677 (40%) were incorrect with respect to dilation. Breaking the appointment blocks down according to exam length shows that similar percentages (31% - 39%) were scheduled incorrectly. Finally, 14/24 (58%) blocks intended to be blank were scheduled with appointments. Since errors with were occurring frequently—30 – 40% of the time—it was important to determine which errors had the biggest impact.
Figure 4 shows the impact of scheduling errors for short, medium and long exam block errors. The most significant impacts were for short block errors: they increased wait time by 1.5 minutes for each 10% increase in errors (p = 0.008) and they increased clinic length by 4.8 minutes per every 10% increase in errors (p = 0.005) as predicted by the linear regression models shown in the figure. The only other significant impact was on clinic length for medium block errors: those decreased the session length by 5.5 minutes for each 10% increase in errors (p = 0.01).

Looking at just the short scheduling errors, we see that scheduling longs in a short exam block had the most significant impact as shown in Figure 5. For every 10% increase in the number of long exams in short blocks the wait time increased by 1.5 minutes (p = 0.01). There wasn’t a significant increase in session length for long exams in short blocks, nor was there a significant impact of scheduling medium exams in short blocks on wait time or session length.

Discussion

The new clinic schedule did not significantly decrease patient wait time or clinic session length over baseline for the entire 3 months we evaluated, presumably because the schedule was not followed for all clinic sessions. This highlights the difficulty in implementing scheduling changes predicted to be optimal based on simulation models. Analyzing the clinic sessions’ schedules for errors gives us insight into improving the scheduling process in the future. In particular, we found:

<table>
<thead>
<tr>
<th>Appointment Blocks</th>
<th>Number</th>
<th>Predicted</th>
<th>Actual</th>
<th>Incorrect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dilation Status</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dilated</td>
<td>277</td>
<td>367</td>
<td>383</td>
<td>82 (30%)</td>
</tr>
<tr>
<td>Undilated</td>
<td>400</td>
<td>310</td>
<td>294</td>
<td>188 (47%)</td>
</tr>
<tr>
<td>Total</td>
<td>677</td>
<td>270</td>
<td>4</td>
<td>240 (36%)</td>
</tr>
<tr>
<td>Exam length</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Short</td>
<td>162</td>
<td>191</td>
<td>63</td>
<td>39 (39%)</td>
</tr>
<tr>
<td>Medium</td>
<td>340</td>
<td>331</td>
<td>105</td>
<td>105 (31%)</td>
</tr>
<tr>
<td>Long</td>
<td>161</td>
<td>155</td>
<td>58</td>
<td>58 (36%)</td>
</tr>
<tr>
<td>Blank</td>
<td>14</td>
<td>24</td>
<td>14</td>
<td>14 (58%)</td>
</tr>
<tr>
<td>Total</td>
<td>677</td>
<td>270</td>
<td>4</td>
<td>240 (36%)</td>
</tr>
</tbody>
</table>

Table 2: Scheduling Errors. The schedule was not followed for both exam length and dilation: 39.9% of blocks were incorrectly scheduled for dilation and 36.2% of blocks were incorrectly scheduled for exam length.

Figure 4: Impacts of Short, Medium, and Long Block Scheduling Errors. Scheduling short blocks incorrectly had the most significant impact on wait time and session length (p = 0.0009 and p=0.005, respectively). Scheduling medium blocks incorrectly also significantly impacted session length (p = 0.01), but not patient wait time. Long block scheduling errors did not significantly impact patient wait time or session length.
1. Complex templates that simulation predicts as optimal can be challenging to follow. At best, 60% of blocks were correct for dilation and 64% were correct for exam length. From our discussions with schedulers, we determined the new template was too difficult to schedule for both exam length and dilation, there weren’t enough dilation blocks (only 277 were designated as dilated blocks, but there were 383 actual dilated patients), not all patients had provider predictions, and there are competing priorities for maintaining clinic volume. When patients cancel, the appointment blocks must be filled, usually with new patients who require long exams. This highlights the substantial challenges for implementing new scheduling templates and the importance of evaluation for determining their benefits.

Table 3 provides a breakdown of the potential sources for scheduling errors. Some appointments were scheduled prior to the implementation of the scheduling template in April of 2016 and these appointments weren’t changed if they violated the schedule. For example, 14.5% of the errors in short blocks were due to this. Also, when patients canceled close to the appointment time, there is pressure to fill the appointment block; this caused 29.1% of the short block errors. Missing provider predictions for appointments also results in errors; this caused 78.2% of the short block errors. Some of these were new patients who didn’t have predictions (Schedulers know they are always longs) or patients whose previous appointment was prior to April of 2016, before the provider started documenting her predictions. The remainder were appointments that should have had predictions, but were missing. Nevertheless, even with predictions, there were still scheduling errors. This highlights the need for more provider predictions and better training for schedulers to follow them. It also motivated us to find ways to simplify the schedule.

2. Secondary use of EHR data allowed for a thorough analysis of a complex scheduling template and suggested ways to simplify it. We determined dilation block scheduling errors do not appear to worsen wait times or session length and it is difficult to schedule for both dilation and exam length. For both of those reasons, we simplified the template to exclude dilation, which we hope eases the schedulers’ burden.

3. When the template was followed, scheduling according to exam length significantly improves patient wait time and session length, especially when short exams are scheduled at the start of the clinic. Wait time decreases by 3.7 minutes per 10% increase in correct exam length blocks and session length decreases by 7.7 minutes per 10% increase in correct blocks. Further, scheduling short patients at the start of the clinic helps the clinic stay on time and reduce patient wait times for subsequent patients. It is most significant to avoid scheduling long exams in the short exam blocks, but even medium exams can have an impact. This result is important since it validates our simulation models and previous studies⁸,⁹ and confirms that our schedule can improve clinic efficiency if it is
4. These insights can lead to better scheduling and potentially innovative dynamic scheduling strategies. There are currently many opportunities and possible innovations for improving scheduling in healthcare. For example, because scheduling long patients at the start of the clinic has the biggest impact on patient wait time, all attempts should be made to avoid doing this, even when an early block becomes available for a long patient (e.g. a new patient). Schedulers should attempt to shift shorter exams to this block or double book the new patient later, leaving the short block empty. The unfilled blocks (typically 1-2 per clinic session) due to no shows or late cancels can help offset this double booking. Also, since the clinic has a high number of early patients, any that have short or medium exams can be worked in to fill that early blank spot. Other possibilities include innovative dynamic scheduling strategies where patients are scheduled for a given clinic session, but not assigned an appointment time until just prior to the session. That way, patients could be arranged so that the shortest appointments were first, followed by the mediums and longs.

5. EHR timing data allows for detailed evaluation and improvement of clinic workflow modifications. EHR data can be used to measure metrics such as average patient wait times, session lengths, and adherence to scheduling templates which allows for a detailed evaluation of a new scheduling strategy. While this study focuses on a new scheduling template in an outpatient ophthalmology clinic, EHR timing data can be used to analyze any workflow modification in any setting. This timing data can determine if a new modification is having the desired impact and why, leading to potential improvements. Without using EHR data, metrics would be gathered manually through observations, which is prohibitive for any extended test period and may not provide as rich of insights.

Limitations

This study has the following limitations: 1) It uses EHR timestamps for measuring exam times, calculating patient wait times, and session lengths. If the provider’s or staff’s EHR use doesn’t coincide with their exams, the wait time and session length calculations could be wrong. Previous studies have validated this approach, but errors may still occur. 2) The new schedule was studied for a 3 month time period—its performance may vary if a longer test period including multiple seasons was used. We plan to evaluate the schedule again, particularly after more scheduler training and after relaxing the dilation criteria. 3) We focused our evaluation to quantitative methods, but qualitative evaluation will provide more insight into the benefits and challenges of the new schedule implementation. A followup qualitative study is planned to address this limitation. 4) OHSU’s policy of equal length scheduling blocks eliminates other potential optimal scheduling policies, such as variable length appointments. 5) We assume that long appointments can be scheduled near the end of the clinic; if this is not the case, then other strategies would need to be used to mitigate the potential delays incurred by long patients at the start of the clinic. Future studies are needed to investigate strategies such as variable length appointments or empty “catchup” blocks for early long appointments.

Table 3: Sources of Scheduling Errors. A few scheduling errors occurred because of appointments that were scheduled prior to the new template’s implementation: it accounts for 14.5% of the short block errors. Others occurred because of late scheduling, but the majority of errors happened without provider predictions (78.2% of the short block errors), but even with predictions there were still errors (21.8% of the short block errors).
Conclusions

Simulation models predict that scheduling according to length—short appointments first, followed by longer appointments—significantly decrease patient wait times and session length. Real world implementations of these scheduling rules, however, can be challenging, but using EHR timing data can evaluate the effects of compliance. Secondary use of EHR data allows for a detailed analysis of a new scheduling template, with the goal of identifying critical aspects of the modification and those that can be eliminated. This study focuses on a scheduling template in outpatient ophthalmology, but the data analysis techniques apply to the evaluation of any workflow modification in any setting.

Acknowledgements

This study was supported by grants K99 LM12238 and P30 EY10572 from the National Institutes of Health (Bethesda, MD), and by unrestricted departmental funding from Research to Prevent Blindness (New York, NY).

References

A First Step Towards Behavioral Coaching for Managing Stress: A Case Study on Optimal Policy Estimation with Multi-stage Threshold Q-learning

Xinyu Hu, MS1, Pei-Yun S. Hsueh, PhD3, Ching-Hua Chen, PhD3, Keith M. Diaz, PhD2, Ying-Kuen K. Cheung, PhD1, Min Qian, PhD1

1 Department of Biostatistics, Columbia University, New York, NY, USA; 2 Center for Behavioral Cardiovascular Health, Columbia University, New York, NY, USA; 3IBM T.J. Watson Research Center, Yorktown Heights, NY, USA

Abstract
Psychological stress is a major contributor to the adoption of unhealthy behaviors, which in turn accounts for 41% of global cardiovascular disease burden. While the proliferation of mobile health apps has offered promise to stress management, these apps do not provide micro-level feedback with regard to how to adjust one’s behaviors to achieve a desired health outcome. In this paper, we formulate the task of multi-stage stress management as a sequential decision-making problem and explore the application of reinforcement learning to provide micro-level feedback for stress reduction. Specifically, we incorporate a multi-stage threshold selection into Q-learning to derive an interpretable form of a recommendation policy for behavioral coaching. We apply this method on an observational dataset that contains Fitbit ActiGraph measurements and self-reported stress levels. The estimated policy is then used to understand how exercise patterns may affect users’ psychological stress levels and to perform coaching more effectively.

Introduction
Psychological stress has long been shown as the source of unhealthy behaviors and in turn contributes to the increase of cardiovascular disease risk [1]. Public health experts have been advocating on the importance of behavioral factors on cardiovascular risk, which account for 41% of global cardiovascular disease burden [2, 3]. Improving on exercise behaviors, in particular, has been shown as a cost-effective solution compared to pharmacological treatments. Recent research has demonstrated the benefit of regular exercise and physical activities on reducing stress and improving emotional well-being [4, 5]. Yet its implementation often fails due to the lack of systematic monitoring for effective coaching. Early research in clinical decision support systems has attempted to integrate multimodal sensor input to start guiding the management of psychological stress. However, its clinical uptake has been slow due to the lack of intuitive reasoning and explanation capability [6].

In fact, the stress-behavior mechanistic pathway is a process that is highly individualized, in which interpretations of the environmental demand and adaptive capacity ("appraisals") are important in determining psychological, behavioral, and physiological responses to stress [7]. Previous clinical trials have demonstrated the effectiveness of individualized stress management for obtaining certain goals, e.g., hypertension control [8]. The recent proliferation of mobile health and self-tracking mobile apps has offered new promise to further integrate daily monitoring data into the process of individualized stress management. However, a most recent review of existing stress management apps [9] has shown that there is still a lack of apps that can provide user behavior-oriented feedback such as prompting for a specific goal. Therefore, how to leverage individual observations to provide incremental micro-level feedback has been a key challenge to the development of behavioral coaching systems.

In many other domains, recent technological advances in mobile health have used smart phones and wearable devices to collect data and deliver interventions over time. Evidence has emerged for the potential of drawing sequential patterns from patient-reported outcomes (PRO) [10] and Ecological Momentary Assessments (EMA) [11] to make predictions and recommendations. This is favorable by behavior scientists and healthcare researchers, since mobile devices provide an interactive platform that has great potential to promote health behaviors and achieve better health outcomes [12, 13]. Despite that simply measuring users’ calories intake and exercise patterns has only limited effect on outcomes, such as long-term weight loss [14], it still has been expected that the development of mobile-based behavioral coaching systems would be beneficial to understand the implicit user preferences and barriers so as to guide users for goal attainment with personalized training plans. In addition, behavior interventions can be tailored based on users’ changing needs and ongoing performance. The delivery of such interventions has started being attempted in the area of substance use disorder [15], physical activity [16], mental health [17, 18] and diabetic self-management [19]. Yet the methodology is under development for these behavior intervention studies.
As a first step towards an interactive and personalized behavioral coaching system for stress management, we need to build on the observations of actions and self-reported outcomes over time to derive recommendations of beneficial behaviors and to guide users to achieve healthy outcomes. Intuitively, this can be represented as a sequential decision-making problem. Under similar problem specifications, reinforcement learning algorithms [20], for example Q-learning, can be applied to solve this problem by estimating an optimal policy, i.e., a set of decision rules for selecting actions that maximize long-range cumulative reward. However, the “black-box” nature of reinforcement learning algorithms makes it less appealing for stress management coaching systems that require explicit explanations for recommended actions -- as demonstrated in the previous clinical decision support research [6]. In many other fields, such as clinical trials, explicit explanations are also preferred for decision-making. Statistical methods have been developed for prescribing adaptive treatment regimes in clinical trials [21, 22, 23]. In this paper, we apply a Multi-stage Threshold Q-learning (MTQL) method and construct interpretable policies that map from subjects’ up-to-date observations to recommend actions adaptively for stress reduction coaching. We make model assumptions for the Q-learning algorithm and incorporate a threshold-finding step to identify optimal sub-goals for recommendation. In the rest of this paper, we will first introduce the stress data used in the study and describe the MTQL method. Then we will demonstrate and discuss the interpretable policies derived from the MTQL method on the dataset.

Method

Data Collection

In this paper, the stress dataset used was collected in a longitudinal study during Jan 2014 to July 2015 from 79 subjects. The observations used to evaluate the MTQL method were from the first twelve weeks of the study. 75 subjects had complete data for that period of time.

Subjects in the study were healthy individuals aged 18 or older who responded to fliers posted throughout the buildings of Columbia University Medical Center and who on phone screening reported only intermittent engagement in exercise (exercise 6-11 times per month, not on a regular basis), had daily access to a computer with Internet, and had an iPhone or Android phone. Excluded were individuals who had previously been told by a healthcare professional to restrict physical activity, were deemed unable to comply with the protocol (either self-selected, by indicating during screening that he/she could not complete all requested tasks), were unavailable during the following continuous twelve months, had serious medical comorbidity that would compromise their ability to engage in usual physical activity, had occupational work demands that required rigorous activity or would make responding to the EMA dangerous, or were unable to read and speak English.

Physical activity was continuously and objectively measured using a wrist-based model of the Fitbit (Fitbit Flex; Fitbit, Inc., San Francisco, CA) [24]. The Fitbit Flex is a microelectromechanical triaxial accelerometer that tracks the wearer’s daily physical activity including steps, intensity of activities (sedentary, light, moderate, or vigorous) and energy expenditure. We and others have shown that the Fitbit Flex is a valid and reliable device for measuring physical activity in adults [25, 26].

Minute-by-minute activity data were extracted using Fitabase (Small Steps Labs, San Diego, California). We defined a “social day” as the period from 3am one day until 2:59am the next day. Non-wear was defined as greater than 60 consecutive minutes with fewer than 10 steps. Only days with a minimum of 10 hours of wear time were included in analyses. Each valid day was then classified as an exercise or non-exercise day. Our objective measure for exercise was defined as at least 24 minutes of moderate or vigorous physical activity (MVPA) within any consecutive 30-minute period, thereby allowing for up to 6 minutes of below threshold physical activity (e.g., rest). Such an instance is referred to as an “MVPA bout”. This definition was adapted from conventional accelerometer processing approaches used in many population-based studies [27, 28, 29] and suggested by best practice recommendations [30] wherein a healthful bout of physical activity is defined as a 10-minute or longer bout of MVPA with an allowance of 2 minutes below threshold (e.g. 8 out of 10 consecutive minutes). The 8 out of 10-minute bout was extrapolated to 30 minutes for the purposes of this study to be consistent with physical activity guidelines which recommend exercise in bouts of 10 minutes or more for at least 30 minutes a day, while accommodating interruptions.

An electronic diary that used the subject’s own iPhone or Android phone was used to capture momentary and summary aspects of their perceived stress. Each morning upon rising, the subject responded to a question on a browser asking them, “Overall, how stressful do you expect today to be?”, answered on a scale from 0 (Not at all) to 10 (Extremely). Similarly, each evening, the subject responded to a question on a browser asking them, “Overall, how stressful was your day?”, answered on the same scale. In addition, the diary system was programmed to query
the subject via text message or email 3 random times per day over their preset hours of wakefulness (e.g., 7am-10pm), with the specific time period programmed individually according to the subject’s own sleep schedule. Notifications were separated by at least 1 hour. Each of the 3 daytime momentary assessments included questions concerning key sources of stress (one screen listing sources of stress including work, argument, traffic, deadline trouble, paying bills, running late, none, or other, with the subject checking all that apply) [31], stress appraisal using the four-item Perceived Stress Scale [32]. Before each prompt, the subject responded to a question asking “how stressful did you feel?”, answered on the scale specified before. Data transmission was secured via SSL (i.e., https) and sent to a managed server.

The data used for the analyses includes duration of daily MVPA bouts, perceived stress levels, exogenous and environmental factors (such as day in a week, daylight time, temperature and precipitation).

Problem Formulation for the Multi-stage Threshold Q-learning Method

We observed a sequence of data discussed in the last section and aim to maximize the mean stress level reduction from baseline over a given time period. At baseline, the subjects’ characteristics and stress information were collected. We defined the baseline stress level as the mean stress level over the first four-week of the study. After four weeks, daily actions (e.g., MVPA pattern) and daily health outcomes (e.g., stress level) were observed at each stage. The action was binary and defined as if the mean duration of daily MVPA bout over the time period of the stage was greater than 30 minutes then the action was denoted as 1 otherwise 0. Mean stress level was used as the health outcome of interest, which was a continuous value ranging from 0 to 10. Everyday the stress level was assessed three times during the wakefulness time of the subjects and the overall stress level was assessed at the end of the day. If the value of the overall stress level was not missing, then the daily stress level was defined as the overall stress level. Otherwise, it was defined as the average of the stress levels assessed during the wakefulness time. The stress level at each stage was calculated by averaging over the daily stress level during the time period of the stage. In a T-stage study for each subject $i$, we observe the data:

$$\{O_{it}, A_{it}, O_{it+1}, \ldots, O_{iT}, A_{iT}, O_{(T+1)}\}$$

$O_{it}$ is a baseline stress level, $O_{it}$, where $1 < t \leq T + 1$, is a stress level at the $(t-1)$-stage, and $A_{it}$ is a binary action at the $t$-stage. Since stress levels were self-evaluated, subtracting the baseline stress level from the following stress levels evaluated helps to adjust for heterogeneity of self-evaluation across subjects. We define $R_{it} = O_{it} - O_{it+1}$ as the stress reduction from baseline at the $t$-stage, which is also the reward at the $t$-stage. We use $H_{it}$ to denote historical information for subject $i$ before the $t$-stage, for example $H_{i1} = \{O_{i1}\}$ and $H_{i2} = \{O_{i1}, A_{i1}, O_{i2}\}$.

Q-learning is a commonly used method of reinforcement learning introduced by computer scientists to construct high quality policies [20]. Q-learning models the interaction between an agent and environment. In our setting, the agent refers to each subject in the study, and environment refers to the system of human body and external source of observations. The Q-learning algorithm uses a backward induction to estimate the optimal policy $\pi$, which consists of a sequence of decision rules $\pi_1, \ldots, \pi_T$, under which the cumulative reward is maximized. At the $t$-stage, a decision rule $\pi_t$ takes the historical information of a subject $H_{it}$ as an input and outputs a recommended action $A_{it}$. The $t$-th subject with historical information $H_{it}$ takes an action $A_{it}$ and gains a reward $R_{it}$. In the computer science field, researchers focus on solving infinite time horizon decision-making problems [20], while in the statistical field, methodologies were developed to solve finite time horizon problems using statistical models [21-23, 33-36]. The MTQL method combines statistical methods with the Q-learning algorithm to construct interpretable decision rules. In the following, we use lower-case variables to denote the realizations of the corresponding upper-case random variables. We define a value function at the $t$-stage:

$$V_t(h_t) = E_\pi \left( \sum_{k=0}^{T} r_{t+k} | h_t \right)$$

Then the optimal value function is the one maximized over all possible policies $\pi$ in the policy space $\Pi$, i.e., $V_t^\ast(h_t) = \max_{\pi \in \Pi} V_t(h_t)$. We define the optimal Q function at the $t$-stage:

$$Q_t^\ast(h_t, a_t) = E[r_t + V_{t+1}(h_{t+1}) | h_t, a_t]$$

By definition, the relationship between a value function and a Q function is $V_t^\ast(h_t) = \max_a Q_t^\ast(h_t, a) = E[r_t + \max_{a_{t+1}} Q_{t+1}^\ast(h_{t+1}, a_{t+1}) | h_t, a_t]$. 

932
The optimal Q function $Q^*_t$ (referred to as the Q function in the following) is estimated based on the Bellman equation [37] backwards through time. The optimal policy is defined as $\pi^*_t(h_t) = \arg\max_{a_t} Q^*_t(h_t, a_t)$.

In our proposed MTQL method, we assume a regression model for the Q function. The benefit of assuming a regression model is to make the estimated Q function and policy interpretable. A threshold-selection step is added into the learning process. The thresholds are set for the health outcome variables and selected by maximizing the expected cumulative reward. The estimated thresholds are considered as goal-setting options of the health outcome. Our goal is to estimate a sequence of optimal decision rules that maximizes the expected health outcome.

We illustrate the proposed method in multi-stage studies and discuss the effect of setting different numbers of stages on the expected health outcome. Our goal is to estimate a sequence of optimal decision rules that maximizes the mean stress reduction from baseline over a four-week period. This can be formulated as a four-stage study ($T=4$) with one week for each stage, or a two-stage study ($T=2$) with two weeks for each stage, or a one-stage study ($T=1$) with four weeks for one stage. We describe the variables used in modeling Q functions in Table 1.

<table>
<thead>
<tr>
<th>Variable name</th>
<th>Variable label</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Y$</td>
<td>Mean stress level reduction from baseline over a four-week period.</td>
</tr>
<tr>
<td>$O_t$</td>
<td>Baseline stress level (mean stress level at the first four-week of the study).</td>
</tr>
<tr>
<td>$A_t$</td>
<td>Binary action at the $t$-stage, $t = 1, \ldots, T$.</td>
</tr>
<tr>
<td>$R_t$</td>
<td>Stress level reduction from baseline at the $t$-stage, $t = 1, \ldots, T$.</td>
</tr>
<tr>
<td>$X_t$</td>
<td>Stress level reduction from the previous stage at the $t$-stage, $t = 1, \ldots, T$.</td>
</tr>
<tr>
<td>$I_t$</td>
<td>Indicator of whether $X_t$ exceeds a threshold, $t = 1, \ldots, T$.</td>
</tr>
</tbody>
</table>

Table 1. List of variables used in modeling Q functions.

The Q function consists of the main effect and the interaction effect. The interaction effect assesses the association of the historical information and the value of the Q function under different actions. We use $H_{10} = (1, R_1, A_1, R_2, A_2, \ldots, R_T)$ to denote the design matrix for the main effect and use $H_{11} = (1, I_{t-1}, A_{t-1}, I_t, I_{t-1}I_t, I_{t-1}A_{t-1}, A_{t-1}I_t)$ to denote the design matrix for the interaction effect. Let $\theta_t$ be a vector of regression coefficients which consists of the coefficients of the main effect $\theta_{t0}$ and the coefficients of the interaction effect $\theta_{t1}$, i.e., $\theta_t = (\theta_{t0}, \theta_{t1})$. The model of the Q function at the $t$-stage, for $t = 2, \ldots, T$, is of the form:

$$Q_t(h_t, a_t; \theta_t, c_t) = h_{t0}\theta_{t0} + (h_{t1}a_t)\theta_{t1}$$

For the first stage ($t = 1$), let $H_{10} = (1, O_1)$ and $H_{11} = (1, 1(O_1 > c_t))$, thus

$$Q_1(h_1, a_1; \theta_1, c_1) = h_{10}\theta_{10} + (h_{11}a_1)\theta_{11}$$

The optimal policy is estimated from the last stage to the first stage using a backward induction. The optimal policy at the $t$-stage is of the form:

$$\pi^*_t(h_{t1}; \theta_{t1}, c_t) = \arg\max_{a_t} Q^*_t(h_{t1}, a_t)\theta_{t1}$$

The parameters of a policy consist of the regression parameter $\theta_t$ and the threshold parameter $c_t$. $\theta_t$ is estimated as a least square estimator. More specifically, at final stage $T$ given an estimated threshold parameter $\hat{c}_T$,

$$\hat{\theta}_T = \arg\min_{\theta_T} \sum_{i=1}^{n} (R_{iT} - Q_T(h_{iT}, a_{iT}; \theta_T, \hat{c}_T))^2$$

At a previous stage $t$:

$$\pi^*_t = R_{it} + \max_{a_{i(t+1)}} Q_{t+1}(h_{i(t+1)}, a_{i(t+1)}; \hat{\theta}_{t+1}, \hat{c}_{t+1})$$

933
The estimated threshold parameter $\hat{c}_t$ is chosen as the one that maximizes the mean stress level reduction from baseline $Y$. We introduce an inverse probability weighting estimator (IPWE) [34], which is an expected outcome under a given policy. IPWE does not depend on model assumptions, thus is robust to model misspecification. It is used to evaluate the estimated parametric policies. IPWE is defined as

$$\frac{\sum_{i=1}^{n} W_i Y_i}{\sum_{i=1}^{n} W_i}$$

where the weight

$$W_i = \frac{\prod_{k=1}^{T} 1(a_{ik} = \pi_k(h_{ik}; \theta_t, c_t))}{\prod_{k=1}^{T} p_k(a_{ik} | h_{ik})}$$

$p_t(a_{it} | h_{it})$ is a propensity score estimated using logistic regression. The logistic regression model is built based on variable selection using P-value less than 0.2 as a criterion, and it contains the variables of stress level, action, gender and mean precipitation. The threshold parameter $c_t$ is estimated using the genetic algorithm [38] to maximize the mean stress reduction from baseline. Genetic algorithm is a heuristic searching algorithm and has computational advantages for solving optimization problems.

**Policy Estimation from the Observational Data**

In the observational study, we consider a binary action of whether exercising more than 30 minutes daily on average (i.e., daily MVPA bout on average is greater than 30 minutes) over the time period of a stage, and define “active” as exercising more than 30 minutes daily on average, and “inactive” otherwise. The health outcome of our interest is the self-evaluated stress level on average over the time period of a stage. We use the data of the first four weeks of the study for assessing the baseline information, the data of the second four weeks as the training data and the data of the third four weeks as the test data. We apply the MTQL method on the training data to estimate the optimal policy and evaluate it on the test data. The estimated tree-structured policy is presented below in Figure 1-3 respectively. Our goal is to compare the expected outcomes estimated in a one-stage study, in a two-stage study and in a four-stage study in order to evaluate the effect of setting different numbers of intervention stages.

**Results**

![Figure 1. Estimated optimal policy for the one-stage study.](image-url)
Figure 2. Estimated optimal policy for the two-stage study.
Figure 1 shows the estimated optimal policy in a one-stage study over the time period of four weeks. That is, if a subject has perceived baseline stress level higher than 4.4 (on a scale of 0-10), then for obtaining the optimal stress reduction outcome, this subject should be recommended to stay “active” (i.e., exercising more than 30 minutes daily on average) in the next four weeks. This recommendation is similar to what has been offered by the existing stress management apps (as surveyed in [9]). The threshold learned here can help subjects and their care team to understand whether keeping active would be expected to yield a positive impact on the overall stress reduction for them.

As we expect that a fine-grained level of policy will reveal more actionable micro strategies for subjects and their care team, we further derive the optimal two- and four-stage policy from the data. Figure 2 shows the optimal policy estimated using the data of the first two weeks as one stage and the second two weeks as the other stage. The stress level 3.7 is learned as the baseline threshold to decide if we should recommend being active in the first two weeks, and 1.3 is learned as the threshold for the mean stress level increased from baseline to differentiate whether a target subject can potentially obtain an optimal outcome by adopting a sequential active-to-inactive strategy: i.e., staying active in the first two weeks, and then take a break in the next two weeks. For those subjects who have baseline stress level lower than or equal to 3.7, it would be important to observe whether being active in the first two weeks followed by an increase of mean stress level in the next two weeks. If the mean stress level increases greater than 1.3, then it is better not to suggest the active action any further to accommodate the individual behavioral preferences and barriers.

Figure 3 shows the finer-grained policy (one week as one stage) learned for making recommendations for subjects. It is important to know whether the target subject starts with a perceived mean stress level higher than 5.4, whether the subject is active during the first week and whether exercising helps reduce the subject’s mean stress level. The observed intermediate outcomes and actions would be important to determine whether to recommend being active or not in the next stage. To make the fourth stage recommendation, it is critical to keep observing for the mean stress level change in the second week and whether the user is active or not in the third week.

In addition to the empirical interpretation of the learned policy for curating actionable micro strategies, we are also interested in learning the quantifiable impact of a finer-grained policy on the overall outcome (i.e. the mean stress reduction from baseline). Applying the learned policies on the test data, the mean stress level reduction from baseline in the one-stage study is 0.04, the one in the two-stage study is 0.58, and the one in the four-stage study is 0.97. The distribution of the mean stress level reduction among the study subjects is shown in Figure 4. The estimated values of the mean stress level reduction based on different numbers of stages are indicated in Figure 4. The results show that dividing the study period into multiple stages and incorporating more sub-goals for micro-level feedback potentially help subjects achieving better stress reduction.
Discussion

There exist needs to propose new methods that can support the building of simple but interpretable decision making models and the discovery of optimal sub-goals so as to facilitate decision-making processes, instead of dictating interventions to subjects. In addition, a concern frequently raised for applying behavioral coaching is often that the evidence may not adequately reflect individual differences among users. In the past, psychologists and behavioral scientists have developed a variety of ideographic approaches for single-case experimental designs [39]. To extend the ideographic approach in experimental designs of intervention, researchers have further developed N-of-1 trials to help patients make health decisions that are informed by highly relevant, evidence-based information [40, 41, 42]. In this paper, we extend this school of thoughts to investigate how to further tie the insights learned from user-generated health data into an adaptive decision-making process. In particular, we develop a MTQL method to learn how to curate effective micro strategies for recommendations, with an implicit consideration of user preferences and barriers. The proposed method uses a simple form of regression models to approximate Q functions, and constructs an interpretable form of policies using the Q-learning algorithm. The policy learned using the MTQL method provides sub-goal setting options, which may affect the intervention process. In many other domains, such as tutoring, micro-level strategies have been validated to be beneficial in pragmatic settings [43, 44]. The proposed method in this paper provides a novel way to start exploring a more systematic monitoring and self-experimentation framework to derive sequential micro-level strategies for behavioral coaching.

Conclusion

In this paper we described a MTQL method to estimate the optimal policy in order to maximize the mean stress reduction for healthy adults over a four-week period. We implement the MTQL method on the stress data to illustrate the interpretability of the estimated policies and the way to provide recommendations based on the policies. The insights we gain from the stress data are three-fold: First, tracking whether intermediate stress reduction is achieved and what actions have been taken by a target user are important; the observations will affect the choice of different strategies of stress reduction in a longer term. Second, individualized stress management can benefit from the agility introduced by micro-level strategies. When we assume a finer-grained observation time unit, the system derives more effective micro-strategies that in turn lend support to subjects for achieving better stress reduction in a longer term. Moreover, when the system observes that a user has started going astray from his usual path, the system is still capable to find corresponding recommendations for this user for quick adaptation. Instead of insisting on the same recommendation for target users, the proposed approach is expected to be more flexible and effective in pragmatic settings. Last but not least, personal behavioral coaching recommendations have the potential to help managing people’s stress better by enabling more informed decision making. The derived insights are expected to benefit future collaborative care and self-care applications, such as sense making [19] and persuasive reminders [45]. These insights can also be used to provide feedback into adaptive N-of-1 trials [40, 41, 42] and self-experimentation [46]. Our work is a first step towards a behavioral coaching system, which can help consume systematic monitoring

Figure 4. Histogram of the mean stress level reduction from baseline over a four-week period. The solid line is the mean stress level reduction following the estimated policy of the one-stage study; the dotted line is the one following the estimated policy of the two-stage study; the dashed line is the one following the estimated policy of the four-stage study.
data of target users and transform the curated insights into micro-level feedback that can sequentially guide users through the care management process (or self-experimentation when applicable) to obtain the best possible outcome.

Acknowledgement

This study was supported by the following grant: R01 MH109496, R01 NS072127, R01 HL111195 and R21 MH108999.

References

Framing Electronic Medical Records as Polylingual Documents in Query Expansion

Edward W Huang1, Sheng Wang1, Doris Jung-Lin Lee1, Runshun Zhang2, Baoyan Liu3, Xuezhong Zhou4, ChengXiang Zhai1

1University of Illinois at Urbana-Champaign, Urbana, Illinois, USA; 2Guang’anmen Hospital, Beijing, China; 3National Data Center of Traditional Chinese Medicine, Beijing, China; 4Beijing Jiaotong University, Beijing, China

Abstract

We present a study of electronic medical record (EMR) retrieval that emulates situations in which a doctor treats a new patient. Given a query consisting of a new patient’s symptoms, the retrieval system returns the set of most relevant records of previously treated patients. However, due to semantic, functional, and treatment synonyms in medical terminology, queries are often incomplete and thus require enhancement. In this paper, we present a topic model that frames symptoms and treatments as separate languages. Our experimental results show that this method improves retrieval performance over several baselines with statistical significance. These baselines include methods used in prior studies as well as state-of-the-art embedding techniques. Finally, we show that our proposed topic model discovers all three types of synonyms to improve medical record retrieval.

Introduction

The Obama administration made electronic medical records (EMR) a high-priority initiative, devoting significant amounts of resources to improve their adoption rate in healthcare practices1. As EMR databases are further introduced into daily usage, retrieval systems are increasing in importance. In particular, one important application of an EMR retrieval system is to efficiently parse medical records and identify those that are most relevant to a new patient. Standard information retrieval systems receive string queries as input, compute numeric scores that determine how well each database document matches the query, and output a ranked list of documents. Ideally, a doctor can query a system with a new patient’s symptoms and receive a set of relevant patient records. These records can serve as an informative baseline to prescribe suitable treatments for the new patient.

However, due to synonyms that occur in the medical vocabulary, the original search query may not be complete, and thus may not retrieve optimal results. There are three categories of synonyms of interest to medical record retrieval:

1. **Semantic synonyms** are medical terms that have identical meanings. For instance, “halitosis” and “fetor oris” are semantic synonyms because they are different terms that refer to the same symptom. Because doctors will typically record only one of these, queries that contain “halitosis” will not properly match records that contain “fetor oris”, and vice versa. Semantic synonyms can be mined with natural language processing techniques and straightforward statistical measures2.

2. **Functional synonyms** are terms that are not identical, but co-occur more frequently than random. “Arthritis” and “hypertension” are functional synonyms due to their comorbid relationship (a decade-long study showed that nearly half of elderly arthritis patients also suffered from hypertension)3. Thus, if a hypothetical query consists of only the term “arthritis” to describe an elderly arthritis patient with hypertension, the retrieved patient records may not contain treatments optimally suited for the query patient. Functional synonyms can be inferred from treatment synonyms, which are described next.

3. **Treatment synonyms** are drug-symptom pairs in which the drug treats the symptom. For example, “ibuprofen” and “fever” are treatment synonyms. Treatment synonyms can be obtained by mining medical publications4.

If a query consists entirely of symptoms, then semantic and functional synonyms are also symptoms. We show that augmenting an original patient query with all three synonym types can capture relevant but mismatched documents, thus improving retrieval performance.
In prior work, Zeng et al. performed query expansion in a similar medical record retrieval setting using synonyms and topic models. Their synonym-based query expansion utilized the Unified Medical Language System (UMLS), which is a compendium of biomedical vocabularies, to map query terms to their semantic synonyms. We refer to this method as the dictionary-based query expansion to avoid confusion. On the other hand, their topic-model-based query expansion trained on patient records to jointly find all three synonym types. However, using this standard topic model, symptoms and treatments are grouped together with no distinction in the medical records, which may decrease performance.

Rather than jointly mine these three types of synonyms, we separately modeled the symptom and treatment synonyms. Our approach is based on traditional monolingual topic models, but instead views the symptoms and treatments of a medical record as generated by distinct languages. Thus, outputted topics will be aligned across the two languages and will contain synonyms of all three types. This is because symptoms in the same topic are likely to be semantic or functional synonyms, while symptoms and treatments in aligned topics are likely to be treatment synonyms. Synonyms of query symptoms can then be used to augment the original query during retrieval. To the best of our knowledge, our proposed method is the first to model symptoms and treatments as separate languages in electronic medical records. We also compare with two embedding methods that jointly mine all three synonyms.

We evaluated our approach on a traditional Chinese medicine (TCM) medical record collection. We chose this dataset because functionally synonymous symptoms are even more prevalent in the TCM field. Thus, if our method could improve retrieval performance on this dataset, it would also work well for EMR datasets in other domains. We show that our method can improve over baseline methods, as well as state-of-the-art embedding methods, in query expansion.

Problem Formulation

Given a database of patient records \( R = \{r_1, \ldots, r_n\} \), the \( i \)th patient record \( r_i \) consists of a set of diseases \( D_i \), a set of symptoms \( S_i \), and a set of treatments \( T_i \). From this database, we wish to retrieve the set of patient records most relevant to some new patient \( p_{\text{new}} \) who is not in the database. To achieve this, we first reformulate \( p_{\text{new}} \)'s symptoms as a query. Thus, given \( p_{\text{new}} \)'s set of symptoms \( S_{\text{new}} = \{s_1, \ldots, s_j\} \),

\[
Q_{\text{new}} = S_{\text{new}} = \{s_1, \ldots, s_j\}
\]  

By performing query expansion on \( Q_{\text{new}} \), we add query terms to better match relevant records and thus improve the retrieval performance:

\[
Q'_{\text{new}} = \{s_1, \ldots, s_j, q_1, \ldots, q_l\}
\]

Here, \( \{q_1, \ldots, q_l\} \) is the set of expansion terms that are added to the original query. Although the original query \( Q_{\text{new}} \) only contains symptoms, the expansion terms can contain both symptoms and treatments. Expansion terms can be obtained with a variety of methods, which we discuss in the next section.

We hypothesize that the expanded query, \( Q'_{\text{new}} \), will retrieve more relevant documents because in practice, \( Q_{\text{new}} \) is usually not comprehensive. In our medical setting, this is analogous to situations in which the list of symptoms that a doctor identifies in a new patient is incomplete, which may be due to a combination of two major factors.

1. The doctor uses one of many possible synonyms, including semantic, functional, and treatment synonyms, to describe a patient’s condition.
2. The database is incomplete, so a query symptom may simply not appear in existing medical records, resulting in poor query matches.

We expect the first factor to have a larger impact on query quality, particularly due to unique variations of symptoms that are prevalent in TCM.

Methods

With each technique that we used in our experiments, we conducted query expansion, which is a form of pseudo-relevance feedback, to improve retrieval performance. We augmented each query with synonym terms selected by
each method, and then performed the retrieval on the existing database of medical records.

Overall, we used five different methods of query expansion. First, we addressed two baselines used in previous work\(^5\): the dictionary-based query expansion and the topic-model-based query expansion. In our dataset, the dictionary-based method incorporated an external treatment-symptom knowledge graph to add manually curated treatment synonyms. The topic-model-based method trains topics on the patient record database to add expansion terms that co-occur in the same topics as the given query. Although the previous study used a third method, predicate-based query expansion, we did not utilize this method due to a lack of high-quality TCM ontology databases. Furthermore, the predicate-based method was unable to outperform the topic-model-based method in prior work.

Next, we explored two network embedding techniques, Med2Vec and diffusion component analysis (DCA). Med2Vec is an embedding algorithm that learns efficient representations of medical records and concepts by using EMR datasets. On the other hand, DCA performs network embedding on the knowledge graph utilized in dictionary-based query expansion to obtain vector representations of nodes in the graph. The key difference between these two methods is that Med2Vec does not depend on expert medical knowledge, while DCA does.

Lastly, we discuss our method, which mines semantic, functional, and treatment synonyms by considering symptoms and treatments to be separate languages in a topic model.

**Dictionary-Based Query Expansion**

Dictionary-based query expansion utilizes a ground-truth, treatment-symptom TCM dictionary. This dictionary was curated from a TCM textbook containing known relations, interactions, and treatments. For example, the “crow-dipper” herb has multiple entries, treating symptoms from vertigo to breathing difficulties. We constructed a knowledge graph, in which an undirected edge \(\{t, s\}\) indicates that a treatment \(t\) treats a symptom \(s\) in the dictionary. There are 1,995 treatments, 1,635 symptoms, and 27,824 treatment relations in the dictionary, which translated to a total of 3,630 nodes and 27,824 edges in the resulting knowledge graph. There are no treatment-treatment or symptom-symptom edges. To perform query expansion on a query \(Q_{new}\), we add all treatments from the knowledge graph that are directly connected to at least one symptom in \(Q_{new}\).

**Topic-Model-Based Query Expansion**

In prior work, topic-model-based query expansion performed the best in a similar medical record retrieval task in terms of recall and F-measure\(^5\). Specifically, the authors used latent Dirichlet allocation (LDA)\(^7\) to derive topics from their database of EMRs. With LDA, each document is characterized by a mixture of topics. In turn, each topic consists of mixtures of words. In our study, we also used LDA to train topics from the dataset.

After training \(k\) topics, from topic \(i\)’s per-word distribution \(\phi_i\), we refer to the set of 100 words with the highest probabilities as \(H_i\). For a query \(Q_{new}\), we then perform the following multiplication:

\[
\phi'_i = |Q_{new} \cap H_i| \cdot \phi_i
\]

With this operation, we scale each word’s probability in \(\phi_i\) by the number of query terms that are in the top 100 words of \(\phi_i\). Finally, we sum each word’s probabilities across the scaled distributions, \(\sum_{i=1}^{k} \phi'_i\), and receive a new weight for each word. We experimentally chose to identify five topics from our dataset. The top five words with the highest weights were designated expansion terms.

**Med2Vec-Based Query Expansion**

Med2Vec is a state-of-the-art embedding method designed specifically for EMRs\(^8\). It discovers efficient representations of “medical codes” (symptoms and treatments, in the case of our dataset). To learn embedding from patient records, Med2Vec’s optimization function is similar to that of word embedding methods that use the skip-gram model, such as word2vec\(^9\). The authors stated three major reasons for word2vec’s failure to accommodate medical data:

1. Healthcare datasets have unique structures in which the visits are temporally ordered, but the medical codes
within a visit form an unordered set.

2. Learned representations should be interpretable.

3. The algorithm should be scalable to handle real-world datasets of millions of patients.

In particular, the first reason is of greatest relevance to our experiment setting. Med2Vec maximizes the likelihood of observing a medical code (symptom or treatment) given the codes in the same visit. In other words, a medical code’s vector representation predicts its neighboring medical codes. By obtaining vector representations of all medical codes as well as computing their pairwise similarities, Med2Vec jointly discovers semantic, functional, and treatment relationships.

We ran Med2Vec on our training corpus and obtained a set of low-dimensional vector representations for each symptom and treatment in the dataset. Given a query \( Q_{new} = \{s_1, \ldots, s_j\} \), we computed the cosine similarity between each query term \( s_i \)’s Med2Vec representation and every non-query term’s Med2Vec representation. Thus, for each candidate expansion term, we summed \( j \) similarity scores, one for each query term. We took the five terms with the highest score sums as expansion terms.

**DCA-Based Query Expansion**

DCA is a network embedding technique that takes a network as input and then learns low-dimensional vector representations of the input’s nodes\(^{10}\). DCA has been shown to achieve excellent results in learning network structure for gene function prediction\(^{11}\).

DCA ensures that two nodes have very similar low-dimensional representations if they are topologically close in the network. Thus, related medical concepts tend to have similar output vectors. Like Med2Vec, DCA jointly mines all three synonym types. We used the network constructed in dictionary-based query expansion as the input to DCA. After learning vector representations for each node in the network, we computed cosine similarity scores as in Med2Vec-based query expansion, again taking the top five terms with the highest score sums as expansion terms.

**BiLDA-Based Query Expansion**

In our data, symptoms and treatments are labeled and separated in each patient record. We hypothesize that a topic model that explicitly captures this structure will improve performance over standard, monolingual topic models.

Polylingual topic modeling (PLTM) finds latent cross-lingual topics in a multilingual corpus\(^{12}\). These text collections can either be direct translations or theme aligned\(^{13}\). Direct translations occur in sentences of two documents that are meant to be translations of one another. An example of a direct translation is “Romeo and Juliet” in English and Chinese. On the other hand, theme alignment occurs in documents that are not necessarily direct translations, but discuss the same topics in similar sections. An example of theme alignment is the Wikipedia pages on “Romeo and Juliet” in English and Chinese.

Our method considers EMRs to consist of two separate “languages”: symptoms and treatments. Thus, patient records are theme aligned in the sense that a patient’s symptoms and treatments are generated by the same set of diseases. Furthermore, the symptoms and treatments are varied according to the same syndromes, which are latent factors not explicitly stated in the records. Standard monolingual topic models are unable to represent these separate “languages”, since symptoms and treatments are grouped together. This removes the ability to differentiate, and therefore translate, between the two.

The output of the PLTM is a set of cross-lingual topics, including per-document topic distributions and per-topic word distributions in each of the languages. This model assumes that each topic consists of a discrete distribution of words for each language. Thus, there are two language-specific topics, \( \Phi^S \) and \( \Phi^T \), each of which is drawn from its own symmetric Dirichlet distribution with parameters \( \beta^S \) and \( \beta^T \), respectively.

Next, we discuss the generative process. Each EMR is represented as a mixture over topics, and is generated by first sampling from an asymmetric Dirichlet prior with concentration parameter \( \alpha \) and base measure \( m \):
Figure 1: The plate notation of our proposed model, framing electronic medical records as bilingual documents. A variable’s superscript $S$ indicates symptoms and $T$ indicates treatments. $\alpha$ and $\beta$ are the parameters of the Dirichlet priors on the per-document topic distributions and the per-topic word distributions, respectively. $\theta_d$ is the topic distribution for document $d$. $\phi^S_k$ and $\phi^T_k$ are each language’s corresponding word distributions for topic $k$. $z$ is the latent topic assignment for each observed word $w$.

$$\theta \sim \text{Dir}(\theta, \alpha m)$$  \hspace{1cm} (4)

Then, a latent topic assignment is drawn for each word in the corresponding “language” (i.e., symptoms and treatments).

$$z^S \sim P(z^S | \theta) = \prod_r \theta_{z^S_r}$$  \hspace{1cm} (5)

$$z^T \sim P(z^T | \theta) = \prod_r \theta_{z^T_r}$$  \hspace{1cm} (6)

The individual symptoms and treatments are then drawn using language-specific topic parameters.

$$w^S \sim P(w^S | z^S, \Phi^S) = \prod_r \phi^S_{w^S_r | z^S_r}$$  \hspace{1cm} (7)

$$w^T \sim P(w^T | z^T, \Phi^T) = \prod_r \phi^T_{w^T_r | z^T_r}$$  \hspace{1cm} (8)

With two languages, PLTM reduces to Bilingual Latent Dirichlet Allocation (BiLDA) (Figure 1). We obtained $k$ joint topics that align $k$ symptom topics and $k$ treatment topics. As with monolingual LDA, we experimented with different values of $k$, finding $k = 5$ to yield the best results. To train topics with BiLDA, we used the MAchine Learning for Language Toolkit (MALLET)\textsuperscript{14}, which performs inference with Gibbs sampling. We conducted query expansion the same way we performed LDA-based query expansion, selecting the five terms with the highest sums.

Evaluation

We evaluated and compared the five different query expansion methods, as well as the baseline with no query expansion, by performing retrieval on our dataset. We first describe our dataset, then discuss the evaluation process in the following two sections.

Data Description

We used a large, real-world EMR database containing 7,553 anonymous patient visits, obtained from the department of gastroenterology at Guang’anmen Hospital in Beijing, China. The same doctor treated all patients in the record,
which has the advantage of consistent treatment, but the simultaneous disadvantage of potentially systematic errors or incompleteness. All patients had some variety of stomach illness. Each record contains a detailed list of symptoms, treatments, and diseases. Each patient had an average of 9.08 symptoms and 1.63 diseases. The disease information was used as ground truth labels in the evaluation stage, and was therefore not included when finding query expansion terms. We elected to use only the first visit for each patient to prevent cases in which a patient’s query returns other visits of the same patient. This left us with 3,750 patient visits.

Cross-Validation

We split our dataset into ten random training and test sets as per $k$-fold cross-validation. Thus, the training records were functionally a database of EMRs. Each held-out test patient was then regarded as a new, unseen patient. For each of the test patients, we retrieved relevant patient documents from the training set.

Each test set contained 375 patient records. We excluded all details from the test set except for symptoms. Using a given test patient’s symptom set as a query, we performed each query expansion method in three ways: adding symptoms, adding treatments, and adding both. We refer to these methods as “symptom expansion”, “treatment expansion”, and “mixed expansion”, respectively. The only exception to this was the dictionary-based query expansion, which is only capable of treatment expansion.

Retrieval Tests

For each query in the held-out test set, we performed medical record retrieval. To score a document in the training corpus given a query patient, we used Okapi BM25 as our ranking function, which is one of the most effective retrieval methods. The Okapi BM25 score of a document $D$ given a query patient $Q = \{q_1, \ldots, q_n\}$ is defined as

$$\text{Score}(D, Q) = \sum_{i=1}^{n} \text{IDF}(q_i) \cdot \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \cdot \left(1 - b + b \cdot \frac{|D|}{\text{avgdl}}\right)}$$

In our experiments, $f(q_i, D)$ was always 1 if $q_i$ appeared in $D$, since no patient record contained duplicate symptoms or treatments. $|D|$ is the length of document $D$, and avgdl is the average document length in the training corpus. For the symptom expansions and the baseline with no query expansion, $D$ contained only symptoms. For treatment and mixed expansions, $D$ contained all symptoms and treatments of the patient. In the absence of parameter optimization, we chose the default values of $k_1 = 2$ and $b = 0.75$. Additionally, the inverse document frequency is defined as

$$\text{IDF}(q_i) = \log \frac{N - n(q_i) + 0.5}{n(q_i) + 0.5}$$

where $N$ is the total number of documents in the training corpus, and $n(q_i)$ is the number of training documents containing the term $q_i$. With this ranking function, we returned a ranked list of retrieved documents given a query $Q$.

Relevance Measure

To evaluate the performance of each retrieval task, we assigned an objective measure of relevance to a retrieved patient given a query patient. Conveniently, the list of diseases the doctor assigned to each patient was recorded in our dataset. We used these disease lists as ground truth labels for the corresponding patients. Thus, we define the gain of a document $D$ given a query patient $Q$ to be the following:

$$\text{Gain}(D, Q) = \frac{|D_{\text{disease}} \cap Q_{\text{disease}}|}{|D_{\text{disease}}||Q_{\text{disease}}|}$$

Here, $D_{\text{disease}}$ and $Q_{\text{disease}}$ refer to the set of diseases belonging to $D$ and $Q$, respectively. In the traditional vector space model, this gain is the cosine similarity between the document and query vectors, which is a useful metric for determining similarity between two documents. Thus, we can use normalized discounted cumulative gain (NDCG),
Table 1: Retrieval results for various query expansion methods. Bolded values indicate the highest NDCG@k. BiLDA mixed expansion performed best for all choices of k.

<table>
<thead>
<tr>
<th>Expansion Method</th>
<th>NDCG@5</th>
<th>NDCG@10</th>
<th>NDCG@15</th>
<th>NDCG@20</th>
</tr>
</thead>
<tbody>
<tr>
<td>No query expansion</td>
<td>0.1673</td>
<td>0.1675</td>
<td>0.1674</td>
<td>0.1677</td>
</tr>
<tr>
<td>Dictionary</td>
<td>0.1633</td>
<td>0.1647</td>
<td>0.1652</td>
<td>0.1659</td>
</tr>
<tr>
<td>LDA (symptoms)</td>
<td>0.1686</td>
<td>0.1682</td>
<td>0.1681</td>
<td>0.1690</td>
</tr>
<tr>
<td>LDA (treatments)</td>
<td>0.1689</td>
<td>0.1669</td>
<td>0.1667</td>
<td>0.1677</td>
</tr>
<tr>
<td>LDA (mixed)</td>
<td>0.1690</td>
<td>0.1671</td>
<td>0.1668</td>
<td>0.1679</td>
</tr>
<tr>
<td>Med2Vec (symptoms)</td>
<td>0.1636</td>
<td>0.1637</td>
<td>0.1648</td>
<td>0.1652</td>
</tr>
<tr>
<td>Med2Vec (treatments)</td>
<td>0.1682</td>
<td>0.1673</td>
<td>0.1684</td>
<td>0.1677</td>
</tr>
<tr>
<td>Med2Vec (mixed)</td>
<td>0.1678</td>
<td>0.1671</td>
<td>0.1685</td>
<td>0.1675</td>
</tr>
<tr>
<td>DCA (symptoms)</td>
<td>0.1518</td>
<td>0.1534</td>
<td>0.1556</td>
<td>0.1560</td>
</tr>
<tr>
<td>DCA (treatments)</td>
<td>0.1689</td>
<td>0.1702</td>
<td>0.1712</td>
<td>0.1719</td>
</tr>
<tr>
<td>DCA (mixed)</td>
<td>0.1510</td>
<td>0.1537</td>
<td>0.1557</td>
<td>0.1565</td>
</tr>
<tr>
<td>BiLDA (symptoms)</td>
<td>0.1709</td>
<td>0.1706</td>
<td>0.1713</td>
<td>0.1716</td>
</tr>
<tr>
<td>BiLDA (treatments)</td>
<td>0.1684</td>
<td>0.1681</td>
<td>0.1681</td>
<td>0.1679</td>
</tr>
<tr>
<td>BiLDA (mixed)</td>
<td><strong>0.1752</strong></td>
<td><strong>0.1739</strong></td>
<td><strong>0.1747</strong></td>
<td><strong>0.1736</strong></td>
</tr>
</tbody>
</table>

A standard method of evaluating search engines,[17] to compute the quality of our ranked list. The DCG at a particular rank k, for query Q, which returns a ranked list of D_1, ..., D_N is defined as

$$DCG@k = \sum_{i=1}^{k} \frac{Gain(D_i, Q)}{\log_2(i + 1)}$$  \hspace{1cm} (12)$$

where Gain(D_i, Q) is defined in Equation 11. NDCG@k is defined as DCG@k divided by the DCG of the ideal ranked list for query Q, thus making it a metric comparable across queries and suitable for our 10-fold framework. We show results for k ∈ {5, 10, 15, 20}. We exclude precision, recall, and the F-measure due to their inability to incorporate rankings.

Results and Discussion

The results of the evaluation are shown in Table 1. In order to analyze the significance of the NDCG values, we performed the paired t-test on the ranking metrics between pairs of expansion methods.

BiLDA-based mixed query expansion achieved the best retrieval performance among all expansions. For NDCG@5, 10, 15, and 20, it performed better than the baseline with no query expansion with p-values of 2.842 × 10^{-3}, 4.784 × 10^{-3}, 6.852 × 10^{-7}, and 1.929 × 10^{-6}, respectively. Furthermore, BiLDA mixed expansion performed better than all of the runner-up methods at the 5% significance level.

Mixed expansion was only the best-performing expansion type for BiLDA. This is due to the fact that all other methods do not separately mine symptom and treatment synonyms. On the other hand, BiLDA-based query expansion considers symptoms and treatments to be from separate topics, and therefore it successfully added in mixed query terms.

Dictionary-based expansion’s poor performance can be explained by the fact that it adds too many treatment synonyms, which dilutes the original query’s symptoms. On average, dictionary-based expansion nearly doubled each query in size.

We show an example of a mixed query expansion from BiLDA. In the patient query in Table 2, the five expansion terms include three symptoms and two treatment herbs. Fluttering pulse was an expansion term for this patient, and is indeed an indicative symptom of the patient’s disease, chronic gastritis.[18] Dark, red tongue is a functional synonym of yellow, greasy tongue coating, both of which commonly appear in patients with chronic gastritis.[19] Fullness is a
Table 2: Example of an expanded query created by the BiLDA method. Different query terms are separated by semicolons.

<table>
<thead>
<tr>
<th>Disease Label</th>
<th>Original Query</th>
<th>Expansion Terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chronic gastritis</td>
<td>yellow, greasy tongue coating; epigastric chills; heartburn; bloating; belching; stomachache; acid reflux; dry mouth</td>
<td>fluttering pulse; dark, red tongue; fullness; bitter orange; crow-dipper</td>
</tr>
</tbody>
</table>

semantic synonym of bloating. Bitter oranges are known to treat abdominal bloating\textsuperscript{20}, and are furthermore known to treat chronic gastritis\textsuperscript{21}. Lastly, crow-dippers are also known to treat bloating in chronic gastritis patients\textsuperscript{22}. Indeed, crow-dipper was actually prescribed to this particular patient.

Related Work

Zeng et al. performed a study of synonym, topic model, and predicate-based query expansions. They used monolingual LDA as their choice of topic model and determined it to be the best-performing method\textsuperscript{5}. Our work builds upon their study in the context of traditional Chinese medicine, while also comparing additional methods, showing BiLDA to be an even more effective method. A major difference between their work and ours is that while both systems aim to return the most similar patients to a query, their experimental queries consisted of a single primary disease (“PTSD” and “diabetes”), while our query consists of the complete set of symptoms per patient. Furthermore, we refine the choice of evaluation from traditional measures of precision, recall, and $F_1$ to the more comparable metric of NDCG@$k$. Choi et al. developed a method of learning efficient representations of medical codes, Med2Vec, which we used as one of this study’s expansion methods\textsuperscript{8}. In a previous work, we also performed DCA on a prior knowledge dictionary, but in the context of patient record matrix enrichment for subcategorization of TCM syndromes\textsuperscript{6}. Jain et al. also performed medical record retrieval with query expansion on a patient’s symptoms\textsuperscript{1}. However, they used a knowledge base by integrating domain ontologies and automatic semantic relationship learning, similar to Zeng et al.’s predicate-based query expansion. Due to the lack of TCM ontologies, this method was infeasible.

Conclusions and Future Work

In this paper, we studied how medical record retrieval can improve with query expansion. Prior work showed topic-model-based query expansion to perform the best\textsuperscript{5}. We presented an improved topic model that frames symptoms and treatments as distinct languages.

We performed query expansion on EMR retrieval experiments with latent Dirichlet allocation (LDA), a treatment-symptom dictionary, Med2Vec, diffusion component analysis (DCA), and a polylingual topic model. LDA and dictionary synonyms were studied in prior work and thus used as baselines in this paper. Med2Vec is an EMR-specific embedding approach that learns interpretable representations of medical concepts. DCA is a network embedding method that incorporates prior TCM knowledge to also learn low-dimensional representations of medical concepts. Our experimental results showed that our method performs the best by normalized discounted cumulative gain, with significant $p$-values computed by paired $t$-tests.

Future work includes experimenting with other methods of query expansion. For instance, pointwise mutual information (PMI) has shown promising treatment-symptom pairings. Another potential method is the Weighted Exclusivity Test (WExT), which computes triples of medical concepts as an extension to PMI\textsuperscript{23}.

Lastly, a fundamental change to our problem would be to reframe the retrieval task as a treatment recommendation system. Like before, the system would take a test patient’s set of symptoms as the input query. However, instead of retrieving patient records relevant to the query, the system would recommend a set of drugs to treat the query symptoms. We can evaluate the new system by counting the number of recommended treatments that match the actual prescribed treatments for the test patient. With this framework, we can skip the step in which the doctor analyzes the set of returned patients in the retrieval task and instead directly recommend treatment. In fact, the embedding and knowledge graph-based methods, in addition to PMI and WExT outputs, already have explicit treatment-symptom relationships that would enable this new framework.

947
Acknowledgments

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE-1144245. This research was also supported by grant 1U54GM114838 awarded by NIGMS through funds provided by the trans-NIH Big Data to Knowledge (BD2K) initiative (www.bd2k.nih.gov). Our work also received support from National Key Technology R&D Program grants 2013BAI02B01 and 2013BAH06F03, National S&T Major Project of China grant 2013BAI13B04, and Special Research Project of TCMS by State Administration of Traditional Chinese Medicine grant 201407001.

References

Extracting Actionable Recommendations for Modifying Enterprise Order Set Templates from CPOE Utilization Patterns
Nathan C. Hulse, PhD1,2, Jaehoon Lee, PhD1,2
1Intermountain Healthcare, Salt Lake City, UT; 2Department of Biomedical Informatics, University of Utah, Salt Lake City, UT;

Abstract
As part of an enterprise-wide rollout of a new EHR, Intermountain Healthcare is investing significant effort in building a central library of best-practice order sets. As part of this effort, we have built analytics tools that can capture and determine actionable opportunities for change to order set templates, as reflected by aggregate user data. In order to determine the acceptability of this system and set meaningful thresholds for actual use, we extracted recommendations for additions, removals, and change in initial order selection status for a series of thirteen order sets. We asked local clinical experts to review the changes and classify them as acceptable or not. In total, the system identified 362 potential changes in the order set templates and 186 were deemed acceptable. While further enhancement will sharpen the efficacy of the intervention, we expect that this type of utility will provide useful insight for content owners.

Introduction
Multiple studies in the medical informatics literature that have characterized CPOE system implementations have identified order sets, or standardized groupings of orders targeted to specific clinical conditions or scenarios, as a determining factor in driving user acceptance of CPOE1,2. Physicians often find benefit from using them in that the system can provide logical, pre-populated views of commonly grouped orders for routine procedures and scenarios that they face in the course of their work3. Users often find them quick, convenient and thoughtful in helping them remember all of the nuanced details that can accompany the ordering process for a clinical situation. In similar fashion, organizations have a unique opportunity through order set creation/customization to create value and standardize care by ‘making it easy to do the right thing’ in treating patients. They can build into these templates the kinds of practice elements that represent current best practice4-6. Order set templates often drive usage and adherence, potentially leading to lower overall variation in clinical care processes7.

These benefits, however, come with tradeoffs. Researchers have cited the difficulties of building up a comprehensive library of order sets1, keeping it current as medical knowledge evolves, and the management difficulties that can come with accommodating highly personalized derivatives of order set templates (sometimes known as ‘favorites’ or ‘personal order sets’)8. Clinical knowledge management systems are intended to help organizations identify, centralize, evaluate, refine, and share knowledge resources throughout a given company. Knowledge management strategy as it relates to Computerized Physician Order Entry (CPOE) systems poses a particularly complex challenge. A wide variety of knowledge components must be accounted for, ranging from orderable item libraries, drug databases, order set definitions, decision support rule bases and content dependency trees.

Compiling a centralized knowledge base of order sets has been described as a labor-intensive process. The amount of effort associated with building and maintaining this type of content has led previous research groups to explore methods of automating the creation and refinement of this content6,11. Whether an organization opts to purchase and customize third-party content or develop the order set content in-house, there are significant investments of time and expertise necessary to make it work. Not only are there up-front tasks to populate the catalog and customize it to local practice and availability, but there are ongoing costs associated with the effort such as a) regular review and update existing content to bring it into alignment with current knowledge, b) additions to the catalog to address gaps in coverage and c) removal of obsolete or low-utility content. Researchers have noted that with the exponential growth in medical knowledge, that there is a real risk that once implemented, order sets may be inadequately maintained; in essence, driving caregivers to practice outdated medicine on a widespread basis12. Experts have recommended standard, regular review processes before initial implementation of order set content, but also consistent, periodic review that involves the clinical sponsors as well as allied clinical services (nursing, laboratory, radiology, etc)13. In these reviews, committees should analyze variance from established clinical standards, consistency of care recommendations across order sets, potential for adverse events, and establish priorities for associated clinical decision support rules.
Background

Intermountain Healthcare is a not-for-profit integrated delivery network that serves the populations of the Intermountain West (Utah and southern Idaho). It has 22 hospitals, over 150 clinics, a medical group of over 700 employed physicians and an insurance plan that serves the needs of the people in the region. Collectively, it accounts for roughly half of the healthcare given in the region and insures a little more than a quarter of the population.

Intermountain Healthcare has an established legacy of informatics excellence, dating back to some of the earliest efforts at building electronic medical records in the United States. Recently, Intermountain has opted to transition over to a new EHR offering, in partnership with the Cerner Corporation. Our enterprise rollout of the Cerner EHR product is approximately two thirds complete, scheduled for completion by the end of the year. The change in strategy from building our own EHR to using one provided by a vendor has necessitated shifts in our clinical knowledge management strategy. In particular, many of the content authoring tools have shifted from homegrown solutions to that of using content builders and wizards afforded by the vendor. The two most immediate priorities that have come from the decision related to knowledge management are that of building up and providing visibility into usage patterns surrounding an enterprise library of order sets and decision support rules. In this manuscript, we will focus on the efforts of the former, specifically tooling and processes surrounding order set management.

The effort to create an enterprise repository of order sets has been conducted under the organizational structure of Intermountain’s Clinical Programs. These groups cover 10 specific domains of medical care at Intermountain including cardiovascular, intensive medicine, pediatric, and oncology. Clinical Programs have been in place at Intermountain since 1997 and have a primary charge to identify, develop, and deploy best practice protocols in the enterprise. Among the chief roles that they have is to set and drive towards specific clinical performance targets, and create enterprise content that supports these initiatives. As Intermountain has worked to prepare for and roll out a new EHR platform built on a vendor-supplied infrastructure, care teams of clinical champions, data managers and quality experts have collaborated in building out a library of over 2,000 order sets. (an example is given in Figure 1 below) Most of these order sets have been iterated upon and refined multiple times in the course of preparation and actual usage. This library continues to grow and iterative enhancements are regularly incorporated into the content as stakeholders give feedback and as the content adapts to changes in understanding of current best practice. The maintenance burdens that come with building and maintaining a library of this size are significant. At Intermountain, well over 100 authors have been personally involved in the creation and maintenance of this content, with hundreds of others giving feedback directly or in committee as to enhancement and refinement.

Figure 1- Order set example. This particular order set focuses on the treatment of dehydration in infants.
Rationale

As part of this effort, business and Clinical Programs leadership have prioritized the development of tools that give visibility into the order set catalog. Specific questions around the catalog, associated metadata (authors, review teams, publication dates, sponsoring departments) and usage patterns have been the focus of a tool known as DOT (Dashboard for Order set Transactions). The development and usage of this product has been previously described\(^\text{15}\).

As authors grew accustomed to this very granular level of visibility into the content usage and variation related to their order sets, they became increasingly interested in specific scenarios in which the prescribed order sets content and user behavior didn’t align well. In particular, four scenarios were identified in which they wanted the system to assist them by identifying particular opportunities for either change to the order set template or active education and change management processes to redirect user behavior. These included:

1. **Change to Unselected**: An order is pre-selected in the order set template (it shows up automatically selected upon loading) but users are actually ordering it infrequently. This would identify scenarios in which users are actively going out of their way to not order preselected content.
2. **Change to Preselected**: An order is initially unselected in the order set template, but a majority of users end up selecting and ordering it anyway.
3. **Additions**: An order that is not in the order set template at all is frequently added to the order list as part of the ordering encounter. This may identify scenarios in which thoughtful and clinically useful additions might be made to the order set template.
4. **Removals**: Orders that are in the order set template are never or very seldom used. Given that users can always revert back to looking up any individual order from the orderable item catalog, there is some benefit in considering removals from the order set template if they are seldom used. This can optimize users’ experience with the order set by reducing the number of items that they need to scan and process during the ordering episode.

While there are clearly other elements to utilization of order sets that could be approached in this type of optimization effort (order sentence detail, organization and co-location of order elements, etc.) these particular emphases were agreed upon as initial priorities. The overall effort behind this type of utility would be to reinforce the clinical review processes intended for systematic review of content with detailed, actionable feedback from the system that reflects usage patterns. Additionally, clinical sponsors recognized that in building this type of tool, not all feedback would result in template change. Ordering patterns that deviated from expected care might be just as useful for identifying specific areas where clinical leadership may need to build specific goals and education plans to bring ordering patterns closer to expected levels.

The process of navigating and using an order set has been previously described as a collective mathematical sum of both cognitive and physical costs.\(^\text{16}\) Cognitive costs involve the reading, scanning, and logical processing of the recommendations given by an order set. Physical costs involve the manual efforts of clicking checkboxes, dropdown menus, selection, and scrolling through the content itself to fully review and utilize it. In essence, the process of order set optimization is intended to account for both optimizing these costs while maintaining the clinical and operational effectiveness and utility that come from order set usage.

Methods

**Report Generation** – In order to balance these needs, we decided to build a content advisor utility that analyzes order set templates from the knowledge repository, associated usage data from the enterprise data warehouse and makes active recommendations for change along the four change ‘axes’ described above. Order set template data was extracted from Cerner’s database tables to Intermountain’s knowledge repository and organized into logical groupings by version. In similar fashion, encounters for which these order sets were used in clinical care were identified, and the orders derived from these sessions were grouped and aligned against the order set template. Under advisement from clinical partners, we set four initial thresholds for identifying actionable change items in the content.

1. Orders would be flagged as ‘candidates for unselected status’ if they were initially selected in the order set template, yet ordered in less than 50% of the encounters that used that order set template.
2. Orders would be identified as ‘candidates for preselected status’ if they were initially unselected in the order set template, yet ordered in more than 50% of the encounters that used that order set template.
3. Orders would be identified as ‘candidates for addition’ if they were not in the order set template at all, yet were added ‘ad hoc’ by ordering providers more than 5% of the time.

4. Orders would be identified as ‘candidates for removal’ if they were in the order set template, yet ordered 1% or less of the time.

We built database scripts that align the data from the sources described above, identifies order elements that match these thresholds and organized four lists in a document that details all recommended content changes per order set. These documents contain four sections, and in each section candidate orders for preselection, deselection, addition, and removal are detailed, along with the corresponding data that was intended to direct the users attention to the specific details of how the orders were used. In particular, the reports showed:

- Order set title and version
- Number of overall ordering instances for which the specific version of the order set template was used
- Orderable title
- Percentage of overall ordering instances for which the particular orderable item was used
- Actionable recommendation (change selection status to preselected/deselected, addition, removal)

**Report Validation** – In order to test the utility of these reports and understand whether these initial thresholds were optimal for this purpose, we sampled a series of thirteen order sets relevant to three separate clinical programs, and asked content and data managers to review the recommendations. These order sets were chosen as order sets of interest by three Intermountain Clinical Programs and are listed in Table 1 below.

**Report Selection & Creation Criteria** – This type of feedback mechanism is dependent upon a minimal threshold of content being in place. As the overall number of encounters for which a specific version of an order set increases, so does the sampling by which the recommendations are representative of general use. As a rule, we opted to analyze the most current version of the content in the system, unless there were fewer than 20 instances of use of the order set template. In this circumstance, we reverted to the version immediately prior the current version for analysis. Table 2 contains the # of ordering instances that correspond with each order set included in this analysis. It is noteworthy to mention that while it might be expected that the number of ordering instances of a hip replacement and hip replacement post-operative order set would be expected to be the same, the data does not bear this premise out. This is primarily due to the fact that these order sets are independently maintained and updated in the database. As such, some of the content was updated during this timeframe, affecting the overall number of ordering instances per version. The numbers reflected below are specific to one single version of the content in the database.

<table>
<thead>
<tr>
<th>Musculoskeletal Clinical Program</th>
<th>Pediatric Clinical Program</th>
<th>Critical Care Clinical Program</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Order Set Title</strong></td>
<td><strong>Order Set Title</strong></td>
<td><strong>Order Set Title</strong></td>
</tr>
<tr>
<td>Hip Replacement</td>
<td>89</td>
<td>Dehydration</td>
</tr>
<tr>
<td>Hip replacement post op</td>
<td>40</td>
<td>Asthma</td>
</tr>
<tr>
<td>Knee replacement</td>
<td>130</td>
<td>Bronchiolitis</td>
</tr>
<tr>
<td>Knee replacement post op</td>
<td>90</td>
<td>PED General Admission</td>
</tr>
<tr>
<td>Shoulder replacement</td>
<td>35</td>
<td></td>
</tr>
<tr>
<td>Shoulder replacement post op</td>
<td>20</td>
<td></td>
</tr>
<tr>
<td>Sepsis ICU Admission</td>
<td>201</td>
<td>General Critical Care Admission</td>
</tr>
<tr>
<td>Partial Critical Care Admission</td>
<td>440</td>
<td>Respiratory Failure / Pneumonia</td>
</tr>
<tr>
<td>PED General Admission</td>
<td>659</td>
<td></td>
</tr>
</tbody>
</table>

*Table 1 - Clinical Programs, Selected Order Sets and Corresponding Numbers of Ordering Instances for the Order Sets Analyzed in this Study*
From the metadata associated with each order set in the database, we contacted the main author and content stewards to see if they would be willing to review these system generated recommendations and provide us with feedback about the system-generated recommendations. They were given a spreadsheet in which they were asked to indicate whether they thought that the suggestion was acceptable as is, worthy of discussing in committee (but not yet implementing) or not acceptable. For those deemed not acceptable, we asked them to provide specific feedback as to why they felt that the feedback was not acceptable or useful. We then aggregated the results, tallied the responses, and stratified the results by suggestion type.

Results

Overall, the system made 53 recommendations for preselection, 23 recommendations for deselection, 32 candidate orders as ‘additions, and 255 candidates for removal. Given that both a response of ‘accept’ and ‘would consider in committee’ are positive indicators of how the suggestion was received, we aggregated those groups into a common category indicating that the respondents felt that the suggestion was reasonable. Table 2 contains the results of each of the suggestion categories made by the system, as well as the corresponding ratings from the owners and content managers.

<table>
<thead>
<tr>
<th>Suggestion Type</th>
<th>#</th>
<th>Y</th>
<th>M</th>
<th>N</th>
<th>Total ‘reasonable’</th>
<th>% ‘reasonable’</th>
</tr>
</thead>
<tbody>
<tr>
<td>Change to preselected</td>
<td>53</td>
<td>20</td>
<td>16</td>
<td>17</td>
<td>36</td>
<td>68%</td>
</tr>
<tr>
<td>Change to unselected</td>
<td>23</td>
<td>1</td>
<td>2</td>
<td>20</td>
<td>3</td>
<td>13%</td>
</tr>
<tr>
<td>Additions</td>
<td>31</td>
<td>11</td>
<td>14</td>
<td>6</td>
<td>22</td>
<td>81%</td>
</tr>
<tr>
<td>Removals</td>
<td>255</td>
<td>56</td>
<td>69</td>
<td>130</td>
<td>125</td>
<td>49%</td>
</tr>
</tbody>
</table>

*Table 2 - Suggestion Type Analysis - Total Counts, Ratings and Percentage of ’Reasonable’ Suggestion Numbers are shown*
For the respondents that indicated ‘no’ as to whether or not the recommendation was acceptable, we gathered additional narrative input about their reasons for not wanting to implement the feedback. For purposes of presentation, we grouped the comments according to common themes and aggregated them into counts that reflect those reasons. The reasons that were more commonly given for the various suggested edits are summarized in Table 3.

<table>
<thead>
<tr>
<th>Reason</th>
<th>Change type</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Needs to be a conscious decision by provider because of costs and/or risks</td>
<td>Preselection</td>
<td>3</td>
</tr>
<tr>
<td>Too close to 50/50. I would make the physician think about it before selecting.</td>
<td>Preselection</td>
<td>5</td>
</tr>
<tr>
<td>No consensus</td>
<td>Preselection</td>
<td>3</td>
</tr>
<tr>
<td>Is a part of the standard care and should be ordered</td>
<td>De-selection</td>
<td>8</td>
</tr>
<tr>
<td>Order is facility-specific will only show up in specific locations. Rates are expected to be lower, even though it is selected in the order template</td>
<td>De-selection</td>
<td>6</td>
</tr>
<tr>
<td>Decision Support rule that triggers this action isn’t currently working, resolving that issue will change utilization rates</td>
<td>De-selection</td>
<td>3</td>
</tr>
<tr>
<td>PRN, want to keep</td>
<td>Removal</td>
<td>107</td>
</tr>
<tr>
<td>Want to keep as a reminder</td>
<td>Removal</td>
<td>8</td>
</tr>
<tr>
<td>Keep as advisory item</td>
<td>Removal</td>
<td>5</td>
</tr>
<tr>
<td>Facility-specific item, so lower usage is okay</td>
<td>Removal</td>
<td>2</td>
</tr>
</tbody>
</table>

Table 3 - Common reasons for deeming feedback to change order set elements as ’not acceptable’, along with change suggestion type and corresponding counts

Discussion

The content recommendation system produced a significant number of recommended changes across the order sets, with an average of just under 28 recommended changes per order set in the study. In total, just over 51% of the recommended 362 changes were deemed ‘accepted’ or ‘pending committee review’. Overall the system has exhibited higher rates of acceptable suggestions for suggesting changes to make orders preselected and additions to the order sets. Content owners were less likely to consider removals from the system (closer to 50%) and far less likely to consider unselecting currently-selected orders (only 13% of the time).

Opportunities for goal setting - One of the key findings of this effort is that the orders identified by the system may either be opportunities for change to the order set template itself or outliers from expected behaviors. Several of the orders flagged for ‘deselection’ came as surprises to the content owners and stewards. They were under the impression that users were ordering pre-selected orders close to 100% of the time, only to find out that they were in fact ordering them less than half the time. This serves as an important notice to the content author, even if it does not result in an immediate change to the order template. It represents a disconnect between the authors’ expectations around user behavior and the actual behaviors themselves. While individual reasons are different for each scenario, some possible reasons for these disconnects include a) different priorities between caregivers and content authors b)
presumed workflows that bear out differently in actual practice c) and local workarounds and user treatment behaviors that may deviate from centrally-designated best practice.

**Identification of broken components** - One of the suggestions for order set template change came from unexpectedly low rates for ordering head circumference measurements for pediatric patients. This suggestion showed up in three of the four pediatric order sets reviewed. Upon further inspection of the reasoning behind these low utilization rates, a knowledge engineer identified that an associated decision support rule intended to trigger and drive higher usage of these orders was currently not working. Although not a direct intent of the tool itself, the resultant suggestions from the tool helped to identify a different knowledge artifact in the system that was malfunctioning. Separately, in extracting data for the analysis process, it became clear that some of the order set definitions themselves pointed to orderables that were no longer active in the orderables library. While not considered a specific objective in this study, these particular points could also be brought forward to content authors in order to facilitate cleanup and maintenance of order set content.

**Facility-specific items** – Our recommendation system was not attuned for the fact that some orderable items are configured to only be available in specific locations. These types of orderables often correspond with a given location’s abilities to perform orders based on availability of equipment, specialized personnel, formulary, or other constraints. Since our system did not account for these variations, it sometimes recommended modifications that would not be universally useful in scope.

**Strength of suggestion** – There is some evidence in the data we gathered to indicate that the strength of the content recommendation is stronger with a greater overall ‘denominator’ of order set content usage. Content reviewers felt that a recommendation to remove an order that had not been ordered in a set of a few dozen episodes of care was far less impactful than a recommendation to remove an orderable that had not been ordered from the template from hundreds of order episodes. We may use this feedback to only present ‘candidates for removal with a lower threshold of usage (absolutely no usage) and with a required denominator of at least several hundred ordering instances before the recommendation was made. In similar fashion, recommendations to either preselect or deselect orders near the 50% threshold were less acceptable to users. Going forward, we may require 65-70% usage of a currently unselected order before recommending it for preselection, as users responses to those recommendations were more absolute. Several of the responses near the 50% threshold were deemed as too inconclusive to really make a change either way.

Overall, there appears to be more willingness on the part of content authors to preselect and add than there is to remove and unselect. This may be problematic if not tempered, in that overselecting and overloading any particular order set can make the order set more cumbersome to use. Overuse of preselection status within an order set may lead to overutilization of labs, imaging, and even medications. Our hope is that the approach taken by content authors to feed these recommendations back to a larger committee for review would temper and offset some of these concerns.

**Limitations**

**Size and scale** – In this study, we have only sampled thirteen order sets from our order set library of over 2,000 separate documents. We realize that this is a small sample and that it involves only a small number of reviewers. We hope to incorporate the lessons learned from this early sampler of content into the behavior of the system before rolling it out for broader use among the other Clinical Programs and content owners. Furthermore, our study has some real limitations in that we are bringing in data only from the sites that are currently live on our new EHR platform. Nearly half of the care delivery volumes at Intermountain Healthcare are not yet live on this platform and will undoubtedly impact both the usage numbers but also behavior patterns exhibited in the data set.

**Personalized order sets or ‘favorites’** – In our new EHR implementation, users do have the ability to derive personal order sets with modifications from the enterprise order set templates and save them as ‘favorites’ from which they can order. As of yet, our system does not account for either the definition of these templates, nor the usage patterns derived from their use. We do anticipate including both of these dimensions of content derivation and use in future iterations of the content recommendation tool itself. We feel that there is a rich amount of data stored in the definitions of these personalized order sets and that machine learning techniques will help us to identify clusters of user behaviors in this space.

**Consolidation of feedback recommendations across template versions** – In the current design of the content recommendation tool, we are not accounting for broader usage numbers that could be achieved with cleaner reconciliation of common orders across versions of order sets. While the content differences between versions of an
order set may be either subtle or substantial, we anticipate that we should be able to employ some grouping logic to allow the framework to more gracefully span order set versions and hopefully make recommendations that are more crosscutting as it relates to overall usage.

Next Steps

Going forward, we expect to extend the content recommendation tool to account for recommendations derived from the authoring of and usage surrounding personalized order sets. We expect to account for recommendations that span versions of the document, as detailed above. Further, we intend to pursue similar recommendation patterns for other portions of the knowledge captured in order set definition, including order sentence details (that often contain form elements specific to the orderable), sequence, redundancy (identical or similar orders placed in multiple locations in the template), and decision support integration.

In studying the acceptability of the recommendations themselves, we intend to pursue the recommendations that go to content committees and record how they are received there and whether the content changes eventually make their way into the order set definitions themselves. As described earlier, we plan to use altered ‘thresholds’ for triggering these recommendations, both in terms of the percentages and minimum number of ordering instances needed before the reports can be created.

Finally, we intend to conduct time-based studies to see if these types of changes affect the overall efficiency of our end-users. One of the main objectives of the research is to build a framework that can maximize efficiencies in users’ interactions without compromising the value and clinical quality that order set usage can bring.

Acknowledgments

We would like to acknowledge our colleagues at Cerner for their help in identifying data relevant to the order set templates, corresponding metadata, and the order set instance derivatives created from the use of these plans in the clinical data repository. Additionally, we would like to thank the authors, data managers, and clinical leaders who spent time reviewing suggested edits from the system and providing feedback.

Conclusion

As part of a broader initiative aimed at creation and effective use of clinical knowledge content inside our EHR, we have built a content recommendation system that is capable of identifying opportunities for optimizing order set definitions inside our system. Our preliminary testing has shown that the system produces recommendations that users generally find reasonable, although its performance in recommending additions and preselections is higher than that of deselections and removals. We expect that our efforts in trying to automate and summarize user-derived feedback systems will leave content owners better informed and more empowered to maintain their content in useful ways over time.

References


Replicability, Reproducibility, and Agent-based Simulation of Interventions

R. Stanley Hum, MD, MA1, Samantha Kleinberg, PhD2
1 Columbia University, New York, NY; 2 Stevens Institute of Technology, Hoboken, NJ

Abstract

Secondary use of medical data and use of observational data for causal inference has been growing. Yet these data bring many challenges such as confounding due to unobserved variables and variation in medical processes across settings. Further, while methods exist to handle some of these problems, researchers lack ground truth to evaluate these methods. When a finding is not replicated across multiple sites, it is unknown whether this is a failure of an algorithm, a genuine difference between populations, or an artifact of structural differences between the sites. We show how agent-based simulation of medical interventions can be used to explore how bias, error, and variation across settings affect inference. Our approach enables users to model not only interventions and outcomes, but also the complex interaction between patients with different risks of mortality and providers with different observed and latent treatment effects. Ultimately we propose that such simulations can be used to better evaluate the behavior of new methods with known ground truth and better calculate sample size for EHR-based studies.

Introduction

Electronic health records (EHRs) have made it possible to study large populations over longer timescales than has previously been feasible, yet the observational nature of these data lead to a number of challenges, including difficulties in both reproducing studies and understanding when we should expect results to be replicated in a new setting. While confounding due to unobserved variables is a known challenge for causal inference (and methods for inferring latent variables have been developed to address this), the effect of variation in medical processes and unobserved treatments on false positives and negatives is less appreciated and can lead to the same outcomes as selection bias – even without hidden common causes. The problem is further compounded when we repeat studies in multiple settings, whether as part of a multi-site trial or as an attempt at reproducing a particular finding. Standards of care, documentation, and variation in both treatments and patients may differ, and may all contribute to a failure to reproduce a genuine finding. Studies may also fail to show an effect due to being underpowered, and in some cases effective power may be lower than expected due to latent effects that are not accounted for.

For example, a study may aim to test the role of ventilators in treating pneumonia, but ventilators are combined with antibiotics in some cases, and this choice may depend on the population, a clinician’s background, and the patient’s treatment preferences. While ideally we will have data on all interventions so as to control for such differences, documentation practices and the nature of EHR data mean that such information may be systematically missing. That is, if only text-based data are used, and that is where interventions such as ventilation are documented, while medications are prescribed electronically and stored in a structured database, then work focusing solely on free text will not be able to capture the role of antibiotics.

However, it is difficult to tease apart the effect of each possible source of error and bias from EHR data directly, especially given the lack of ground truth for evaluating results. At the same time, replicability, or the lack thereof, is receiving increasing attention across many fields. We propose that simulations can be used to test algorithms for robustness in the face of these challenges, and to develop solutions to address the unique difficulties of reusing medical data. Here we present the results of an agent-based simulation of medical interventions, showing the individual and compounded impact of hidden interventions and process features. In particular, we examine how combinations of latent and observed interventions may lead to reduced study power, and show how agent-based modeling can help researchers explore these issues.

Background

Reproducibility and causal inference

A key part of the scientific method is that hypotheses should be falsifiable and results should be consistent. A protocol developed by one person should be able to be followed by another and should produce the same results under the same
circumstances. Yet recent articles have been raising concerns that many scientific results cannot be replicated, and that the problem may be getting worse. Attempts by pharmaceutical companies to replicate academic findings led to only 20-25% of findings being repeated and a review of 53 major findings in cancer biology found only 6 were replicable. Recent works such as the Many Labs project have investigated this in social psychology research by attempting large-scale replications. The Open Science Collaboration conducted detailed replications of 100 psychology studies, with only about half being replicable.

While much of the focus has been on the lack of reproducibility in experimental studies, much of the work in biomedical informatics relies on secondary use of observational data. Instead of prospectively collecting data to test a specific hypothesis, this research is primarily data driven, using massive EHR datasets and other sources to generate new hypotheses. Reproducing work is both more important (given the increased possibility for confounding and error in these observational data that were not primarily generated for research purposes) and more difficult in this case. One study looked at 52 claims from observational studies and found none were reproduced in later randomized controlled trials (RCTs). However, one of the challenges of reproducing claims, particularly in biomedical informatics, is the lack of information about experimental protocols. Ioannidis looked at a sample of 441 biomedical articles, finding only one made a full protocol available and none shared their raw data. Raw data often cannot be shared due to privacy and ethics reasons, though some benchmark datasets in related areas such as pharmacovigilance have been shared.

Reproducibility in biomedical informatics has primarily been studied in terms of how portable algorithms for identifying phenotypes are between institutions. However, few studies have been done to evaluate the reproducibility of machine learning results across settings. The difficulty in reproducing studies is further compounded when we aim to infer causal relationships and not simply correlations. Here we must be sure that any effects we observe are not the result of hidden common causes, such as a latent variable causing both a treatment to be selected and outcome to be observed.

One work attempted to replicate a data-driven study of risk factors for congestive heart failure in two EHR systems. However, when using existing, observational, data, there are often many differences between two settings – data formats, population characteristics, medical practice – and it can be impossible to disentangle which of these are responsible for differences in results. As a result, it was both logistically difficult to reproduce the main approach for identifying cases and controls (as different data were collected) and difficult to evaluate whether results were reproduced. Many of the same problems plague multi-site randomized controlled trials (MRCTs), where standardization across sites is critical to finding the true impact of an intervention. Chesworth et al. assessed treatment fidelity for a specific trial, finding that in many cases actual procedures differed substantially from the protocol. Spirito et al. examined the role of differences in sample population and study protocol in a 6-site MRCT, and found that among other factors attrition varied by site and accounted for some outcome differences.

Medical simulations

The studies described in the previous section have shown difficulties in reproducing results between settings, and have retrospectively analyzed reasons for lack of reproducibility for individual experiments. However, the same incorrect finding can be made in two places, and when we are testing a new method, a failure to reproduce does not mean the method is incorrect (as the difference may be due to differences in data or the population). To systematically test how factors such as error in documentation or variation in protocol affect inference algorithms, we need 1) ground truth (so we know what should be found) and 2) variation (systematically changing features of the problem). Simulations allow us to replicate real-world scenarios such as a hospital stay, and enable one to learn about how outcomes would differ under various parameters (e.g. how mortality rate would differ if doctors had less variation in their processes). Importantly, all variables are observed and we have perfect ground truth in the simulation, and can test what happens in other cases (e.g. latent treatment) by removing some variables from the output analyzed.

Simulations have primarily focused on modeling spread of infectious diseases such as influenza, with less work on simulating medical data itself. The Observational Medical Dataset Simulator created simulated longitudinal medical data, using real data as input to determine realistic population characteristics. The approach generates populations where individuals each have a set of diseases and treatment assignments, then simulates the effect of treatment on outcomes. This approach is similar to the one we take, using probabilistic modeling of treatments...
Figure 1: Overview of system. A patient has an underlying physiological state, measured with some error. A provider makes qualitative observations, observes the noisy measurements, and chooses interventions based on these. Interventions affect both underlying physiology and measurements. Finally, documentation is a noisy and incomplete recording of a patient’s status, and affects intervention choice.

Figure 1: Overview of system. A patient has an underlying physiological state, measured with some error. A provider makes qualitative observations, observes the noisy measurements, and chooses interventions based on these. Interventions affect both underlying physiology and measurements. Finally, documentation is a noisy and incomplete recording of a patient’s status, and affects intervention choice.

and outcomes and aiming to capture the confounding that makes inference so difficult. However, OSIM2 models only patients and treatments, and not the bias and latent variables that make causal inference challenging. On the other hand, we aim to capture the heterogeneity and error introduced in the medical process through providers and documentation, and directly model unobserved factors that may affect outcomes.

Methods

In this section we describe the architecture of our simulation, data generation approach, and analysis methodology.

Simulation

We propose an agent-based approach to simulating medical interventions, which can capture the complex interactions between a patient’s state, a provider’s translation of their observations into interventions, and the documentation of these processes (that become input to other cognitive processes). One of the advantages of an agent-based approach is being able to simulate individual characteristics of providers, versus having the same parameters for all. This approach will ultimately allow us to model the interaction between providers (nurses, doctors) over time. An overview of the proposed approach is shown in figure 1. The idea is that we can clearly separate the different sources of error and bias so they can be systematically varied in combination and alone. A patient (purple segment at top) has a true underlying physiological state, which is distinct from measurements of this state (e.g., laboratory tests), which may be error-prone. In the clinician layer (green, middle), a clinician’s assessment of a patient is dependent on their skill and experience, and these plus noisy measurements and documentation guide intervention choices. Thus we can simulate how an error in a note could lead to errors in intervention (e.g., incorrect medication dosage). Finally, EHRs and other documentation are not simply a faithful recording of all events, but depend on skill and experience, and this documentation (orange, bottom layer) feeds back into intervention choice.

In this work we focus on the middle layer in the diagram: modeling provider features. We are interested primarily in features of the medical process that can lead to erroneous inferences, rather than the specifics of any particular illness. Thus we develop an agent-based simulation to model interventions at a high-level, using Repast Simphony, an open-source toolkit for agent-based simulations. To simplify the problem, we focus on the case where each patient has a single medical encounter with a single provider, and can receive (or not) an observable treatment and can also receive (or not) a latent treatment. A latent treatment is one that is not documented (unobserved). The effect of both the latent

1https://repast.github.io/index.html
Figure 2: Detail of simulation process. Individual patients and providers are drawn randomly from the relevant defined distributions. For patients, these are overall risk of death and for providers they are efficacy of interventions. Then, at each time step instantiations of treatment are drawn from the distributions. Patient, provider, and treatment characteristics all modify overall risk rate.

and observed treatments can vary in a provider or site dependent way, enabling simulation of some of the challenges of multi-site RCTs. For example, there may be systematic differences in standard of care or adherence to protocols, which can confound results, such as finding no difference between intervention and control if the intervention protocol has significant variation. While at a large scale such differences are assumed to average out, cluster randomized trials for example have lower effective sample size, as conditions within a site are not independent.

The flow of the simulation is shown in figure 2. There is a set of patient and provider agents, each of whom has their own features drawn from a distribution of distributions. Instead of describing a treatment in terms of only its mean effect and standard deviation (s.d.), we draw the mean and s.d. from distributions. For each encounter between a patient and provider, the intervention effects for latent and observed interventions are chosen from the intervention distributions. Finally, each patient’s outcome is simulated by combining the factors that modify their relative risk (personal risk rate, latent treatment, observed treatment). In this work, the outcome for each patient at the end of the simulation is survival, so each factor modifies risk of mortality.

The key parameter distributions in the simulation are as follows:

- Risk of death (mean and standard deviation)
- Treatment effect mean (mean and standard deviation)
- Treatment effect standard deviation (mean and standard deviation)
- Latent effect mean (mean and standard deviation)
- Latent effect standard deviation (mean and standard deviation)

Again, rather than a single distribution for each treatment parameter, the key parameters themselves vary according to user-defined distributions. This allows groups of patients (risk of death) and groups of providers (treatment and latent effects) to be modeled automatically. For example, if there is one group of providers being studied, such as in a single center study, the treatment effect would draw upon a single mean and a single standard deviation. If there are several groups of providers being studied, such as a multi-center study, then each center would have a treatment mean with an associated standard deviation. We can choose the four variables that affect treatment mean and treatment standard deviation to describe the overall distributions that inform each center’s treatment effect parameters.

We can thus simulate effects which are uniform or highly variable (normally centered around a single mean average effect) and which are evenly or unevenly applied (also normally centered around a single mean standard deviation).
Table 1: Parameters across experiments. The two risk of death mean values shown as high mortality (low mortality in parentheses) are constant across experiments. Brackets indicate iteration with a given step size.

For example, we can simulate an intervention that has minimal variation in its standard deviation at each site, but large variations in its mean effect. The treatment effects and risk of death are expressed as relative risk (RR), which is the probability of death under exposure divided by the probability in a non-exposed group. An RR of 1 leaves risk unchanged, while effects < 1 reduce risk and those > 1 increase it.

The latent treatment effect can be used to express a number of unrecorded or unobserved qualities, such as actual unobserved treatments or factors affecting outcome such as provider skill. For example, a latent effect mean of 1 and standard deviation of 0 mean the provider outcome is exactly the same as that of the treatment. However, one can also imagine cases with mean < 1 but wide standard deviation, meaning that on average the provider’s outcomes improve upon the treatment effect, but that there is high variance.

In reality, many important factors appear as latent for varying reasons. In some cases data may not be recorded or available, such as when using only claims data rather than the full EHR. In other cases, data on interventions may be recorded with a delay, so we do not know when a treatment happened relative to a change in a patient’s condition and cannot use such data to identify a causal relationship. With this simulation we aim to be able to systematically vary such factors, to test how well algorithms can handle these cases and develop new approaches that are more robust.

Data simulated

To demonstrate our approach, we create three types of datasets that illustrate features of our simulation and the difficulties posed for even simple comparisons. In each experiment we simulate 1000 patients, varying the risk of death from the illness itself (high vs. low mortality rate) as well as the efficacy of both the latent and observed treatments. Due to the randomness inherent in the simulation, we run 10 simulations for each parameter setting and average the results. Parameters for the three datasets are shown in table 1. In all cases, risk of death is given as a relative risk. Risk of death is that at each time step, so that for a simulation with $t$ time steps, the actual risk is $(1 - \text{risk})^t$. In this work we use $t = 10$, so the high mortality settings have an overall mortality rate of 40%, across the 10 time steps. The primary outcome of each simulation is the mortality rate, which we compare across simulations, aiming to replicate what happens in a multisite trial. We aim to determine under what conditions we can find statistically significant differences between groups, and how much unmeasured latent effect is required to draw erroneous conclusions.

Varying latent treatment effect: We begin with the common case of determining whether there is a difference in mortality rate between two conditions. In this example we fix the treatment as having no effect on risk of death, and compare two population risk of death means (0.05 for high mortality, 0.005 for low mortality) while varying how effective the latent treatment is (iterating from 0.1 to 1.0 by 0.1). This case can occur when there are unknown risk factors for a disease, such as environmental exposure, that differ between sites in a multi-site trial.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>No Tx effect, vary latent</th>
<th>Low/Med/High Tx effect, vary latent</th>
<th>Med Tx, vary latent s. d.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Risk of death mean</td>
<td>0.05 (0.0005)</td>
<td>0.05 (0.0005)</td>
<td>0.05</td>
</tr>
<tr>
<td>Risk of death s. d.</td>
<td>0.003</td>
<td>0.003</td>
<td>0.003</td>
</tr>
<tr>
<td>Tx effect mean (mean)</td>
<td>1</td>
<td>0.9, 0.5, 0.3</td>
<td>0.5</td>
</tr>
<tr>
<td>Tx effect mean (s. d.)</td>
<td>0</td>
<td>0.03</td>
<td>0.03</td>
</tr>
<tr>
<td>Latent effect mean (mean)</td>
<td>[0.1:0.1:1]</td>
<td>[0.1:0.1:1]</td>
<td>0.7</td>
</tr>
<tr>
<td>Latent effect mean (s. d.)</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Latent effect s.d. (mean)</td>
<td>0.01</td>
<td>0.01</td>
<td>[0.5:0.5:2.5]</td>
</tr>
<tr>
<td>Latent effect s.d. (s. d.)</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Mixed latent and observed treatment effects: Once again we use a sample size of 1000 patients, but now include an observed treatment effect. All settings are the same as for the previous set-up, except now there are high (RR=0.3), medium (RR=0.5), and low (RR=0.9) observed treatment effects simulated in each of the high and low mortality conditions. As before, we iterate over latent treatment effects from 0.1 to 1.0 by increments of 0.1 and average the results of 10 runs with each setting.

Constant mean effect, varying standard deviation: Finally, we examine the effect of variance in latent treatment effects. Here we fix mortality as high (0.05) and treatment as moderately effective (RR=0.5), and use a latent effect of 0.7, based on results of the first two simulations. These parameters are now fixed as we iterate over latent effect standard deviation mean from 0.5 to 2.5 by increments of 0.5. Once again we use 1000 patients in each run and 10 runs for each parameter setting.

Analysis

The primary outcome from each dataset generated is the mortality rate. We first aim to compare groups to determine if we can distinguish between more and less effective treatments. We test statistical significance using a Fisher exact test and threshold of $p < 0.05$ for statistical significance. We also assess observed study power (using actual mortality rate of the simulations and other simulation parameters), as it can differ from expected power.

Results

We now examine results for each of the three simulated scenarios, where the primary outcome is mortality and we aim to determine whether there is a statistically significant difference in mortality rate between treatment groups.

Effect of latent treatment

In the first case, we simulated a totally ineffective treatment, but varied the latent treatment and base risk of mortality. In this case we ask, at what point will there be a statistically significant difference between groups? Figure 3 shows the total number of survivors (out of 1000 patients) as we vary how effective the latent treatment is, holding constant an ineffective observed treatment under the two mortality conditions (high and low). When base mortality rate is high, the latent treatment has a substantial impact on total number of deaths, and a statistically significant effect on mortality once RR=0.8. This is because with a higher mortality rate, there is more opportunity to show a difference. On the other hand when mortality rate is low, it takes a much stronger latent treatment effect to show a difference in mortality (compared to a treatment effect of 1, meaning no impact). It is only once the relative risk reaches 0.7, that the difference between that and the condition where the latent treatment has no effect (RR=1) reaches statistical significance using a Fisher exact test, though the effect size remains small. This means that if we are comparing two groups with a high mortality condition (e.g. such as stroke), even if the treatment we are interested in has no effect on outcome, a relatively modest latent effect may confound our results.
Mixed latent and observed treatment effects

Now we combine observed and latent treatments, studying the impact of latent treatment effects with varying efficacy of observed treatments. Results are shown in figure 4. Different y-axis scales are used to better highlight subtle differences within the low mortality condition. First, within the high mortality condition, when the latent treatment effect is at least 0.6, the high and medium treatment conditions are no longer statistically distinguishable ($P > 0.14$). That is, even though there is an actual difference between these two treatments, if these were two groups in an RCT, we would no longer be able to find a difference between them because of the impact of the unobserved latent treatment. Further, in the low treatment effect condition, the latent effect makes the intervention seem to have a bigger effect on mortality than it actually does, which could lead to erroneous conclusions about the efficacy of an intervention.

In the low mortality condition (note that the range for the Y-axis is 970-1000), none of the differences between high and medium treatment effect are significant (P-values $> 0.6$), as the number of deaths ranges from 0 to 4. Thus, even though the observed treatments have a different effect on mortality in the ground truth of the simulation, we are unable to distinguish between them. Further, once the latent effect reaches 0.6, the medium and low treatment effects are no longer distinguishable (P-value $> 0.2$). The fact that the effects are indistinguishable does not represent the fact that the effects are somehow the same (the ground truth is that they are different). It simply means that the experiments are underpowered to detect a difference among them. If we were to increase the total number of patients in the experiment we would eventually reach a number where we could detect these differences.

Using a standard power calculation, the power for a study with 1000 patients in each group (high and medium efficacy treatment) using the actual true effect sizes, is 97.6% with an alpha of 0.05. Thus power calculations that ignore latent effects may lead to significantly underpowered studies. The actual observed power for detecting a change between these treatment groups, as a function of the strength of the latent effect is shown in figure 5. Once the latent effect reaches 0.7, power drops significantly, and critically is below the commonly used 0.80 threshold. Comparing the low and moderate treatments, power drops into an unacceptable range when the latent treatment effect reaches 0.4. We propose that power calculations may be augmented with simulations such as ours to explore how various unmeasured factors can affect effective power, and provide a more realistic estimate of the necessary sample size.

Effect of variance

Finally, we fix both treatment (RR=0.5) and latent effects (RR=0.7) along with mortality rate and now vary the standard deviation for the latent effect. This captures the scenario when the latent effect is incorporated in power calculations, but its variation is not properly accounted for. In most cases, we assume the treatment will have the same effect in different settings, but provider characteristics may vary in important ways between sites such as degree of adherence.
Figure 5: Actual power using $\alpha = 0.05$, when comparing two treatments, as a function of latent treatment effect’s impact on relative risk.

Figure 6: Result of variation in latent effect on mortality rate.

to protocol, level of experience, or workload. Figure 6 shows the result, where increased variation in treatment effect leads to significant differences in mortality rates within and across experiments. With zero standard deviation (so treatment effect is always mean of 0.5 and latent effect is mean of 0.7), mortality averages less than 3%, but this increases to 10% with s.d. 1.5, and 19% with s.d. 2.5. If these were multiple sites in an RCT, then significant differences may seemingly be observed even though the treatments have exactly the same mean effect in every experiment.

Discussion

Our simulations suggest that traditional sample size calculations may lead to underpowered studies in cases where there are latent effects or multiple sites. Further, due to the potentially complex interaction of multiple latent effects, some of which may be positive (high adherence to protocol) and some negative (high variance in clinician skill level), sample size calculations that account for these interactions may be prohibitively complex. Even when actual characteristics of an observed intervention are completely known (versus hypothesized, as in actual practice), we showed that when latent factors are not considered, studies become underpowered as latent effect size grows. Yet, these factors are rarely considered. One review found that only 41% of a set of pediatric ICU RCTs considered these nuisance parameters, leading to reduced estimates of variance and lower sample sizes than would actually be needed\(^{21}\). Another review found that many studies were either underpowered or overpowered (which can waste resources and time, and expose more patients to risks) due to incorrect assumptions about nuisance parameters\(^{22}\), which usually are not precisely known. While computational work on causal inference has aimed to identify latent confounders, the effects observed
here are not of the usual form considered (where a latent common cause may lead to inferring spurious relationships between its effects). This is exactly the type of variation that the randomization in RCTs aims to remove the effect of, but in research (1) from EHRs or (2) multi-site trials where effective sample size is much smaller due to site-related dependence, we cannot easily remove such effects through experiment design.

**Recommendations**

We propose that there are some key benefits for the simulation of medical processes and make suggestions for future work. First, while we focus here on a single latent treatment, the simulation introduced allows any number of such factors, enabling exploration of the impact of multiple latent effects which may have nonlinear interactions. This can enable better understanding of the effect different variances, treatment effects, and interactions have on sample size determinations. We propose that this can be used to complement standard power analysis and explore how effective sample size may differ under various conditions. Further this approach may be useful for analyzing RCT results (both the trial data and associated EHR data) to distinguish between intervention effects and other effects. As the simulation is extended to include other factors in EHR recording, latent factors like physician variation may have an identifiable signature. Second, it is customary to account for the effect of known co-interventions in any study. However normally only aggregate measures are available and used for analysis. With the advent of EHR documentation, one can envision a study where all of the co-interventions, known or unknown, are documented. Unfortunately, our ability to use this data is limited. The correctness of time varying analyses that aim to account for variations in co-interventions will need to be tested. Only with data for which ground truth is known can the validity of these algorithms be truly tested. By extending our simulation to include intermittent documentation of latent treatments, one can validate these time varying algorithms against with ground truth. Finally, simulation can be used to estimate the distribution of these latent effects to create data patterns that are similar to real world EHR data patterns and thus provide possible hypotheses of the underlying data generation processes that then can be tested. Ultimately, this approach will separate the underlying truth of a patient’s state from our sporadic and noisy measurement and documentation of this state, allowing better use of EHRs for research.

**Limits of Simulation**

The key limitation of our simulation approach is that it is meant to test the behavior of algorithms and contain similar structure to real scenarios, but it cannot be used to answer counterfactual queries about specific treatments. That is, we aim to capture the structure of the system (e.g., features leading to error) rather than to create an exact model of a specific disease process. Thus it cannot be used for developing intervention strategies. Further, some confounding factors may not be removable in practice. However, if they can be explicitly documented, it may be possible to account for their effects using EHR data and computational algorithms. In addition, we have chosen to use normally distributed parameters to model risk of disk and effects, for the sake of simplicity and because RCT power calculations are often based on such distributions. It is possible that the actual structure of the effects have bimodal or multimodal effects, which should be studied in future extensions to this work.

**Acknowledgments**

This work was supported in part by the NLM of the NIH under Award Number R01LM011826.

**References**


2. Prinz F, Schlange T, Asadullah K. Believe it or not: how much can we rely on published data on potential drug targets? Nature Reviews Drug Discovery. 2011 8;10(9):712.


Thinking Together: Modeling Clinical Decision-Support as a Sociotechnical System

Mustafa I. Hussain, MS¹, Tera L. Reynolds, MPH, MA¹, Fatemeh E. Mousavi, MS¹, Yunan Chen, PhD¹, Kai Zheng, PhD¹

¹University of California, Irvine, Irvine, CA, USA

Abstract

Computerized clinical decision-support systems are members of larger sociotechnical systems, composed of human and automated actors, who send, receive, and manipulate artifacts. Sociotechnical consideration is rare in the literature. This makes it difficult to comparatively evaluate the success of CDS implementations, and it may also indicate that sociotechnical context receives inadequate consideration in practice. To facilitate sociotechnical consideration, we developed the Thinking Together model, a flexible diagrammatical means of representing CDS systems as sociotechnical systems. To develop this model, we examined the literature with the lens of Distributed Cognition (DCog) theory. We then present two case studies of vastly different CDSSs, one almost fully automated and the other with minimal automation, to illustrate the flexibility of the Thinking Together model. We show that this model, informed by DCog and the CDS literature, are capable of supporting both research, by enabling comparative evaluation, and practice, by facilitating explicit sociotechnical planning and communication.

Introduction

Computerized clinical decision support systems (CDSSs) are designed to improve patient safety, but may cause harm if implemented improperly.¹ The determinants of CDSS effectiveness are poorly understood, and the problem space is complicated by heterogeneity in CDSS system design.² Although CDSS taxonomies exist that acknowledge the importance of sociotechnical context,³,⁴ which we use here to refer to the collection of relationships between actors and artifacts, we have found in this work that this context is often missing in peer-reviewed CDSS research. Specifically, we found that 1 out of 7 articles did not define or describe how suggestions were delivered to the user (e.g., “a pop-up alert,” “a phone call,” “a fax”). Additionally, although we identified 9 articles documenting a system in which a pharmacist filtered automated suggestions for presentation to a physician, none of these provided enough information to model the sociotechnical system in which the computerized subsystem was situated, missing key sociotechnical information, such as how advisories were presented (individual pop-up alerts or a list), and what other advisory systems were already in place. This makes comparative evaluation difficult. Additionally, while it is possible that sociotechnical design is considered in practice and simply not reported, it is also possible that this reflects a scarcity of sociotechnical consideration in practice.

In this theoretical development paper, we aim to expand the understanding of CDSSs as sociotechnical systems, where humans and computers think together to steer the course of clinical care. Explicit sociotechnical consideration may also improve adoption and clinician responses to computerized decision-support advisories, which may, in turn improve workflow, care coordination, and ultimately quality of care and patient safety.⁵ Inadequate sociotechnical consideration can disrupt work and even lead to patient mortality.¹

To this end, we present the main contribution of this paper, the Thinking Together model, a diagrammatical representation for communicating the sociotechnicalities of CDSSs. It provides a common diagrammatical language for guiding the sociotechnical design of CDS systems, conducting research studies of these systems, and reporting research results. Widespread usage would provide researchers with a uniform body of comparable literature. This would ultimately benefit practice indirectly, by providing evidence upon which to guide the design of sociotechnical systems. The Thinking Together model also provides a more direct benefit to practitioners: Monk and Howard similarly created the Rich Picture as a tool for designers of business information technology to think, talk, and negotiate sociotechnical design by making it explicit.⁶ We have developed the Thinking Together model specifically for the sociotechnical context surrounding CDSSs, to benefit both research and practice.

We developed the Thinking Together model using a systematic approach, by conducting a comprehensive review of articles documenting CDSSs published in the past decade. We used the theoretical lens of Distributed Cognition (DCog)⁷ to interpret descriptions of CDSSs, because DCog is well-suited to studying the complex, collaborative, and dynamic nature of clinical work,⁸,⁹ and it has been widely used to study health informatics applications, such as interruption management in the intensive care unit¹⁰ and whiteboard use in a trauma center and an operating room.⁹
Rather than focusing on the individual actors (e.g., an internist, a pharmacist), DCog emphasizes the sociotechnical system, which includes actors, roles, and artifacts. In this study context, actors perform roles. Human actors include physicians, nurses, and pharmacists. We also speak of computer systems as actors—such as a computer program that fires alerts. These actors think together by representing and conveying information through artifacts, such as pop-up alerts, faxes, and spoken words. Artifacts have intrinsic properties with implications for design; an utterance is immediate, intangible, and ephemeral, while a fax alert must be discovered, can be annotated, and persists until destroyed. By applying DCog to understand CDSSs, we are able to more clearly depict important sociotechnical features, and how they work together to achieve a common goal: delivering high-quality patient care.

In the next section, we briefly present the history of CDSSs which led up to this work. Then, we describe the methods that we used to develop and validate the Thinking Together model instances. In the section titled Two Case Studies, we present two resultant Thinking Together model instances, based on CDSSs selected from the literature, to illustrate how the Thinking Together model can be used to represent CDSSs as sociotechnical systems. We follow these with a discussion of uses, benefits, and limitations of these model instances, and some concluding remarks.

**Background**

Long before the computer became widely available, a clinician’s decision making was supported by artifacts such as reference manuals and dose charts, and by talking to other clinicians. The advent of the computer presented the possibility of using computing methods to automatically generate decision-support advisories. There was a significant effort to develop expert systems, usually based on artificial intelligence methods, to automate the diagnosing process. This attempt fell short, partially because these AI-based applications were unable to reason anatomically or temporally, and because they were unable to explain their reasoning. These shortcomings persist today in machine learning applications, and much computerized clinical decision support instead tends to use explicitly defined, step-by-step rules, usually from evidence-based clinical guidelines. Today’s applications of CDSSs include antibiotic stewardship, opioid management, and adjusting doses for renal insufficiency. The suboptimal usage of these CDSSs, however, has been an enduring issue. The lack of usability is well documented, where low rule specificity and a lack of accurate patient data result in too many irrelevant alerts. When alerts are no longer perceived to be valid problem-indicating cues, they become habitually ignored.

Today’s CDSSs are highly heterogeneous, with different capabilities for patient information access, rule programming, advisory display, and shortcuts for immediate action. Various taxonomies have been developed to organize this heterogeneity. For example, Wright et al. developed a taxonomy to classify the problems that computerized CDSSs address, and some user interface characteristics. Berlin et al. presented a taxonomy with 26 axes and 108 descriptors to classify clinical context, information sources, the underlying recommendation-producing mechanism, information delivery, and how it relates to clinical workflow, noting that one should be cautious when generalizing results of randomized controlled trials of a CDSS intervention across different clinical or workflow settings. Over a decade later, we find that generalization remains difficult, because although these taxonomies exist, many research papers do not use them to report CDSS characteristics. While the question of why these taxonomies are underused remains open, we speculate that this may be because taxonomies do not depict systemic configuration, and so may be less than useful in CDSS design. Rather than provide another taxonomy, our goal in this work is to provide a means of depicting systemic configuration that practitioners find useful, and can be just as easily repurposed to provide context when submitting manuscripts for publication. The Rich Picture, a similar approach, has seen success in business-oriented software design.

**Methods**

In order to develop a flexible representation system—the Thinking Together model—for characterizing CDSSs that depict their systemic configurations, we first needed to arrive at a reasonably inclusive set of features that differentiate CDSS systems. We used a systematic review approach, illustrated in Figure 1. In the first stage, we developed search queries for PubMed/MEDLINE, EMBASE, CINAHL, and Cochrane literature databases. The queries included keywords such as “decision support,” “alert,” and “error.” Since this is intended to model contemporary systems, we restricted the queries to literature published in the last 10 years. The final queries returned a total of 2,760 results, of which 996 were duplicates. We reviewed the remaining 1,764 unique papers for relevance. Of these, 255 met our criteria for inclusion, which were that a paper must be (1) peer reviewed, (2) about a CDSS that targeted clinicians, and (3) written in English. When the same research team reported multiple times on the same system, we kept only the most thorough report.
Figure 1. Literature examination process.

We attempted to identify descriptions of the CDSS in all of these papers, in particular what clinical role was targeted (e.g., physician or pharmacist), how alerts were generated and filtered (e.g., automatically or by a pharmacist), how the advisory was presented (e.g., pop-up dialog or human phone call), and what action shortcuts it provided (e.g., cancel or substitute medication), if any. Notably, approximately 1 in 7 did not define or describe any of this key information. This eliminated 38 articles. For the 217 papers that remained, we inductively identified categories that characterize the sociotechnical aspects of CDS systems, using the constant comparative method. More specifically, we started with an empty list of categories, and for each paper, we compared the described CDSS to the current list of categories, such as “interruptive pop-up alert” and “tab highlights yellow to draw attention, user clicks tab and views list of alerts”. If the system did not fit in any of the existing categories, we extracted or generated a brief description to add to the list. Otherwise, we categorized the system accordingly. This analysis was independently performed by two of the authors, Hussain and Reynolds.

In the second stage, we iteratively developed the properties of the Think Together model by generalizing phrases from the categories into properties, such as by generalizing the phrases “interruptive,” “non-interruptive,” “searched database” (among others) into a broader means of discovery and retrieval. Some of these properties formed groups; for example, means of discovery and retrieval and means of destruction (again, among others) into a broader means of interaction. Properties that characterized content mapped to Grice’s maxims of communication, so we formed these into a group as well. We then organized these properties into theoretical elements by mapping them to DCog theoretical elements. The DCog concepts of systems, roles, and artifacts became prominent. Any ambiguities were resolved through discussion among the two coders. Finally, we selected two sample CDSS systems to validate the model. We purposefully selected one system that is almost fully automated, and another that is almost entirely manual, in order to “stress test” the model using extreme cases to ensure its robustness and representativeness.

Results

Before describing the theoretical elements and their properties, we will provide a general overview of the Thinking Together model. Roles are played by actors; these actors manipulate artifacts. Together, they compose a sociotechnical system. These concepts are holdovers from DCog, where a sociotechnical system performs cognition by transiting information between specialized roles, which process information, collectively determining the course of action. Information is not only distributed among the heads of actors, but also on external artifacts, enabling problem-solving and team coordination. In this work, we differentiate between human and automated actors.

Theoretical elements and properties

As previously mentioned, we identified theoretical elements and properties relevant to CDSS sociotechnical consideration, based on the systematic review and the DCog theory. The results are presented in Table 1, and discussed in this section. Key concepts are italicized in this section. These concepts are important to understand before interpreting instances of the Thinking Together model, which we present in the case studies that follow this section. Below, we explain these theoretical elements and their properties.

Sociotechnical Systems

The Systems element addresses the fundamental premise underlying a CDSS: What is the purpose for this system’s existence? Is it there to support clinical care, or to avoid legal consequences? If the latter, then it is worth considering whether other, more clinically valid CDS alerts are more likely to be ignored in the same sociotechnical system due to a perception of pop-up alerts as an irrelevant nuisance, contributing to alert fatigue.
Table 1 Theoretical elements and their properties.

<table>
<thead>
<tr>
<th>Systems</th>
<th>Roles</th>
<th>Artifacts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Purpose</td>
<td>Human or automated</td>
<td>Physical location</td>
</tr>
<tr>
<td>Inception</td>
<td>Level of expertise</td>
<td>Medium: Virtual or physical</td>
</tr>
<tr>
<td>Maintenance</td>
<td>Area of specialization</td>
<td>Shortcuts for action</td>
</tr>
<tr>
<td>Follow-ups</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Means of interaction**

- Discovery and retrieval
- Generation and processing
- Storage and sending
- Destruction

**Content (Grice’s maxims)**

- Quantity
- Quality
- Relation
- Manner

Also, what is the story of this system’s *inception*? Was it by administrative mandate, or regulatory requirements? Did clinicians push for it? Whether manually-programmed or machine learning algorithms are used, were they calibrated to the locality? Finally, how is the system *maintained*? Is it updated to reflect emerging, improved understandings of medical practices and/or the particular sociotechnical environment? Is there a regular, ongoing review of consequences that results in safety-improving updates to the sociotechnical system?

**Roles**

There are many roles in healthcare, including physicians, nurses, clerical staff, administrators, software developers, and patients. In this work, we draw a distinction between *human* actors and *automated* actors. Human information processing may be characterized as ranging from fast and subconscious to slow and conscious. Automated actors (in the form of computer programs) can apply rule-based or statistical methods quickly, but their results can be lacking due to an absence of context-dependent understanding, and machine learning applications tend to suffer from “black boxing,” wherein a human-understandable rationale may not be readily available. Also, whereas humans tend to be better conversationalists than chat-bots, computers excel at creating precise drawings and producing formatted information quickly.

While examining the literature, it became apparent that, whether it is an experienced brain surgeon, or an advanced analytic application that monitors patients for signs of sepsis based on algorithms trained with historical data, a role has a *level of expertise* and an area of *specialization*.

**Artifacts**

Artifacts include physical objects such as phones, dose calculators and conversion charts, checkbox forms, injection site maps, exam room flags, and patient schedules, as well as virtual objects, such as pop-up alerts, lists of suggested doses presented during order entry, and online reference entries. The artifact’s properties are largely reflected in Berlin’s taxonomy; here we focus on contextualizing it as a core element of a sociotechnical system.

Physical artifacts are designable and configurable by nearly any user. With a paper chart, one can fold a corner, jot a note, stick a colored dot. This flexibility allows for spontaneous creation and modification of external representations, which, in turn, allows for problem reframing, an essential step in problem solving.

Virtual artifacts, on the other hand, are interesting because they are not bound to the same laws as physical artifacts. For example, a modal dialog or modal window is a pop-up message that is not escapable, and which usually locks down several on-screen background artifacts until an action is taken. The modal dialog does not accommodate situations in which information must be retrieved from these frozen background windows, but the needed information is scrolled out of sight. Further, once these dialogs disappear, they are usually gone forever, and cannot be retrieved again at will for reference, that is, they are transient, not permanent. This type of brittle interactivity violates Nielsen’s heuristic of *user control and freedom*. Careful design can resolve many usability issues.

972
Artifacts are manipulated by actors. In the literature, we found that an actor may discover an existing artifact in different ways (perhaps the “new email” notification sound plays), or they might retrieve it from a storage system. They may process an existing artifact, or generate a new one. Finally, they may store, send, or destroy it. Many virtual artifacts found in health IT systems, such as physician notes, cannot be modified or destroyed, for legal reasons rather than those relating to clinical care.29 Modal dialogs, eager to self-destruct, are a curious exception. It may be wise to approach object permanence, and destructibility on a case-by-case basis.

An artifact’s medium of conveyance determines much of what can be done with the information it represents. A pharmacist’s phone call about a medication error may be easy to discover because the phone is ringing, but the content may need to be transcribed for future reference, perhaps on a pad of sticky notes. It may take some time for this phone call to arrive, because the pharmacist is human, and so their information processing tends to be slow.

Though delayed, the content of a human-generated phone call may conform better to Grice’s maxims of communication30 than a computer-generated phone call. Grice’s maxims of communication were reflected in the literature we examined, underscoring the relevance of broader social sciences research to the design of health IT systems. Grice’s maxim of quantity states that one should be no more and no less verbose than is required to convey the information. When an action is suggested, is a rationale provided? If not, the content may be too sparse. Is the entire text of a literature review displayed to explain a particular computer-generated decision-support advisory? This may be too much. Grice’s maxim of quality refers to the accuracy of the information presented. Is a suggestion based on a study in rodents? Were there conflicts of interest that may diminish the legitimacy of the results, and therefore the recommendation? Grice’s maxim of relation refers to the relevance of the content to the task and context at hand. Is an alert for drug-allergy interaction presented because the patient is known to have an allergy, or is it presented for all patients “just in case?” If a phone call about a prescribing error interrupts note-taking, is it important enough in relation to note-taking to justify the potential loss in productivity and accuracy in note-taking?29 Grice’s maxim of manner states that communication should be clear, brief, orderly, and specific. Tabular formats may be more appropriate than prose in certain CDS pop-up alerts,32 possibly because the spatial order and the discrete nature of the information displayed can be more easily retrieved.

If a document is represented in an electronic health record (EHR), one may need to access a computer to interact with it; its physical location is constrained, and this limits the contexts in which suggestions may be displayed. The dimensions and resolution of the display also constrain the design of virtual artifacts. On the other hand, virtual artifacts can also provide convenient shortcuts for common actions that one may need to take, such as ordering appropriate consultations or medications upon receipt of an abnormal test result. Smith et al presented shortcuts for such actions within pop-up alerts, rather than simply demanding clinician acknowledgement.5

Some artifacts in the literature were intended to follow up on a previous communication. This sometimes comes in the form of personalized phone calls conducted by humans, and sometimes as pop-up reminders that are displayed at regular intervals until a task is accomplished. When deciding between these options, one might consider that nuisance alerts contribute to alert fatigue.19

The Thinking Together Model

The configuration of information flow is modeled in numbered stages. During each stage, an artifact may be retrieved from storage or generated anew, processed by an actor, and then sent or stored to another actor or a storage system. When using Thinking Together to model a sociotechnical system, we recommend first drawing and labelling all of the roles and storage systems involved, in the order in which they come into play, working clockwise. Next, connect the roles and storage volumes with arrows, showing the flow of information. Finally, label these roles and arrows, one stage at a time, starting with the number one, and labelling each step in each stage as r, p, g, and s, for retrieval, processing, generation, and storing or sending (See Figure 3 and Figure 4). Even if the present study focuses on one prong of a multi-prong intervention, be sure to include other forms of decision support that clinicians receive addressing the same topic. For example, if physicians already receive antibiotic management alerts, and the study at hand addresses the additional intervention of pharmacist review and counseling, document both of these interventions in the diagram, since the interaction between these two interventions may need to be later teased out in an aggregated literature review.

![Figure 2: Recommended cyclical modeling process.](image-url)
Next, interrogate the diagram, considering the theoretical elements and properties in Table 1. For example:

- How is this rule-based advisory system maintained as new clinical evidence emerges?
- How do users point out problematic alerts, so that rules can be changed?
- While making this phone call, does the pharmacist adjust dose on behalf of the physician, for later signing?

If the answer to a question is unknown, it may be worthwhile to find out, such as by interviewing physicians to find out what they do, as one does when generating a Rich Picture. Next, update the design, and continue to ask questions. Continue this cycle, as shown in Figure 2, until no new questions emerge.

Because Thinking Together is a descriptive, rather than prescriptive, modelling system, we do not prescribe rules for evaluating or suggesting changes. Rather, Thinking Together is intended to facilitate interrogative reflection.

**Two Case Studies**

We selected 2 CDSSs described in the literature to illustrate how the Think Together model can be applied. As mentioned earlier, we purposefully chose two systems at the extremes of the automation spectrum to validate the robustness and representativeness of the model. The first system is primarily automated, requiring little to no human involvement, while the second is operated almost entirely by human actors.

**Case 1. Pre-Surgery Lab Alerts**

In Figure 3, we show a rule-based alert system documented by Freundlich et al. It was put into place following an adverse event in which a patient’s partial thromboplastin time (PTT), a lab value that indicates blood coagulation, became abnormal without the awareness of the surgical team. The surgery was performed without additional precautions that may have prevented the patient’s death. To address this, a system was instated that retrieved patients scheduled for surgery, along with their lab values (1r), and compared their lab values with predefined thresholds (1p). It then generated (1g) and sent (1s) alphanumeric pages to the surgical staff concerning patients with abnormal lab values. This alerting process was performed every afternoon for most surgical units, and also in the morning before surgeries in the neurosurgery unit.

The authors reported that feedback was largely positive, and that suggested improvements to the rules encoded in the system had been received and were implemented. However, since we do not know how this feedback was collected, we are unable to draw this portion of the diagram. It is imaginable that this may impact how the rules were updated. This has practical, real-world implications. To name an example, two-way pagers might elicit more contextual feedback than a one-way pager, since email, phone call, or face-to-face conversation require more steps to providing feedback than replying to a page. On the other hand, individual pages might provide less information.

The authors also stated that some clinicians may have perceived the alerts as a redundant nuisance. It is possible that this refers to the less-frequent manual laboratory review prior to scheduling the surgery, which also cannot be diagrammed because it is similarly left unspecified.
A Systematic Oversight

We would like to make it clear that we hold the work of Freundlich et al. in high regard. Their paper was well-written, the methods were reasonable, the alerting system appeared thoughtful, and the technology staff appeared responsive to clinician feedback.

It is easy to overlook the provision of additional context when focusing on a specific problem, and their paper is certainly not alone. We also wished to show a diagram displaying a system in which an automated system generated many alerts that were then whittled down by a pharmacist, who then contacted physicians personally, by phone or email. Hybrid systems like this would serve as an illustrative contrast between a fully automated system, like the one we presented in the preceding case study, and a fully manual system, like the one we will present in the following case study. However, although we identified nine articles that documented such a system, none provided the reader with the full sociotechnical system, missing key information, such as whether suggestions were presented to the pharmacist as pop-up alerts or as a list, whether the physician received other alerts (and in what form), and how follow-ups were conducted, such as via email, face-to-face conversation, or phone call. This suggests that some sociotechnical aspects of these CDSS designs went without consideration. This illustrates a systematic oversight that we have observed in the clinical decision support literature: Although sociotechnical context is widely recognized as an important factor to consider in the implementation and alteration of CDS systems,1,32,33 this information is under-reported to the extent that it is currently difficult to systematically analyze the effectiveness of real-world interventions.

This has importance to practitioners for two main reasons. The first is less direct; comparative evaluation research will benefit practice by providing reliable, useful guidance. The second is more direct; explicitly diagramming sociotechnicalities may help the practitioner reason about, communicate, and negotiate sociotechnical design.6

Case 2. A Human-Operated CDSS

![Diagram](Figure 4. A Thinking Together model of a multidisciplinary CDSS with minimal automation.24)

We selected the CDS presented by Sullivan et al.24 for modeling as well; it is presented in Figure 4. In Stage 1, prescribers write prescriptions, which are stored electronically. In Stage 2, pharmacists retrieve (2r) and review (1p) these prescriptions daily. When they see an error, they take two send/store actions: they call the prescriber (2s1) and log the error (2s2).
In Stage 3, the multidisciplinary team meets every other week, retrieving all error reports (3r), aggregating them (3p), writing feedback emails (3g), and sending them (3s). In this same meeting, the team also retrieves prescribers' emails about systemic concerns (4r), plans corrective action (4p), and contacts the appropriate party (4s). We considered these to be two separate stages because they appeared to be two separate tasks. In Stage 5, prescribers receive (5r) and read (5p) these emails, and write (5g) and send (5s) responses about systemic concerns that may underlie the errors if appropriate. These emails are read during the next biweekly meeting.

We found this case interesting because it is run manually; though information was relayed and stored with computers, it is hard to argue that automated actors were really supporting decision-making by filing data and shuttling emails. It is also difficult to argue that clinical decision-making is not supported in this sociotechnical system. The choice to shun computerization was conscious; the authors cited unintended consequences, alert fatigue, and low effectiveness at preventing medical errors as rationale. Their system serves as a counter-example to CDS definitions that assume computer dependence: one could imaginably replace the databases and email server with file cabinets and couriers, while maintaining similar functionality. Also noteworthy is that feedback from prescribers could be flexibly handled by the multidisciplinary team precisely because they are human; they were able to translate explanations for errors into plans for corrective action (4p), and relay those plans to the appropriate recipients (4s). For example, they had rotavirus vaccine intramuscular removed from the available CPOE options because, after it caused a medical error, a prescriber pointed out in a systemic concern email (5s) that it was not an appropriate option in any case. Also noteworthy is the care with which the multidisciplinary team composed the feedback emails (3g); they were careful not to place blame, in order to maintain workplace safety culture, and they provided details on specific errors and advice for avoidance. Courteous prose and real-world reasoning are hardly a computer’s forte.

Discussion

In this work, we developed the Thinking Together model, a descriptive and flexible means of representing CDSS sociotechnicalities. We are confident in its flexibility due to the contrast between the two case studies presented, as well as its grounding in DCog, which has remained largely the same for over two decades. Usage in future research is likely to provide a basis upon which to compare the effectiveness of CDSS implementations, since sociotechnical context is critical to a CDS implementation’s success. The ability to comparatively analyze the body of work on CDSS in aggregate will allow researchers to make meaningful recommendations to improve CDS systems. It will be difficult to perform such comparative analyses without clear information about the sociotechnicalities of the CDS systems from which individual results originated. The Thinking Together model is also intended to support the examination of the sociotechnical context when implementing and altering CDS systems in practice. The potential of this model to improve both research and practice can only be realized through cooperation among researchers, clinical practitioners, and technologists.

System representation is powerful insofar as it shows only key details, allowing an overview. Many low-level details that may affect system efficacy cannot be shown. For example, a neural network trained on too many dimensions may learn a practically useless identity function, and may be outperformed by one trained on only the most relevant dimensions (as determined by a principal components analysis, for example), which would arrive at an equation that is valid for previously unseen cases. This distinction will not be apparent in a Thinking Together diagram. However, the burgeoning work in automated information processing complements the sociotechnical viewpoint. Likewise, interpersonal relationships can affect healthcare safety; this is also hidden from view by necessity, but is complemented by work in Crew Resource Management.

The Thinking Together model should be thought of as a tool in a toolbox of complementary approaches, from algorithm design to workplace culture. It provides an intuitive means of representing and communicating sociotechnical clinical decision support, to support comparative evaluation in research and design in practice.

We invite practitioners to use the Thinking Together model for integrating CDSSs in sociotechnical design, to reflect on their experiences, and to publish their results. We believe they will be useful in practice, and we hope that this will transform reporting sociotechnical context from a tangential burden into a simple matter of cleaning up and attaching an existing drawing that was used during design. We plan to conduct and publish comparative analyses of CDS system effectiveness as articles documenting both results and sociotechnical design become available.

Conclusion

In this paper, we described the Thinking Together model, which was based on a systematic review of the literature and the theory of Distributed Cognition. Through two case studies, we demonstrated the flexibility of the Thinking
Together model. We believe that practitioners will find this model useful when considering sociotechnical context in computerized decision support system design, and it will support systematic sociotechnical reporting in the literature, enabling comparative evaluation of CDSS efficacy.

References


Differential Data Augmentation Techniques for Medical Imaging Classification Tasks

Zeshan Hussain\textsuperscript{1}, Francisco Gimenez\textsuperscript{2}, PhD, Darvin Yi\textsuperscript{2}, Daniel Rubin\textsuperscript{2}, MD, MS
\textsuperscript{1}Stanford University, Department of Computer Science, Stanford, CA; \textsuperscript{2}Stanford University, Department of Radiology, Stanford, CA

Abstract

Data augmentation is an essential part of training discriminative Convolutional Neural Networks (CNNs). A variety of augmentation strategies, including horizontal flips, random crops, and principal component analysis (PCA), have been proposed and shown to capture important characteristics of natural images. However, while data augmentation has been commonly used for deep learning in medical imaging, little work has been done to determine which augmentation strategies best capture medical image statistics, leading to more discriminative models. This work compares augmentation strategies and shows that the extent to which an augmented training set retains properties of the original medical images determines model performance. Specifically, augmentation strategies such as flips and gaussian filters lead to validation accuracies of 84\% and 88\%, respectively. On the other hand, a less effective strategy such as adding noise leads to a significantly worse validation accuracy of 66\%. Finally, we show that the augmentation affects mass generation.

Introduction

Tremendous progress has been made in using deep learning models for image classification and segmentation. In particular, these methods have been adapted in medical diagnostic tasks, such as the prediction and segmentation of cancerous masses, across different modalities, including lung, liver, and breast scans\textsuperscript{123}.

One important preprocessing method that has been shown to be effective in training highly discriminative deep learning models is data augmentation. Data augmentation was initially popularized by Tanner and Wong in order to make simulation more feasible and simple\textsuperscript{4}. In computer vision, because there are generally millions or even billions of parameters in CNNs, data augmentation is critical to accumulate enough data to attain satisfactory performance. Multiple data augmentation strategies have been proposed to improve vision tasks for natural images. Conventional strategies including horizontally flipping images, random crops, and color jittering\textsuperscript{5}. Krizhevsky et al. employ a technique called fancy PCA, which alters the intensities of the RGB channels in training images\textsuperscript{6}.

While different augmentation strategies and their combinations have been researched heavily for natural image tasks, there has been little work on finding optimal augmentation strategies for medical imaging tasks. Unlike in the natural image domain, where ImageNet and similar datasets provide millions of images, there are far fewer training images available in medical imaging\textsuperscript{7}. This dearth of training data makes it critical to explore methods such as data augmentation, which serves as a regularizer and addresses the data-scarcity problem. In this work, we study different strategies for binary image classification of mass and non-mass mammogram images. We show that some augmentation methods capture medical image statistics more effectively than others, leading to higher training and validation accuracy. Finally, we demonstrate that smarter augmentation may result in fewer artifacts in CNN visualizations.

Methods

We attempt to gain insight into the effect that various augmentation methods have on classification accuracy and visualizations of trained CNNs. Our workflow consists of four main steps, the details of which are given in the following sections. First, for each image $I$ in the entire dataset, we perform initial preprocessing, which includes cropping the full size image and splitting into training and validation sets. Second, for each image $I_t$ in the training set, we perform one of eight augmentations. Third, after performing additional cropping of the augmented images, we train eight VGG-16 nets independently on the eight uniquely augmented sets. Finally, we evaluate performance of the trained CNN by measuring training and validation accuracy of mass/non-mass mammogram classification as well as qualitatively assessing visualizations generated from the CNN.
Dataset & Preprocessing

A set of 1650 mass cases and 1651 non-mass (normal) cases were obtained from the Digital Database for Screening Mammography (DDSM)\textsuperscript{8}. The full sized images were initially cropped to 1000 by 1000 images to improve augmentation speed. Each mass image was cropped around the mass lesion, while a random 1000 by 1000 crop of the breast tissue was taken in normal images. Finally, we split the full dataset into a training set and validation set, where approximately 80\% of the images are in the training set and 20\% are in the validation set.

![Example Mass Case + Augmentations](image)

**Figure 1:** Example Mass Case + Augmentations

Augmentations

Before training the models, we utilize eight augmentation strategies to generate eight new training sets. Each new training set is simply the original training images in addition to the training images augmented by one of the techniques below. Each augmentation is as follows:

- **Flips:** We perform a horizontal and vertical flip for each image $I$ in the training set. Even though only horizontal flips are used in natural images, we believe that vertical flips capture a unique property of medical images, namely, invariance to vertical reflection. Conventionally, for natural images, only horizontal flips of the original images are used, since vertical flips often do not reflect natural images (i.e. an upside-down cat would not generally make a model more discriminative during training). However, a vertical flip of a mass would still result in a realistic mass.

- **Gaussian Noise:** We generate an array, $N$, where each element in the array is a sample from a gaussian distribution with $\mu = 0$ and with $\sigma^2$ in the range of $[0.1, 0.9]$. Then, for each image $I$, we obtain a noisy image, $I' = I + N$.

- **Jittering:** For each $I$, we add a small amount of contrast (+/- 1-4 intensity values).

- **Scaling:** We scale each $I$ in either the $x$ or $y$ direction; specifically, we apply an affine transformation, $A = \begin{pmatrix} s_x & 0 \\ 0 & s_y \end{pmatrix}$ on $I$.

- **Powers:** We take each $I$ to a power. To calculate the power, $p$, we use the following equation, $p = n \cdot r + 1$, where $n$ is a random float taken from a Gaussian distribution with mean 0 and variance 1 while $r$ is a number less than 1. Then, to generated the augmented image, $I_a$, we have, $I_a = \text{sign}(I) \cdot (|I|^p)$. The sign and power are both taken elementwise.

- **Gaussian Blur:** We blur each image $I$ by a gaussian function defined by a variance between 0.1 and 0.9. The filter size is then generated internally by scikit-image\textsuperscript{9}, where the radius of the kernel is, $r = 4 \cdot \sigma$.

- **Rotations:** The following affine transformation, $A = \begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix}$, where $\theta$ is between 10 and 175 degrees, is applied.

- **Shears:** Finally, each image $I$ is sheared, represented by the following affine transformation, $A = \begin{pmatrix} 1 & s \\ 0 & 1 \end{pmatrix}$. $s$ defines the amount that $I$ is sheared, and it is in the range of $[0.1, 0.35]$. 

980
For each augmentation, we vary the augmentation hyperparameters across their specified ranges and generate a final augmented training set of 15673 images. These ranges were chosen such that a single transformation in the range preserves the class while a transformation outside of the range is not guaranteed to be class-preserving. Note that no augmentations were done on the validation set.

**Training & Evaluation**

Before we begin the training, we further crop each image in both the training and validation set to 500 by 500 pixels. We do so to retain as much information as possible before down sampling to 224 by 224 pixels, which is the input image size expected by the VGG-16 CNN. Then, we train eight VGG-16 CNNs on the eight augmented training sets. The specific experimental parameters are described in the next section.

Before evaluating model performance, we first quantify the augmentation by comparing the similarity between the mean image of the augmented training set, \( M' \), and the mean image of the pre-augmented training set, \( M \). The mean image is an intuitive representation of the training set because it captures the basic regularities across all the training images (e.g. general mass shape, location, etc.). Specifically, we compute the mutual information between \( M \) and \( M' \). Mutual information has emerged as an effective similarity metric between two images; it captures the reduction in uncertainty of one variable given that we know the other. We evaluate each model by observing the variation in training and validation accuracy across augmentations.

Additionally, we qualitatively assess visualizations generated by the CNNs. The visualizations for each CNN are generated according to the class visualization method detailed by Simonyan et al. Specifically, we numerically generate an image that represents the class of interest, which in this case is "mass", from an input image of normal breast tissue. We do this generation by performing gradient ascent on the target class.

Formally, let \( I \) be the the input image and let \( y \) be the target class. \( s_y(I) \) is the score that a CNN assigns to image \( I \) for class \( y \). The image, \( I^* \), that maximizes the score for class \( y \) is generated by solving the following optimization problem,

\[
I^* = \arg \max_I s_y(I) + \lambda \|I\|_2^2.
\] (1)

Note that \( \lambda \) is a regularization parameter. We solve this optimization using backpropagation, where we initialize the scores to be a one-hot vector, with the target class set to 1 and the other class set to 0. Letting \( dI \) be the gradients derived from backpropagating from the score layer, the final gradients are defined by the following equation,

\[
dI = dI - 2\lambda \cdot I,
\] (2)

where the second term, \( 2\lambda I \), comes from the derivative of the second term in the objective function.

**Experiments**

We perform eight training experiments, where one experiment consists of training a VGG-16 net on one of the eight augmented sets. For each experiment, the learning rate is set to \( 1e^{-3} \), L2 regularization is set to \( 1e^{-7} \), and the dropout parameter \( p \) is 0.5. The network is then trained over its corresponding augmented training set for 2500 iterations. We train each network for approximately 1.5 epochs, since the classification accuracy generally plateaus around this point.

For the visualization experiments, we find the gradients that need to be applied to image \( X \) at each iteration via backpropagation (derivation not shown). If we let \( dX \) be the gradients derived from the backpropagation at iteration \( i \), then the update step at \( i \) will be \( X = X + \alpha \cdot dX \), where \( \alpha \) is the learning rate. For each of the eight visualizations, the learning rate is set to \( 1e^{-1} \), the regularization parameter applied to \( dX \) is \( 1e^{-7} \), and the number of iterations where we update \( X \) is 350.
Results

To visualize the effect of each augmentation on the training set, in Figure 2, we present the mean image of the training set before augmentation as well as the mean images of each augmented training set. The mutual information between the mean images of each augmented training set and the mean image of the original data set are given in Table 1. These values enable us to explore the relationship between the mutual information of a mean image generated from an augmented training set and the training and validation accuracy of the CNN trained on that augmented set. In other words, we attempt to determine if the extent to which an augmentation captures information about the original training set affects the training and validation accuracies. Because the original training set is representative of the validation set, this effect gives us some intuition on how the model might generalize given some augmentation on the training set. Figure 3 shows how classification accuracies vary with the mutual information of the augmented training set’s mean image.

<table>
<thead>
<tr>
<th>augmentation type</th>
<th>MI</th>
<th>training acc.</th>
<th>validation acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noise</td>
<td>2.27</td>
<td>0.625</td>
<td>0.660</td>
</tr>
<tr>
<td>Gaussian Filter</td>
<td>2.60</td>
<td>0.870</td>
<td>0.881</td>
</tr>
<tr>
<td>Jitter</td>
<td>2.59</td>
<td>0.832</td>
<td>0.813</td>
</tr>
<tr>
<td>Scale</td>
<td>2.67</td>
<td>0.887</td>
<td>0.874</td>
</tr>
<tr>
<td>Powers</td>
<td>2.33</td>
<td>0.661</td>
<td>0.737</td>
</tr>
<tr>
<td>Rotate</td>
<td>2.20</td>
<td>0.884</td>
<td>0.880</td>
</tr>
<tr>
<td>Shear</td>
<td>2.06</td>
<td>0.891</td>
<td>0.879</td>
</tr>
<tr>
<td>Flips</td>
<td>2.70</td>
<td>0.830</td>
<td>0.842</td>
</tr>
</tbody>
</table>

Table 1: Mutual Information between Mean Images + Average Accuracies

Figure 3: Accuracy vs. Mutual Information of Mean Images

Generally, we wish to see how each model performs by observing its classification accuracy over some number of epochs. The training/validation accuracy is computed simply by finding the number of samples classified correctly as mass or non-mass out of the total number of samples in the training/validation set. These accuracies are tracked over 2500 iterations (approximately 1.5 epochs), and the mean accuracies are shown in Table 1. Finally, we also assess the
performance of each model by qualitatively analyzing the image generated from the CNN using the class visualization method described previously, where we iteratively alter an input image to make it look more like a "mass". These images are shown in Figure 4.

![Figure 4](image_url)

**Figure 4**: Visualizations; augmentation used has significant effect on the generated mass image. Artifacts that appear are correlated to the type of augmentation used.

**Discussion**

Looking at the results from Table 1 and Figure 3, we notice that aside from the accuracies for the shear and rotate augmentation sets (the cluster on the top left), a higher mutual information correlates to a higher classification accuracy. Specifically, the augmentations that resulted in mean images that had mutual informations of approximately 2.6-2.7 (flips, scale, jitter, gaussian filter) had accuracies of around 0.85 while those that resulted in mean images with mutual informations of around 2.3 (noise, powers) had much lower accuracies of around 0.65-0.70. Also, in general, the noise and powers augmentations, resulting in mutual information 2.27 and 2.33, respectively, have more variance in their validation accuracies over time compared to augmentations resulting in higher mutual information (plots not shown).

With regard to the shear and rotation augmentations, which result in mean images with relatively low mutual information but high classification accuracies, there is an intuitive explanation as to why this is the case. Namely, rotations and shears do retain many of the image statistics in the original training set, leading to a high classification accuracy, but artifacts in the augmented images may affect the mean image and lead to a lower mutual information. For example, if a mass appears on the edge of a breast, then the 500 by 500 crop of the original image will retain the 0 pixels that appear when rotating the 1000 by 1000 image. Generally, if the mass is near the center of the crop, as is the case for the majority of the training images, then the 0 pixels will be cropped away. These slight artifacts may decrease the overall mutual information, even though the high classification accuracies suggest that rotations and shears preserve mammography image statistics. To solidify the correlation between mutual information and classification accuracy, future work includes investigating other augmentation strategies and deriving a better representation to capture image statistics than the mean image.

Furthermore, we see that the augmentation strategy used has a significant effect on the type of mass that is generated, as shown by the images in Figure 4. In general, we see that the augmentation determines the type of artifact that predominates the mass generation. For example, the artifacts that appear in the generated image from the CNN trained on the rotations set has circular patterns. We see a similar phenomenon for the 'flips' generated image, where we see several masses that look they have been flipped. These results suggest that using a combination of augmentations
that have high mutual information might lead to an ensemble effect where the medical image statistics are holistically captured, leading to generated images with fewer artifacts. We hope to use these findings when applying generative deep learning techniques in our future work.

Conclusion

We show that the mutual information captures the basic image statistics of an augmented dataset and roughly correlates with the performance of a CNN trained on that augmented set. Overall, our work shows that augmentation strategy greatly affects discriminative performance but also drastically affects generative performance, suggesting a strong link between discriminative and generative learning.

References


Measuring Workload Demand of Informatics Systems with the Clinical Case Demand Index

M. Sriram Iyengar, PhD¹, Deevakar Rogith, MBBS, PhD², Jose F Florez-Arango, MD, PhD¹,³

¹Center for Biomedical Informatics, College of Medicine, Texas A&M Health Science Center, Houston, TX, USA; ²The University of Texas Health Science Center at Houston (UTHealth) School of Biomedical Informatics, Houston, TX, USA; ³Escuela de Ciencias de la Salud, Universidad Pontificia Bolivariana, Medellin, Colombia

Abstract:

Introduction: The increasing use of Health Information Technology (HIT) can add substantially to workload on clinical providers. Current methods for assessing workload do not take into account the nature of clinical cases and the use of HIT tools while solving them. Methods: The Clinical Case Demand Index (CCDI), consisting of a summary score and visual representation, was developed to meet this need. Consistency with current perceived workload measures was evaluated in a Randomized Control Trial of a mobile health system. Results: CCDI is significantly correlated with existing workload measures and inversely related to provider performance. Discussion: CCDI combines subjective and objective characteristics of clinical cases along with cognitive and clinical dimensions. Applications include evaluation of HIT tools, clinician scheduling, medical education. Conclusion: CCDI supports comparative effectiveness research of HIT tools. In addition, CCDI could have numerous applications including training, clinical trials, design of clinical workflows, and others.

Introduction

In recent years, there has been increasing interest in assessing and quantifying healthcare provider workload¹-³. A major motivation is that excessive workload has a substantial negative effect on quality of care provided and patient safety⁴-⁶. The introduction of Informatics tools such as Electronic Health Records EHRs, Computerized Prescription Order Entry, CPOE systems, mobile health decision support tools, is known to contribute to workload⁷,⁸ for example by increasing cognitive and physical demands on the provider⁹-¹¹. Therefore, there is a need for methods to quantify the impact of Health Information Technology (HIT) on provider workload and, especially, to perform comparative effectiveness research of informatics tools with respect to workload.

While objective measures, such as time performances, and subjective measures like the National Aeronautics and Space Administration -Task Load Index (NASA-TLX)¹², and the Subjective Workload Assessment Technique (SWAT)¹³ have productive uses in healthcare, they do not explicitly consider an important factor: the nature of the clinical case¹⁴. For example, it seems intuitively clear that 8 hours spent in handling routine and simple outpatient cases impose less clinical demands and workload on healthcare providers than 8 hours in an intense trauma and critical care setting. This aspect of clinical tasks is not captured explicitly or adequately in current workload analysis. This means that evaluation of HIT tools is limited to task efficiency or subjective factors while ignoring the influence of the nature of the clinical problem on effectiveness of HIT tools. Quantifying workload characteristics at the clinical case level enables detailed analysis, modeling, and simulation of clinical cases. For example, when performing comparative effectiveness research of informatics tools using clinical trials it can become necessary to provide patient cases that make comparable clinical demands across study design groups.

In this paper, we present the Clinical Case Demand Index (CCDI), a tool that succinctly combines objective and subjective characteristics of clinical cases with respect to the multifactorial demands they place on clinicians. The CCDI focuses on evaluating clinical cases while the NASA-TLX and SWAT are primarily focused on evaluating subjects¹⁵,¹⁶. A major application of CCDI is to compare two or more cases and determine how similar or different they are in terms of the demands these cases make on providers.

BROADLY SPEAKING,
metric to identify such comparable clinical cases. In the following we describe such a comparative effectiveness study in some detail.

**Related Work: Workload factors in Clinical Care**

Provider workload can be measured objectively by metrics such as physical hours spent in patient consulting, number of patients seen, type of patients / diseases seen. For example, in a study, primary care physicians worked about 50-60 hours per week, with about 18 patient visits per day per physician. These sorts of measures are aggregates focused on quantity of provider services and do not account for the multifactorial demands made on clinicians by specific clinical cases and types of cases.

In addition to the purely objective measures, subjective measures of workload, developed initially for engineering and manufacturing contexts have also been applied in the clinical domain. Notable among these are The National Aeronautics and Space Administration -Task Load Index (NASA-TLX) and the Subjective Workload Assessment Technique (SWAT). The NASA-TLX incorporates mental, physical, and temporal demands as well as the operator’s assessment of his/her performance into its model of workload. The SWAT considers the operator’s time pressures, mental effort, and psychological stress. When assessing the workload, the scenario is distilled into its tasks. In this context, we note that physical workload can be defined as physical strength and resources required for completing a task. Mental (Cognitive) workload is defined as effort required for thought processes, reasoning, calculating, decision-making, memory, learning, and other cognitive activities.

NASA-TLX and SWAT are aggregates of purely subjective perceptions that do not consider the multiple characteristics of the clinical case. Some of these case characteristics are objective, such as the number of clinical tasks needed, whether the case is routine, is an emergency, and so on. Others are subjective such as perceived cognitive and physical demands.

Workload and performance in clinical cases have also been quantified using case complexity measures. Such case complexity measures consider intrinsic factors including clinical, social, and epidemiological factors. These complexity measures are used in training and research. However, these measures do not consider extrinsic factors such as provider skill and available resources, including HIT tools, for solving clinical cases. So far, workload assessment of HIT tools supporting clinical practice has been done using total time spent and conventional human factors assessment of task performance, ignoring the nature and complexity level of clinical cases. It is useful to combine both intrinsic and extrinsic factors since irrespective of its intrinsic complexity, provider factors such as decision-making skills and executing needed tasks are crucial to outcomes.

In addition, the comprehensive relationship between sources of workload and human performance evaluation does not seem to have been investigated in detail for this purpose. Florez-Arango, Iyengar, and Smith (2016) developed an ontological approach to model properties of actions performed by clinicians and related human factors. This ontological approach combines previous work describing task as goals or processes, and human factors derived from NASA TLX.

**Methods:**

**Clinical Case Demand Index**

Clinical cases are a fundamental unit in healthcare, comprising of a complex micro-environment of physical and mental processes. While solving a clinical case a healthcare provider typically performs several actions such as taking blood pressure, performing auscultation, arriving at a diagnosis, providing a treatment plan, and so on. We denote these actions as Tasks. Each task contributes to the overall demand and workload on the provider. With this nomenclature, solving a clinical case is equivalent to performing a sequence of one or more tasks by the provider or team. For example, most clinical practice guidelines (CPGs) are structured as a series of tasks to be performed considering clinical signs and symptoms. In this paper, we are concerned with the use of HIT tools while performing one or more of these tasks.

The CCDI is based on five dimensions derived from a formalism to represent tasks that are widely regarded among researchers and practitioners as contributing significantly to case demand. It includes clinical and cognitive dimensions that are related to use of HIT tools. Physical Demand (PD) and Cognitive Load (CL) are ubiquitous when assessing provide workload. The Time Pressure (TP) dimension is related to urgency as a factor for decision making.
The \textit{Expertise (EX)} dimension is a combination of importance/criticality\textsuperscript{34, 35} and familiarity/uniqueness\textsuperscript{36} as a factor for decision making. The \textit{Task Complexity (TC)} dimension counts the number of tasks required to solve the clinical case, noting that the ability to make clinically meaningful clinical decisions decreases as the number of tasks in a case increases\textsuperscript{37}. An issue here is that even for the same clinical case, the sequence of tasks that is performed can vary widely depending on the institution, the qualifications of the healthcare provider, national and local standards. To account for these variations in clinical practice the Task complexity is normalized over a given cohort of clinical cases. This means that CCDI is most meaningful over that cohort, such as within a specific care context or research setting.

CCDI consists of both a numerical summary, CCDI-S and a visual depiction, CCDI-V, of these 5 dimensions. CCDI-S provides an overall idea of the case demand and lends itself to statistical analysis, while CCDI-V enables rapid comprehension of the differential roles of the five dimensions listed above. We suggest that maximum benefit of the CCDI will be obtained by utilization of both the summary value and the visual component.

Developing the Clinical Case Demand Index

The CCDI is computed as follows based on five dimensions.

1. \textbf{Time Pressure (TP)}: This dimension of CCDI describes the temporal constraints on tasks in case. If a single task in a case must be executed immediately, for example, because the patient is in imminent risk of death, it is an \textit{emergent} case. When a single task is required to be executed to avoid death or permanent sequels it is an \textit{urgent} case. Cases with lack of emergent or urgent task are \textit{non-urgent}.

2. \textbf{Expertise (EX)}: This dimension relates to the level of knowledge or skills required to complete the case. It is rated \textit{Low} if the provider can complete the task by himself or herself without needing additional resources. If any task in the case requires the provider to call on another individual of same knowledge and skills, the task is rated \textit{Medium}. Finally, if specialized expertise, knowledge, or skills is needed, the expertise level is rated \textit{High}.

3. \textbf{Cognitive Load (CL)}: In the context of HIT tool use, this dimension is rated \textit{High} when majority of tasks in the case require recall from long-term memory. A \textit{medium} cognitive load occurs when tasks only require use of working memory such as calculations. Any other kind of cognitive processing, such as mechanical click-through and simple note-taking, is scored as \textit{low}.

4. \textbf{Physical Demand (PD)}: When a task in a case requires heavy manual labor, muscular strength, the whole body, or constantly switching positions PD is scored as high. If a task requires use of both hands to manipulate and solve the problem it is graded as \textit{medium}. When there is little or no physical effort is \textit{low}.

5. \textbf{Task Complexity (TC)}: This component is a function of the cases in the cohort under consideration, as explained above. It is a continuous variable in range 0-3.

\[
TC = \frac{n_t \cdot 3}{\max(n)}.
\]

\(n_t = \) number of tasks in the case \(i\), obtained by heuristic task analysis\textsuperscript{38} \(\max(n)\) is the number of task in the largest case.

Each factor is scored between 0 and 3 as shown in Table 1.

\textbf{Table 1—Case Demand Index Assessment Scale}

<table>
<thead>
<tr>
<th>Axis/values</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time Pressure (TP)</td>
<td>No factor</td>
<td>Non-Urgent</td>
<td>Urgent</td>
<td>Emergent</td>
</tr>
<tr>
<td>Expertise (EX)</td>
<td>No previous knowledge</td>
<td>Low</td>
<td>Medium</td>
<td>High</td>
</tr>
<tr>
<td>Cognitive Load (CL)</td>
<td>Absent</td>
<td>Low</td>
<td>Medium</td>
<td>High</td>
</tr>
<tr>
<td>Physical Demand (PD)</td>
<td>Absent</td>
<td>Low</td>
<td>Medium</td>
<td>Heavy</td>
</tr>
<tr>
<td>Task Complexity (TC)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[n_t \cdot 3\]

\[
\max(n)
\]
Clinical Case Demand Index – Summary (CCDI-S):

Summarizing the 5 axes, a composite numerical measure is calculated as shown below:

\[ CCDI - S = \frac{TP^{EX} + CL + PD + TC}{15} \]

Interpretation: The values range from 0-1. Lower numbers represent lower case demand.

Clinical Case Demand Index – Visual (CCDI-V):

The graphical component of the CCDI is a 5-axis radar plot corresponding to each of the five dimensions, TP, EX, TC, CL, and PD. In the rest of this paper we shall refer to each of this interchangeably as a dimension or an axis. Radar Plots (Figure 1) are a graphical method of displaying multivariate data with axes radiating from a central point. The plot consists of spokes radiating from a central point separated by equal angles. The spokes correspond to the axes. The axis values determine the length of a spoke. The number of axes determines the shape of the plot - from a triangle, pentagon, to full circle. Radial lines connecting the major and minor ticks in the axis can be drawn to help in both plotting and reading the graph. The values of the variables are encoded into the lengths of the axes and normalized to fit in the overall area.

Figure 1 - CCDI-Visual (CCDI-V) component of CCDI

In each of the five axes the value of that axis increases from the origin to the perimeter. A case in which all 5 dimensions have a value of 1 (typically interested as defining low demand in that dimension) would appear as a filled in symmetric pentagon with internal radii equal to 1. Asymmetric radar plots indicate that at least one dimension makes greater demands than others. For example, in the 6 radar plots shown in Figure 2 below, the top left plot shows high demand (need) for expertise and low demand for physical effort.

Evaluating CCDI for grouping cases in a clinical trial for a mobile health tool

An example application of CCDI is a mobile Health study at the simulation lab of Universidad de Antioquia (UdeA), Medellin, Colombia. The purpose of this study was to determine the benefits of presenting clinical practice guidelines in a media-rich step-by-step algorithm presented on smart phones vs the same content on paper. The study design was a randomized cross-over in which each subject managed 15 cases using the mobile health tool and another 15 using paper-based guidelines. To make the comparison meaningful the two sets of cases needed to be “equivalent” and yet different, to prevent learning effects. The CCDI was used to quantify clinical demands on 30 clinical cases and determine 15 comparison of cases per subject between treatments. The cases were selected from an existing case bank at Simulation Center and Clinical Skills lab (CS) in UdeA. Physician experts from UdeA reviewed the cases and trained Community Health Worker (CHW) performed the clinical tasks. The CHWs who performed the task were predominately male (58%), and were qualified as CHW by training or experience, comparable by their training. Mean years of training was 3.24 years.

Three physicians, as a group, conducted ontological informed task analysis over the 30 cases (10 pediatric, 10 trauma, and 10 non-trauma) and came to a consensus on the values of the five factors (axes) for all 30 cases. The CCDI was computed for all 30 cases based on the consensus values developed by the 3-physician group for each of TP, EX, TC, CL, and PD. CCDI-S and CCDI-V were calculated for the 30 cases using a Microsoft Excel spreadsheet. Based on the CCDI, cases were indexed from 1-5 and grouped under 6 categories (A & D – Pediatrics, B& E- Trauma, C&F – non-trauma). They are grouped in such a way that cases A1-D1 are equivalent, A2-D2 are equivalent and so on.
Figure 2 - CCDI-V plots of 6 cases.

Case ID - Row 1: A1, B2, C4; Row 2: D1, E2, F4

Figure 2 shows radar plots of six cases from the sample. The cases illustrated in Figure 2 are A1-Mastoiditis, B2 Burns, C4-Hypertension, D1-Severe Malnutrition, E2-Below-Knee Amputation, F4-Headache. Visual analysis, based on area and length of axes in the CCDI-V enabled grouping of the pairs of cases with respect to case demand. It is seen that A1 and D1, B2 and E2, C4 and F4 are equivalent. Using a combination of CCDI-V and CCDI-S allowed to avoid pairing cases like B2 and A1 with similar CCDI-S values but clearly different CCDI-V.

A total of 50 CHWs were recruited. We selected 5 of these CHWs to pilot the CCDI on 30 cases. Three independent physician observers noted that no additional tasks needed to be added to any cases, and that the evaluation of other axes were appropriate. Based on this review no changes to the scale or study design were made, and the data from the 5 pilot study CHWs were included in the final analysis of 50 CHWs. They were divided into two groups and each group performed 15 cases out of total 30. In Group 1, Cases A-C were included and in Group 2, Cases D-F were included. In Group 1, 21 CHWs, and in Group 2, 23 CHWs completed all the 15 cases allotted. Only these completed cases were included in the final analysis. The groups were exclusive and independent samples with no overlap of cases.

To understand whether CCDI provides meaningful values consistent with workload analyses we compared the case groupings described above with the NASA-TLX values assigned by care providers to the 30 cases as follows. One Laerdal Simman® and one Laerdal Simbaby® human patient simulators at the CS were programmed with the selected cases. The simulated cases were presented to CHW subjects. Cases were completed by CHWs in each treatment (mobile vs paper guidelines) group. At the end of each case, the subjects completed NASA-TLX questionnaire in Spanish. For each case the NASA-TLX was computed as average of all subjects’ score in the group. Consistency of grouping by CCDI with NASA TLX grouping is important to validate CCDI as a predictive measure of case demand. NASA TLX has been validated in health-care contexts as a measure of perceived workload.

We compared the CCDI to NASA-TLX using Pearson correlation, regression analysis and paired t-test. Our null hypothesis is that CCDI and NASA-TLX are not comparable with a 0.05 level of significance. We used R software package for statistical analysis.

Results:

CCDI is a valid, strong predictor of NASA-TLX with R-squared of 0.33, and Pearson correlation is 0.58. The difference between the means of CCDI and NASA-TLX would be between 0.11 and 0.01 with 95% confidence interval. The accuracy of CCDI is 0.06 - i.e. 68% of values predicated from CCDI would lie within 0.06 of the NASA-TLX value in the population. Table 2 shows that the grouping of cases by CCDI, which is a prospective, pre-task scale is consistent with NASA-TLX which is a subjective post-task measure off perceived workload.
Table 2– Statistical Analysis of CCDI vs NASA-TLX

A) Pearson Correlation:
Pearson Correlation coefficient $\rho = 0.5815037$
$t$-statistic $= 3.7823$, df $= 28$, p-value $= 0.000751$

B) Regression
Coefficients:

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>Std. Error</th>
<th>t-value</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>0.34277</td>
<td>0.04760</td>
<td>7.200</td>
<td>0.0000000776</td>
</tr>
<tr>
<td>CCDI</td>
<td>0.29925</td>
<td>0.07912</td>
<td>3.782</td>
<td>0.000751</td>
</tr>
</tbody>
</table>

Residual standard error: 0.06482 on 28 degrees of freedom
Multiple R-squared: 0.3381, Adjusted R-squared: 0.3145
F-statistic: 14.31 on 1 and 28 DF, p-value: 0.000751

C) paired t-test:
t $= -2.8946$, df $= 29$, p-value $= 0.00714$
95% CI $= -0.11200744, -0.01925922$
Mean of differences $= -0.06563333$

Discussion:
The Clinical Case Demand Index is a measure of the multifactorial subjective and objective demands required to handle clinical cases in healthcare. In Human Factor research, workload is defined as “that portion of the operator’s limited capacity actually required to perform a particular task” 40. Measuring workload helps in assessing the capacity required for optimal performance. Workload assessment is typically performed to improve productivity, reduce adverse events, plan resource allocation, training, and for quality management. CCDI serves as a tool to assess workload in healthcare. The CCDI includes both a summary numerical value (CCDI-Summary) and a succinct visual in the form of a pentagonal radar plot (CCDI-Visual). The radar plot generated helps to visualize the workload, while maintaining the weights of the dimensions. Radar plots have been used in healthcare to display multivariate data 41, 42. They show the areas of strength and weakness, as well as a general overall summary. The CCDI-V, among other benefits, provides a rapid way to differentiate cases with respect to the specific factors, TP, EX, TC, CL, and PD. CCDI can be applied in selection of clinical cases for training, evaluation, resource management and research purposes.

The CCDI-V is designed to support visual analytics, a technique that’s growing rapidly in popularity in biomedical informatics. They can be inspected rapidly both in regards to the relative sizes of the axes and the size and shape of the shaded area. For example, the relative roles of each of the five axes, representing the impact of each of the five factors making up the overall CCDI, can be assessed at a glance. If multiple cases are being considered, a side-by-side visual inspection can simply assess their similarity. For example, cases that have high cognitive load (CL) can be
grouped together if the corresponding axes on their CCDI-V plots are similar in length. The size and shape of the shaded area also rapidly yields useful case demand information. If the shape of the shaded areas of the CCDI-V plots for several cases are similar, but the sizes are different it is immediately apparent that these cases make similar demands in terms of the 5 factors, TP, EX, TC, CL, and PD, but the overall demand varies.

One of the applications of CCDI is in comparative effective research of HIT. As described above, we used CCDI to allocate cases with equivalent workload in two treatment groups. CCDI enabled presentation of cases matched with respect to difficulty level from simple to more difficult in each group. This enabled meaningful comparisons of clinician performance, learning and burnout. Thus, CCDI can be applied to create standardized cases, normalized with respect to clinical demand, for clinical student evaluation and training purposes.

Applications of the CCDI include training, evaluation, planning healthcare provider schedules, and quality management. The combined CCDI can be used in training contexts such as evaluation of resident’s skills, allotment of cases for examinations. Comparing cases based on CCDI can be useful in quality assessments, planning clinical workflow, and for insurance purposes.

Limitations:
Formal reliability and validity analyses of CCDI need to be performed. The scale was developed as part of study to train community health workers; thus, it may not represent the workload of another clinical environment. The cases used for testing the scale were subjectively chosen and may be subject to selection bias. The task complexity component involves summary of count of number of tasks, irrespective of inherent complexity of the task itself. The CCDI is designed to be applied only when at least two fully specified cases (using an ontology for example) are being considered.

Future Work:
Immediate future tasks include reproducing the evaluation in a variety of clinical settings with various types of providers including nurses, physicians, and specialists with larger sample sizes. These studies will facilitate factor analysis studies. Also, we are currently designing software tools to generate CCDI from case repositories.

Conclusion
The Clinical Case Demand Index measures the combined effects of five dimension typically contributing to provider workload in solving clinical cases especially when using HIT tools. Unlike currently used measures of workload CCDI takes into account intrinsic and extrinsic factors related to solving clinical cases. In this paper, its utility in comparative effectiveness research of a mobile health tool was demonstrated. In addition, CCDI could have numerous applications including training, clinical trials, design of clinical workflows, and others.

References


Real-time mortality prediction in the Intensive Care Unit

Alistair E. W. Johnson, DPhil1, Roger G. Mark, MD PhD1
1Massachussetts Institute of Technology, Cambridge, MA, USA

Abstract

Real-time prediction of mortality for intensive care unit patients has the potential to provide physicians with a simple and easily interpretable synthesis of patient acuity. Here we extract data from a random time during each patient’s ICU stay. We believe this sampling scheme allows for the application of the model(s) across a future patient’s entire ICU stay. The AUROC of a Gradient Boosting model was high (AUROC=0.920), even though no information about diagnosis or comorbid burden was utilized. We also compare models using data from the first 24 hours of a patient’s stay against published severity of illness scores, and find the Gradient Boosting model greatly outperformed the frequently used Simplified Acute Physiology Score II (AUROC = 0.927 vs. 0.809). We nuance this performance with comparison to the literature, provide our interpretation, and discuss potential avenues for improvement.

Introduction

The intensive care unit (ICU) admits severely ill patients in order to provide radical life saving treatment, such as mechanical ventilation. ICUs frequently have a very high staff to patient ratio in order to facilitate for continuous monitoring of all patients and ensure any deterioration in patient condition is detected and corrected for before it becomes fatal; an approach which has been demonstrated to improve outcomes1. As a result, the ICU is a data rich environment. A major effort was placed in utilizing this data to both quantify patient health and predict future outcomes, and one of the most immediately relevant outcomes to the ICU is patient mortality. The APACHE system was first published in2, and provided predictions for patient mortality based upon data collected in the ICU. While the initial system was based off expert rules, later updates used data driven methods3. Other prediction systems have also been developed, including the Acute Physiology Score (APS) III4, Simplified Acute Physiology Score (SAPS)5, SAPS II6, the Sequential Organ Failure Assessment (SOFA) score7, the Logistic Organ Dysfunction Score (LODS)8, and the Oxford Acute Severity of Illness Score (OASIS)9. For a review of severity illness scores in the ICU, see10, 11. Note that these models were universally agreed to lack sufficient calibration to be used on the individual level12, and research goals were shifted to quantify the performance of ICUs and hospitals in aggregate.

With the recent advances in both machine learning and hardware for data archiving, research has begun to return to building better prediction models using more detailed granular data. Past models have been limited by technical and practical considerations, often using summary data from an entire day of a patient’s ICU stay which was manually documented by trained personnel. Given many hospitals now have electronic data collection as a part of routine clinical practice, a wealth of data is becoming available for use in predictive modeling. The Medical Information Mart for Intensive Care (MIMIC-III) database by13 is an example of such an archive. MIMIC-III is a large collection of de-identified electronic medical records for over 40,000 patients admitted to the Beth Israel Deaconess Medical Center in Boston, MA, USA between 2001 and 2012.

Hug et al.14 investigated the use of real-time prediction models on MIMIC-II, an early version of MIMIC containing patient data up to 2008. They extracted observations for 10,066 patients and used a logistic regression model on all observations, achieving a held-out performance Area Under the Receiver Operator Characteristic curve (AUROC) of 0.885. Mortality prediction was the topic of a PhysioNet Computing in Cardiology Challenge in 2012, specifically the prediction of in-hospital mortality for patients who stayed in the ICU for at least 48 hours15. The winning entry utilized a tree based ensemble in a Bayesian framework, with components of the trees being updated using Markov chain Monte Carlo, and achieved an AUROC of 0.86016. The runner up included an ensemble of six support vector machine models built on balanced subsets of data each using the same set of all positive outcomes17.

Lehman et al.18 applied hierarchical Dirichlet Processes to clinical notes during the first 24 hours and found that the addition of extracted topics to a classifier using only severity of illness improved performance (AUROC = 0.82 versus 0.78 with only SAPS I). Ghassemi et al. extracted topics from the notes and combined these topics with static features from MIMIC in a support vector machine to classify in-hospital mortality (AUROC = 0.840)19. The authors
furthered this approach by extracting dynamics associated with topics using a multi-task Gaussian process, improving upon the AUROC of SAPS (AUROC = 0.812 vs 0.702)\textsuperscript{20}. Caballero et al.\textsuperscript{21} created a latent state model incorporating numeric features, text, and modelled topics achieving an AUROC of 0.866 during the first 24 hours on patients in the MIMIC-II database. Finally, Luo et al.\textsuperscript{22} proposed an unsupervised feature extraction model using non-negative matrix factorization, focusing on creating an interpretable model (AUROC=0.848).

The use of mortality prediction models to evaluate ICUs as a whole has found great success, both for identifying useful policies and comparing patient populations. Furthermore, as mentioned, much research has advanced the current state of the art in mortality prediction. In this paper, we propose to address the task of predicting individual patient mortality using a distinct sampling scheme. The aim is building a model which could be applied continuously for a patient, which departs from the common analysis framework of the first 24 hours or distinct daily models. This paper will proceed as follows: first, the patient population is defined and data extraction steps are outlined. Common mortality prediction systems in the literature (commonly called severity of illness scores) are then compared to machine learning approaches using data extracted from the first 24 hours of a patient’s stay. The framework for evaluating a real-time mortality prediction system is then established and we evaluate various machine learning models in this context, concluding with a discussion on their efficacy and avenues for improvement.

Data

We initially extracted data for 61,533 distinct ICU stays (all stays available in MIMIC-III v1.4). We subsequently excluded patients who met the following criteria: age less than 16 (MIMIC contains neonatal admissions), an ICU stay shorter than four hours, no data present in the patient flowsheet (likely an administrative error resulting in an incorrect ICU admission), and organ donor accounts. We also excluded patients who had an order for limitation of treatment (not full code) on or before ICU admission. Note that we otherwise retained all ICU stays, including readmissions for patients with multiple visits. The final cohort had 50,488 ICU stays. This cohort was used for both experiments (described later). These stays correspond to adult ICU admissions for surgical, medical, neurological, or coronary critical illness. Table 1 provides a description of the study population.

<table>
<thead>
<tr>
<th>Demographic</th>
<th>Mean ± S.D. or percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>63.6 ± 17.2</td>
</tr>
<tr>
<td>Male</td>
<td>56%</td>
</tr>
<tr>
<td>Height</td>
<td>169.4 ± 12.75</td>
</tr>
<tr>
<td>Weight</td>
<td>81.0 ± 24.5</td>
</tr>
<tr>
<td>BMI</td>
<td>28.4 ± 8.05</td>
</tr>
<tr>
<td>Emergency admission</td>
<td>85.6%</td>
</tr>
<tr>
<td>Service</td>
<td></td>
</tr>
<tr>
<td>Medicine</td>
<td>40.6%</td>
</tr>
<tr>
<td>Cardiac medicine</td>
<td>15.2%</td>
</tr>
<tr>
<td>Cardiac surgery</td>
<td>10.0%</td>
</tr>
<tr>
<td>Surgery</td>
<td>8.09%</td>
</tr>
</tbody>
</table>

Table 1: Demographics of the dataset (N=50,488).

Feature Extraction

For each patient’s ICU stay, we extracted data from a fixed window of length $W$ (hours) ending at time $t_{i,w}$ (hours) for each patient $i$. $t_{i,w}$ was set to some value during a patient’s ICU stay, i.e. if $t_{i,adm}$ was the time of the patient’s ICU admission and $t_{i,dis}$ was the time of the patient’s ICU discharge, then $t_{i,w} \in [t_{i,adm}, t_{i,dis}]$. For the initial evaluation of models against severity scores, $t_{i,w} := 24$ hours and $W := 24$ hours, resulting in data being collected from the first day of a patient’s ICU stay. For the second evaluation setting, $t_{i,w}$ was set to a random time during the patient’s ICU stay, and $W$ was varied from 4 to 24 hours.

Features extracted from the window of fixed size $W$ are detailed in Table 2. Features were extracted from a number of physiologic and laboratory measurements. Notably, no explicit data regarding treatment was extracted (e.g. use
of vasopressors, mechanical ventilation, dialysis, etc). Features extracted used consistent functional forms: the first, last, minimum, maximum, or sum (in the case of urine output) was extracted from all measurements of the variable made within the window $W$. As laboratory values are less frequently sampled, this window was extended 24 hours backward for these measurements. Table 2 lists the window sizes used for each of the features extracted. In addition to those listed in Table 2, a set of static features which do not change during an ICU stay were extracted. These included gender, age, ICU service, and whether the hospital admission was an emergency (binary covariate). Service was coded using one-hot encoding for the following services: coronary medicine, coronary surgery, surgery, neuro-surgery, other surgery, trauma, neuro-medicine, other medicine, orthopedic, gastro-utinery, gynecology, and ear/nose/throat. A total of 148 features were extracted.

<table>
<thead>
<tr>
<th>Time window</th>
<th>Feature extracted</th>
<th>Variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>$[t_i, w - W, t_i, w]$</td>
<td>Minimum, Maximum, First, Last</td>
<td>Heart rate, Systolic/Diastolic/Mean blood pressure, Respiratory rate, Temperature, Oxygen Saturation, Glucose</td>
</tr>
<tr>
<td>$[t_i, w - W, t_i, w]$</td>
<td>Last</td>
<td>Glasgow coma scale, Glasgow coma scale components (motor, verbal, eyes), unable to collect verbal score</td>
</tr>
<tr>
<td>$[t_i, w - W, t_i, w]$</td>
<td>Minimum</td>
<td>Glasgow coma scale</td>
</tr>
<tr>
<td>$[t_i, w - W - 24, t_i, w]$</td>
<td>First, last</td>
<td>Base excess, Calcium, Carboxyhemoglobin, Methemoglobin, Partial pressure of oxygen, Partial pressure of carbon dioxide, pH, Ratio of partial pressure of oxygen to fraction of oxygen inspired, Total carbon dioxide concentration</td>
</tr>
<tr>
<td>$[t_i, w - W - 24, t_i, w]$</td>
<td>First, last</td>
<td>Anion gap, Albumin, Immature band forms, Bicarbonate, Bilirubin, Creatinine, Chloride, Hematocrit, Hemoglobin, Lactate, Platelet, Potassium, Partial thromboplastin time, International Normalized Ratio, Sodium, Blood urea nitrogen, White blood cell count</td>
</tr>
<tr>
<td>$[t_i, w - W, t_i, w]$</td>
<td>Sum</td>
<td>Urine output</td>
</tr>
</tbody>
</table>

Table 2: Features extracted during the window examined. Blood gases and laboratory measurements have the same feature extraction (first, last), but are separated for clarity. Note that some features have had their window extended backward an extra 24 hours.

**Methods**

We evaluated multiple machine learning models both to quantify the improvement possible from more flexible approaches and to ensure robustness of our findings across multiple settings. The models used were: logistic regression (LR), logistic regression with an L1 regularization penalty using the Least Absolute Shrinkage and Selection Operator (LASSO), logistic regression with an L2 regularization penalty (L2), and Gradient Boosting Decision Trees (GB).

In addition to the above models a set of five severity of illness scores were calculated on the data. Severity of illness scores are a form of mortality prediction model where the result is an integer score correlated with mortality. Severity of illness scores evaluated here are the APS III\(^4\), SAPS\(^5\), SAPS II\(^6\), SOFA\(^7\), LODS\(^8\), and OASIS\(^9\).

For all models, the outcome evaluated was in-hospital mortality. The area under the receiver operator characteristic curve (AUROC) was used to evaluate the models. The AUROC is the probability of ranking a patient who dies higher than a patient who lives: higher values of the AUROC indicate better model discrimination, and a value of 0.5 is equivalent to random chance. We also evaluated the models using the area under the precision recall curve (AUPRC), which provides a measure of performance which is agnostic to the number of true negatives and can be useful for problems with class imbalance. Hyperparameters were set using an internal three-fold cross-validation within the external five-fold cross-validation. Once hyperparameters were selected using the internal three-fold cross-validation, the model was retrained using the entire training set and subsequently evaluated on the external validation.
fold. This process was repeated five times for each external validation fold. Hyperparameters searched were as follows: regularization penalty (LASSO/L2), the number of trees in the ensemble (GB), the learning rate (GB), and the maximum depth of each tree (GB). The other parameters were set to their default values. We used scikit learn v0.18.1 and XGBoost v0.6 with Python 2.7. Model performance is reported as the average across five held out folds in an outer cross-validation, with the minimum and maximum performance across all folds.

Experiments

Two experiments were conducted which essentially involve defining the window to be used for data extraction for each patient. In the first experiment, the window size was set to 24 hours ($W := 24$), and $t_{i,w}$ was fixed to 24 hours after ICU admission (i.e. $t_{i,w} := t_{i,adm} + W$). This emulates the most common framework for evaluating mortality prediction models used for risk adjustment. As these models aim to capture patient health on admission to the ICU (to reduce the impact of ICU care and acquire an estimate of how severely ill a patient was on admission), the data analyzed is from the first 24 hours (except labs, which are again allowed an additional 24 hours as in Table 2). We refer to this experiment as the "benchmarking experiment".

In the second experiment, $t_{i,w}$ is defined as a random time between ICU admission and ICU discharge or the first instance of a treatment limitation order, whichever is earlier, i.e. $t_{i,w} \sim \mathcal{U}[t_{i,adm}, \min(t_{i,dis}, t_{i,dnr})]$. This experiment is designed to create a model which is applicable at all possible times during a patients ICU stay, rather than traditional models which are only applicable using data from the first 24 hours. Furthermore, in order to avoid the model learning from data corresponding to the withdrawal of treatment rather than severe illness, no data is extracted after a patient’s code status is changed. For example, a do-not-resuscitate (DNR) order indicates that a patient does not desire to be resuscitated in the case of cardiac arrest. While some treatments are still given for patients who have a DNR, we took the conservative approach of excluding this data. Note that if a patient has a code status change, the data before this code status change would still be eligible for data extraction. Patients who had a code status other than full code on admission to the ICU were previously removed from this cohort as detailed earlier. We refer to experiment as the "real-time experiment". We utilized a window size $W := 4$ hours. We experimented with longer windows of 8 and 24 hours, though these did not impact model performance significantly (results not shown).

The models developed in the real-time experiment were also evaluated using windows ending at a fixed time to death. This experiment aims to evaluate the construct validity of the algorithms: as the window approaches patient mortality, we assume that the data will reflect worsening physiology. If the classifier has captured this information then the classifier’s performance should increase. To this end, the models were trained as detailed above using a random time window during the patient’s ICU stay. When evaluating the model, patients who died in the ICU had the end of the window set to 0, 4, 8, 16, and 24 hours before mortality. If the patient did not expire in the ICU, then a random time window during their ICU stay was used as was done during normal model development, and their outcome was set to zero (that is, a low prediction for a patient who dies more than 24 hours after ICU discharge is no longer penalized by the evaluation).

All code for this paper is open source and available online.

Results

Two exemplar patients are shown in Figure 1 with the two types of data extraction used for the benchmarking experiment and for the real-time experiment. The patient in the top plot survived to ICU discharge, while the patient in the bottom plot had a code status change and subsequently died in ICU. The gray background indicates the window used for data extraction, representing the benchmarking experiment in the top plot (first 24 hours of ICU admission) and the real-time experiment in the bottom plot (a random time during the ICU stay which cannot occur after a code status change).

Benchmarking experiment

The LR, LASSO, L2, and GB models were trained using features extracted as detailed for the benchmarking experiment. A total of 50,488 patients were included (the full cohort). Five fold cross-validation was used to obtain held
Figure 1: Two exemplar patients shown with vital sign data with measurement values along the y-axis. The shaded gray region corresponds to the window used for data extraction. In the top plot, the dark rectangular patch represents the window used in the benchmarking experiment (24 hours long), and the end of the window set to 24 hours after ICU admission. The lighter rectangular patch represents the additional window used for extracting laboratory measurements, which can be present before ICU admission. The patient in the top plot survived to ICU discharge.

The bottom plot represents a distinct patient with the patches representing an example window (4 hours) used in the real-time experiment hours and the end of the window set to a random time during the patient’s ICU stay before any code status changed. This patient was made do-not-resuscitate (DNR) and subsequently died in ICU.
out estimates of model performance. These models were compared to severity of illness scores: SAPS II, APS III, LODS, SOFA, and OASIS. AUROCs for severity of illness scores were calculated for each fold, and the minimum, maximum, and average AUROCs are reported, along with AUROCs for models trained in this work and models from the literature (Table 3).

<table>
<thead>
<tr>
<th>Model</th>
<th>AUROC [minimum, maximum]</th>
<th>Cohort definition and size</th>
</tr>
</thead>
<tbody>
<tr>
<td>SOFA</td>
<td>0.739 [0.735, 0.746]</td>
<td>*</td>
</tr>
<tr>
<td>LODS</td>
<td>0.755 [0.748, 0.760]</td>
<td>*</td>
</tr>
<tr>
<td>SAPS</td>
<td>0.758 [0.754, 0.765]</td>
<td>*</td>
</tr>
<tr>
<td>OASIS</td>
<td>0.774 [0.766, 0.780]</td>
<td>*</td>
</tr>
<tr>
<td>APS III</td>
<td>0.784 [0.774, 0.794]</td>
<td>*</td>
</tr>
<tr>
<td>SAPS II</td>
<td>0.809 [0.801, 0.822]</td>
<td>*</td>
</tr>
<tr>
<td>L2</td>
<td>0.897 [0.892, 0.899]</td>
<td>*</td>
</tr>
<tr>
<td>LASSO</td>
<td>0.892 [0.888, 0.897]</td>
<td>*</td>
</tr>
<tr>
<td>LR</td>
<td>0.896 [0.892, 0.899]</td>
<td>*</td>
</tr>
<tr>
<td>GB</td>
<td>0.927 [0.925, 0.929]</td>
<td>*</td>
</tr>
</tbody>
</table>

Table 3: Comparison of models trained here and in the literature with severity of illness scores. Models utilize data from the first 24 hours of a patient’s ICU stay for predicting in-hospital mortality unless otherwise stated. For models trained in this work the average, minimum, and maximum performance across five folds of cross-validation is reported. *These models used all adult ICU admissions greater than 4 hours (N=50,488) split in 5-folds of cross-validation.

**Real-time experiment**

The real-time experiment assessed the performance of models both trained and evaluated using a random time point during the patient’s ICU stay.

The performance of the LR, LASSO, RF, and GB models are shown in Table 4. These models were trained using data extracted from a window of length $W = 4$ hours located at a random time point during a patient’s ICU stay, at least $W$ hours before ICU discharge.

<table>
<thead>
<tr>
<th>Model</th>
<th>AUROC [minimum, maximum]</th>
<th>AUPRC [minimum, maximum]</th>
</tr>
</thead>
<tbody>
<tr>
<td>L2</td>
<td>0.892 [0.888, 0.896]</td>
<td>0.588 [0.568, 0.597]</td>
</tr>
<tr>
<td>LASSO</td>
<td>0.888 [0.882, 0.894]</td>
<td>0.579 [0.557, 0.593]</td>
</tr>
<tr>
<td>LR</td>
<td>0.892 [0.887, 0.896]</td>
<td>0.588 [0.568, 0.597]</td>
</tr>
<tr>
<td>GB</td>
<td>0.920 [0.918, 0.924]</td>
<td>0.665 [0.654, 0.669]</td>
</tr>
</tbody>
</table>

Table 4: AUROC and AUPRC of models trained and evaluated on distinct patients using a single window of data occurring at a random time during a patient’s ICU stay (N=50,488). The average AUROC/AUPRC across 5-folds of cross-validation is reported, with the minimum and maximum value in brackets.

The models trained above were evaluated at fixed time points from patient mortality. Figure 2 shows the AUROC at different lead times from ICU discharge.

Figure 3 shows an example of applying the model to every hour of ICU stays for four patients; two of whom survived to hospital discharge and two of whom died in ICU.
Figure 2: AUROC of models when $t_{i,w}$ is fixed at a certain distance from death for those patients who died in the ICU. If the patient survived past 24 hours from their ICU discharge, $t_{i,w}$ was set to random time during their ICU stay, and their outcome was set to 0 (i.e. they were treated as a survivor).

Discussion and Related Work

The benchmarking experiment demonstrates that the use of GB had a substantial improvement over previous methods. GB achieved an AUROC of 0.920, which is substantially higher than severity of illness scores evaluated on the same data (e.g. SAPS II with AUROC of 0.809). The improvement is likely attributable to a combination of more data and a more flexible modelling approach. The progression of the APACHE models for predicting mortality reflect this as well: later installments of the models had higher dimensionality and incorporated both cubic splines and interactions in order to better predict outcome\textsuperscript{3}. The ability of GB to assign non-linear risk across values of a single feature could also contribute to its higher performance (e.g., low and high blood pressures could be assigned distinct risk estimates).

Interestingly, the LR model was extremely competitive, achieving an AUROC of 0.892. The LASSO model was slightly worse than LR (AUROC=0.888). In medicine, a key component of any decision support tool is interpretability, and many practitioners will not trust a tool if they do not understand how it produces predictions. While higher performance is certainly a desirable aspect of the model, the use of a simpler regression model which can be interpreted still remains a promising avenue of future research.

Comparing this performance to the literature, we see that the GB had a slightly higher AUROC than published for the APACHE IV model (AUROC of 0.89)\textsuperscript{3}. It is worth noting that the models here used purely physiology, laboratory measurements, and minimal demographic variables available on admission, whereas the APACHE system utilized over 100 diagnostic categories, comorbid burden, and treatment. These additional features can often require manual collection which can complicate automatic application of the model. In a comparison of APACHE IV versus SAPS II, data abstraction per patient took much longer for APACHE IV (37.3 minutes) versus SAPS II (19.6 minutes)\textsuperscript{26}, and the difficulty in acquiring features was certainly a large factor there. The use of automatically or routinely collected features has many advantages in the potential practical application of models such as the one presented here. Further, the capability of the models to predict well given only physiology may indicate that with enough measurements regarding the patient manually laborious and potentially ambiguous tasks such as classifying patients into a single diagnosis could be circumvented.

The models developed here resulted in higher AUROCs than those reported in the literature. While it could be claimed that the results here represent the “state of the art” in mortality prediction for ICU patients, there are too many caveats to make this statement. First, we cannot claim to have exhaustively located all mortality prediction models built using the MIMIC database or otherwise. However, more importantly, we cannot even claim that the GB model is the best
of those reported in Table 3, even though all utilized some version of the MIMIC database. This is due to a number of reasons. First, there is a lack of consistency in the dataset definition and extraction, and in particular in the timing of data extraction. While most studies utilize the first 24 hours, some required the patients to stay at least 24 hours, while others utilized 48 hours and required patients to stay 48 hours. These decisions substantially impact the patient case-mix and consequently the difficulty of the problem. This is even highlighted by Luo et al.\(^\text{22}\), who utilize the first 12 hours of mortality as it is a more challenging prediction task. In addition to timing, studies frequently disregard samples due to missing data\(^\text{18, 19}\), inappropriate patients (this study removed organ donor accounts), and in order to reduce heterogeneity in the cohort (Hug et al.\(^\text{14, 22}\) removed neuro-surgical and trauma patients).

As MIMIC is a completely open database, and as tools to facilitate sharing of data analysis have matured, there is a unique opportunity for research in this area to be fully reproducible. We hope that future work will aim to alleviate this issue by releasing source code associated with their analysis. In this spirit, we have made all code for this study public and hope that this facilitates future research to improve upon the benchmark we have established\(^\text{25}\).

The real-time experiment was consistent with the benchmarking experiment in that the best performing model was GB (AUROC = 0.920), with the LR model having competitive performance (AUROC = 0.892). The AUROC of the LR model was slightly lower than the model using data from the first 24 hours (0.896), though not greatly. The L2 model had approximately equivalent performance to the LR model in all experiments.

Figure 3 demonstrates that as the models approach the time of mortality their ability to discriminate mortality increases. This is intuitively sensible and acts as a sanity check that the models are capturing abnormal physiology which relates to mortality, though it is not clear why the GB model had a much lower improvement as opposed to other models.

In the medical domain it is important to consider the interpretability of the final model developed. In Figure 3 we see an example of how the predictions of the model would result in a trend, as if used at the bedside continuously, and there appear a few spikes in the predictions. The use of a LR model would allow easy explanation of why the risk profile changed: the following covariates changed which consequently increased or decreased the final prediction. However, a similar interpretation cannot easily be produced using GB, as multiple features will have changed and translating the impact of these changes within the model is non-trivial. The interpretation of tree based models continues to be an active area of research. The trends shown in Figure 3 also show that there is a great deal of information in the trend which could be used to further improve the model performance.

There are limitations to our study. First, as all data in MIMIC is acquired from a single hospital, these models may not be applicable for outside datasets. However, the applied methodology is relatively simple, and we plan to evaluate a

---

**Figure 3:** Example trends for four patients in the ICU. Two of the patients survived to ICU discharge, while two of them died in the ICU.
similar framework in a large multi-center database. Second, the exact mechanism for utilizing such a model is unclear. While synthesizing patient health as presented could have applications for resource utilization and decision support, integrating such a model into clinical practice is a difficult process which must be made with careful consideration. Finally, the data used was a small subset of what is available in MIMIC. Specifically, we did not utilize any data regarding patient treatment, high resolution waveforms, or notes written during routine patient care. The incorporation of information from these sources may improve model performance.

Conclusion

We have shown that classic mortality prediction models can be improved by the use of flexible machine learning approaches combined with more granular data. A GB model had the highest performance here and than models reported in the literature, but we argue that the field must adopt benchmark datasets and a dedication to making code openly available code in order to both validate this claim and to allow research to progress in this field. Finally, we demonstrated a feasible architecture for developing a real-time mortality prediction system for evaluating patient health. The predictions made may provide clinicians with an accurate and rapid summary of patient health, though further research is necessary in order to ensure models are both interpretable and usable at the bedside.

References


User Requirements for an Electronic Medical Records System for Oncology in Developing Countries: A Case Study of Uganda

Johnblack K. Kabukye, MBChB, MSc1,2,3, Sabine Koch, PhD1, Ronald Cornet, PhD3

Jackson Orem, PhD2, Maria Hägglund, PhD1

1Department of Learning, Informatics, Management and Ethics, Karolinska Institutet, Stockholm, Sweden
2Uganda Cancer Institute, Kampala, Uganda
3Department of Medical Informatics, Academic Medical Center, University of Amsterdam, Amsterdam Public Health Research Institute, Amsterdam, The Netherlands

Abstract
Cancer is a major public health challenge in developing countries but the healthcare systems are not well prepared to deal with the epidemic. Health information technologies such as electronic medical records (EMRs) have the potential to improve cancer care yet their adoption remains low, in part due to EMR systems not meeting user requirements. This study aimed at analyzing the user requirements for an EMR for a cancer hospital in Uganda. A user-centered approach was taken, through focus group discussion and interviews with target end users to analyze workflow, challenges and wishes. Findings highlight the uniqueness of oncology in low-resource settings and the requirements including support for oncology-specific documentation, reuse of data for research and reporting, assistance with care coordination, computerized clinical decision support, and the need to meet the constraints in terms of technological infrastructure, stretched healthcare workforce and flexibility to allow variations and exceptions.

Keywords: Medical Oncology, Developing countries, Electronic Health Records, User requirements, Requirements engineering, User-Centered Design Methods.

Introduction
Cancer is one of the leading causes of morbidity and mortality worldwide. According to the World Health Organization (WHO), there were an estimated 17.5 million cancer cases and 8.7 million deaths globally in 2015.1 Probably more striking is that developing countries bear the largest burden of cancer, a disease which was associated with affluence. Currently over 60% of cancer cases and over 70% of cancer deaths occur in these resource-poor settings and the numbers are projected to increase. Cancer is responsible for more deaths than HIV/AIDS, Tuberculosis and Malaria combined.2
Several reasons can be given to explain the cancer epidemic in developing countries including: 1) improved life expectancy, 2) infectious diseases such as HIV/AIDS, Hepatitis B and C, Epstein Barr Virus (EBV), Human papilloma virus (HPV) which are common in developing countries and are involved in pathogenesis of many cancers including Kaposi sarcoma, Non-Hodgkin lymphomas, Cervical cancer, liver cancers, 3) lifestyles that increase risk to cancer such as smoking, being sedentary, unhealthy diets, and 4) industrialization and pollution increasing exposure to carcinogens1-3.4
In addition, the healthcare systems in developing countries are ill-equipped and ill-prepared to deal with the cancer burden. Healthcare budgets are small and most of the spending is allocated to infectious diseases (cancer is not seen as a priority intervention area). There is also lack of critical cancer care infrastructure and services such as functional pathology and radiotherapy services as well as inadequately trained and insufficient staffing. Low health literacy within the population in developing countries leads to delayed presentation and treatment abandonment, and limited research on cancers (basic science, translational and clinical trials) also negatively affects cancer care.2,5 Health information technologies (HIT) such as electronic medical records (EMR) have been postulated to improve healthcare quality, safety, effectiveness, efficiency and to reduce healthcare cost. EMRs are more accessible and can easily be shared between healthcare providers to the benefit of the patient (timeliness, better care coordination, easy consultations, reduction of unnecessary repetition of tests, etc) and electronic data can be reused for research to improve medical knowledge, and for reporting for better healthcare planning/management. Features and functionalities such as computerized order entry and clinical decision support also improve time efficiency, enhance adherence to guidelines and improve safety, and enhance translation of research knowledge into clinical practice, among other benefits6,7.

Several projects and studies have demonstrated the above benefits in a variety of contexts. In developed countries EMRs have been used in different clinical contexts including primary care, ambulatory care, oncology and for care of other non-communicable diseases; where EMRs have shown to reduce patient waiting times and improve time efficiency8,9 save costs,10,11 improve communication12 etc. In developing countries, most of the experience is in
HIV/AIDS care and EMR have also been shown to improve time efficiency\textsuperscript{13-15} increase adherence to guidelines\textsuperscript{16} improve care coordination and reduction in loss to follow-up\textsuperscript{17,18} among other benefits. These benefits if harnessed could potentially contribute to improved cancer care and reduction of the cancer burden in developing countries. However, the adoption of EMRs in oncology remains poor, especially in developing countries.

One reason for low adoption and/or failure of EMRs in general is that many EMRs do not meet the requirements of the users. Requirements of a system are the attributes, capabilities, characteristics or qualities that a system must have for it to be of value to the user - the services that a system is expected to offer and the constraints under which it must operate\textsuperscript{19-21}. Requirements depend on the context of use i.e. the users (their experiences/skills, attitudes, prejudices, etc), tasks and equipment (hardware, software and materials), and the physical and social environments in which the system is used\textsuperscript{22}.

Oncology is a complex medical specialty with unique workflows, information management needs and challenges (chronic, multidisciplinary and multi-modality care, expensive and toxic treatments such as radiotherapy and chemotherapy, complex schedules and combinations, etc) thus it poses unique requirements on an oncology EMR. In developed countries such as the United States and the United Kingdom, there have been efforts to define the unique requirements for oncology EMRs so as to guide vendors to develop usable EMRs and users to select those that are likely to meet their need.\textsuperscript{23-28} But due to differences in socio-economic status and healthcare system organization/workflows (available technological infrastructure, policies, human resource, standard operating procedures, etc) the requirements for an oncology EMR for developing countries are likely to be different compared to developed settings. Yet there have been no studies on requirements for an EMR for use in cancer care in low-resource settings.

**Purpose:** To elicit and analyze the requirements for an EMR suitable for use in oncology in developing countries. The study was party of the Master's thesis of the first author (JK), and contributes to the ongoing EMR implementation process of a national cancer hospital in Uganda, the Uganda Cancer Institute (UCI).

**Methods**

A qualitative study was carried out, consisting of a focus group discussion (FGD) and follow-up one-on-one interviews. These were held in English. This approach was chosen because of the explorative nature of the study. Requirements engineering (RE) as a key component of the user-centered design (UCD) process\textsuperscript{27} requires understanding the perspectives (views, opinions, attitudes, limitations, etc) of the users and the context of use of the system\textsuperscript{22} and FGDs and interviews allow for this\textsuperscript{28,29}. Since requirements are for the same system, some form of consensus is needed which can be got during FGD in addition to allowing a comprehensive exploration of views from different people, as opposed to interviews with one person. However, in-depth interviews allow for detailed discussion which is important in RE to understand the users' challenges and wishes so as to come up with a usable solution, since often the users don't know exactly what the solution or requirement is or there are misunderstandings and miscommunications.\textsuperscript{30}

The FGD followed a semi-structured format where probes were used that sought to understand the different tasks that the target EMR users routinely do, the information they collect and how it is exchanged/reused, as well as challenges faced in cancer care and how these could be alleviated by the EMR. These probes (available as supplement up on request) were developed prior to the study by JK basing on literature review, and they were discussed and agreed up on with MH and SK who are senior researchers in health informatics. The follow-up interviews were unstructured but followed up on issues that had been raised during the FGD.

Considering that the study site is at an early phase of EMR implementation and the participants have limited experience with EMRs, it was suspected that they might have had limited opinions on requirements. To give the participants more insight about EMR and RE, a two hour workshop was set up prior to the FGD where JK presented about these concepts. In addition, a mockup of RE was done for a chemotherapy management system, and a representative of the EMR vendor Clinic Master International\textsuperscript{†} (an EMR used by several hospitals in Uganda) was invited to give a demonstration of the EMR during this workshop.

**Study site:** The study was conducted at the Uganda Cancer Institute (UCI) in Kampala, Uganda. The UCI is a 100-bed public cancer hospital established in 1967 by the Uganda Ministry of Health and Makerere University in collaboration with the US National Cancer Institute (NCI). It was an important center for research and advancement of cancer treatment especially for combination chemotherapy and description and treatment of Burkitt's lymphoma, Kaposi's sarcoma and other cancers common in this area. The UCI has grown to become the main cancer hospital in

---

\textsuperscript{†} http://clinicmaster.net/
Uganda and in the East African region, with over 240 staff, about half of whom are clinicians (29 doctors, 58 nurses, 7 lab staff and 5 pharmacy staff, among others). Currently the UCI is in the process of implementing an EMR.

**Participants:** Participants were selected strategically and purposively to represent the main target end user of the EMR at the UCI i.e. doctors, nurses, pharmacists, medical records officers, lab technicians, etc. These user groups interact directly with patients hence they understand the work processes and dynamics of patient care at the institute including its challenges. They will also be the ones most likely to directly interface with the EMR once it is implemented.

Nine of the 12 participants approached agreed to take part, with 4 males and 5 females and age range between 27 and 51 years. All had worked at the UCI for at least 2 years each and were familiar with the work processes. All the participants had computer literacy at least to the level of comfortable day-to-day use such as internet and email, Office suite and have had exposure to tasks such as data entry for clinical research. Some had also used the Clinic Master EMR system in the past, but not in the setting of the UCI, rather in other private hospitals.

The table below shows the professions/job of the participants and whether they had used an EMR software before.

<table>
<thead>
<tr>
<th>Number</th>
<th>Gender</th>
<th>Title/Job</th>
<th>Prior exposure to an EMR</th>
</tr>
</thead>
<tbody>
<tr>
<td>01</td>
<td>F</td>
<td>Junior medical officer</td>
<td>No</td>
</tr>
<tr>
<td>02</td>
<td>M</td>
<td>Doctor/Intern</td>
<td>Yes</td>
</tr>
<tr>
<td>03</td>
<td>F</td>
<td>Doctor/Surgeon</td>
<td>No</td>
</tr>
<tr>
<td>04</td>
<td>F</td>
<td>Senior Nursing officer</td>
<td>No</td>
</tr>
<tr>
<td>05</td>
<td>M</td>
<td>Pharmacist</td>
<td>Yes</td>
</tr>
<tr>
<td>06</td>
<td>F</td>
<td>Nursing officer - Pediatrics</td>
<td>No</td>
</tr>
<tr>
<td>07</td>
<td>M</td>
<td>Doctor/Pediatrician</td>
<td>Yes</td>
</tr>
<tr>
<td>08</td>
<td>F</td>
<td>Medical records officer</td>
<td>Yes</td>
</tr>
<tr>
<td>09</td>
<td>M</td>
<td>Nursing officer - Community outreach</td>
<td>No</td>
</tr>
</tbody>
</table>

One FGD lasting two hours was conducted, followed by five follow-up interviews with participants 01, 02, 05, 06, and 08 in the subsequent week after the FGD. The interviews lasted average 25 minutes each.

**Data collection:** Data was collected as audio recordings of the FGD and follow-up interviews that were later transcribed verbatim. The FGD was moderated by an independent moderator who is experienced in qualitative research (a nurse). This moderator was first trained by the author in advance to make sure the moderator fully understands the probes to be followed in the FGD. The moderator also attended the workshop prior to the FGD. The author was present in the FGD as a note taker, and he also conducted the follow-up interviews which followed the notes taken during the FGD - to clarify on issues that were not fully explored in the FGD.

**Data analysis:** The audio recordings were transcribed verbatim by the first author and printed out for manual content analysis. Throughout the process of transcription and from the FGD/interviews, the author gained insight into the contents of the transcripts by the time they were printed. The author then read all the transcripts carefully and thoroughly to identify meaningful units and assigned codes/themes following a deductive approach\(^{1, 2}\). Statements that mentioned tasks, challenges or wishes that could translate into EMR functionality or constraint were identified and a mind map made (available as a supplement). A logical workflow model was also made to summarize the key tasks (see figure 1). The figure summarizes a typical flow but variations and exceptions are common among individual patients e.g. some new patients who aren't very ill can do their staging workup as outpatient and some patients continuing with treatment occasionally get admitted if they become ill due toxicity of chemo or radiotherapy or when they are to undergo surgery.

Requirements were derived\(^ {3, 4}\) following categories of EMR functionalities/features in literature especially in\(^{27}\). Throughout this process, JK had several meetings with MH and SK who were his academic supervisors, to discuss the findings and translate requirements from the transcripts.

**Study period:** Data collection was done in February and March 2016 and analysis was done in April and May 2016.

**Ethics:** The study was approved by the Uganda Cancer Institute Research and Ethics Committee (#UCI REC REF 01-2015). All participants gave written informed consent and their privacy and confidentiality was ensured during the study.
Results
The user requirements were derived from understanding the target users' workflow and tasks they need to accomplish as they described them in the sessions, from the challenges and constraints as well as their wishes and what they envisioned if an EMR had would make their work easier or the adoption and usage of the EMR easier. Detailed descriptions of the requirements with explanation or motivation and illustrative quotes from the participants are available as a supplement up on request. The requirements can be categorized as follows: Documentation, Communication and care coordination, Computerized order entry and results management, Computerized clinical decision support, Inventory/Stock management, as well as "Meeting the constraints (Fit into context)".

1. Documentation.
1.1 Routine documentation: When discussing the workflow and tasks, documentation was repeatedly mentioned as crucial routine. The system should support clinical documentation including identification and demographics, clinical history and findings, treatment details, etc, as these are routinely recorded in any clinical setting. The added benefit expected from the EMR however is alleviation of the challenges that come with paper documentation such as paper getting lost or torn or being bulky to store and being inaccessible especially in the complex setting of multidisciplinary and multi-service cancer care. Another challenge that was pointed out is that the current paper files are not structured and documentation not standardized with resulting inconsistence and incompleteness, making the data hard to compute and reuse downstream e.g. in CDSS or for research and automatic report compilation. The EMR should thus address this challenge.

"...we have no unified documentation system. Every oncologist reviews the patient the way he wants. You can write 2 lines, I can write 4 lines, another can write 8 lines. 4, no standardized [documentation]... I feel that if you have a standard template of what we want, every clinician to ask a patient of a class of disease, then such issues of omissions will be minimized... you'll have a consistent standard information... the computer... should be able to give us all the detailed headlines of what we are supposed to do... such that we have consistence of information. If am reviewing the patient, I must enter something in all these areas, otherwise it will not go to the next page." -- Participant 07, Pediatrician
"We write on papers... that you just punch and fix in the files, so many times they can pluck out... So many times you find that you look for a chemotherapy recording paper, you cannot find it. this patient has been on treatment for like 3 months so that means there should be some recording and you can't find it." -- Participant 06, Nurse

"... right now we don't have enough space in Records [storage room]..." -- Participant 08, Medical records officer

"...we have a challenge of files getting lost... and that usually comes for example a file will go to OPD and then somebody picks it for example for research -- Participant 08, Medical records officer

1.2 The EMR should support oncology specific documentation: Oncology as a medical specialty has differences from other specialties and these also show in the documentation. There are oncology specific data such as cancer diagnosis, tumor descriptions (size, location, metastasis, etc), cancer stage, cancer risk factors, special investigations (e.g. biopsy/histology results and metastatic workup), treatment (chemotherapy, radiotherapy), in addition to emphasis put on "routine" data such as patient height and weight which are important for calculating body surface are onto which chemotherapy prescriptions are tightly based, vital signs which often change frequently in cancer patients due to the severity e.g. neutropenic sepsis, drug dose to track ceiling/life time doses, etc.

"... the UCI is a specialized center, .. it's not... where a malaria will walk in... It's a specialized center. So, the system should be tailored to an oncology center... frequency of investigations that are done at the UCI. It's much more...and the importance. Down there [referring to an non-cancer referral hospital] sometimes people request. RFTS, LFTS, just for just. But here, we request things which we actually want to use...the system should incorporate something that can help us to analyze our investigations. Like, you know, CBC and chemistries, ok? And also a way to sort them out... which one is acceptable, which one is not acceptable, which one needs attention..." - Participant 01, Junior doctor

"...a provision where we can fix our observations as a nurse... the chemotherapy sheets or checklist or whichever or road map as per patient treatment schedule.. and also this other supporting treatment cause, well most of them are admitted not for chemo but for other reasons for example neutropenic sepsis..." - Participant 06, Nurse

The quote above also points out an important aspect about cancer care in that patients commonly have multiple medical conditions and co-morbidities besides the malignancy making documentation complex. In addition, care is chronic and information intensive so the participants mentioned a wish for the EMR to offer efficient ways of documentation e.g. by having pre-populated menus which they just check off.

"...if there was a check-check-check...that would be very easy." -- Participant 02, Doctor (Internist)

2. Communication and care coordination
2.1 The system should enable/facilitate communication between the different care teams including the clinical service points (Outpatient department and wards), laboratory, imaging/radiology, radiotherapy, surgery, etc which are not co-located. The EMR should connect and transmit messages between all these service points, so that each knows what the other is doing for the patient for better coordination, to avoid delays, conflicts and medical errors (such as doing surgery in a patient who has just received chemotherapy or radiotherapy, or giving chemotherapy to a neutropenic patient), and to ensure continuity of care.

"...the time spent during movement, usually we have very ill patients and as a clinician you're moved to walk to the lab to get the report ... there's a lot of time lost that could be reduced." - Participant 02, Doctor (Internist)

"... many patients are sometimes lost up within the system because as you know this place UCI has more than three campuses ... and some patients tend to cut across all these areas. For instance a patient is seen by a surgeon, the surgeon does mastectomy ... sends the patient for radiation at the lower campus... the radiation oncologist .. attends to the radiation part of it and if the patient is not well informed that you need further chemotherapy, that patient may be lost to follow-up" -- Participant 02, Doctor (Internist)

"...I would have loved having a prescription coming to me electronically with lab results ... that is very important because before I go ahead to prepare chemo I must know that the patient is suitable for chemo..." -- Participant 05, Pharmacist
2.2 Scheduling: Participants highlighted challenges with scheduling of patients, with "patients tend to walk in sometimes on days that are not theirs" and thought that "the computer system would be important in helping with patients' scheduling dates.. when they're supposed to come back". The EMR should therefore have functionalities to assist/coordinate scheduling of patients for the different appointments and care activities during active treatment and follow-up, according their care roadmap, which is fairly standard.

2.3 Communication with patients: Also due to the complexity of the treatment roadmaps and chronic nature of cancer care, loss to follow-up are common at the UCI and participants need a system that can assist in tracking patients and communicating with them e.g. through calls or SMS reminders. This also enables collecting survivorship data in addition to allowing better preparation and coordination of clinic visits.

"...but we have people who got lost to follow up... people who died at home... we have to be able to track patients... a way of contacting patients.. maybe somewhere where we can incorporate a patient's contact.. maybe phone numbers.. to be able to call them.. also because our patients come periodically we have to be able to incorporate into the system a scheduling system for instance, if someone is on a 3 weekly schedule of chemotherapy, the system should be able to .. I don't know how it works but it should be able to say patient A, patient B, patient C, patient.. D are the ones who are supposed to come tomorrow or the other day, so that it helps one, the records people to prepare the files, 2, the nurses to know how many patients are coming, 3, to deal with people who.. unscheduled visits" - Participant 01, Junior doctor

3. Order entry and result management
The system should assist in making orders for investigations and in managing the results - quickly selecting the orders (e.g from a predefined list) or automatically generating orders basing on previous information on a patient such as their histological diagnosis; communicating the order to appropriate target (e.g. lab or imaging) and communicating back the results to the sender in time. This is key because there is a myriad of tests done along the roadmap of care for the cancer patient, some at particular times and others routinely and more efficient ways of ordering them and transferring of results is needed to avoid wastage (e.g. in repetition of tests) and delays.

"... communication between department, If I order for hemogram or chemistry, how quickly do I get results?...you need the renal function but you find it was ordered yesterday you don't have it yet you order it again, it is lost somewhere either is in the machine not printed yet because no paper or it is printed and kept in the file or printed and lost on the ward or printed and somebody's pocketed it. So some information that has been ordered for doesn't affectively get to the patient chart, which delays care and leads to misuse of resources..." --Participant 02, Doctor (Internist)

Other orders such as instruction sets for a particular task or patient were also pointed out as sources of inefficiency that the EMR could alleviate by automating or reusing existing data or order sets, as exemplified by this quote:

"There is a lot of repetition of things for example commonly we do feeding gastrostomies and there's a feeding protocol after, which is actually similar in all the patients, at least most of them. So you write for this patient Day 1 you give this and this and this day 2... then you if you've done 3 patients, maybe 5 patients on that day, you do that for each of the patients that you've done" - Participant 03, Surgeon

4. Computerized clinical decision support (CDSS) and safety checks
4.1 Adherence to protocols/guidelines and filling in knowledge gaps (expert system): Participants gave several examples of scenarios in which they would need assistance from the EMR in form of CDSS. The want the EMR to assist clinicians in making a cancer diagnosis and stage (according to clinical and investigations), and in making treatment decision (according to stage, co-morbidities, drug availability, drug interactions and contra-indications, etc). The CDSS is needed because some clinicians "lack specialized knowledge" or "oncology specific training", in addition to the complexity and large amount of information that has to be synthesized and acted up.

"... a patient comes with breast cancer and you order for chest x-ray of abdominal scan, hemogram, liver function, renal function,. the essence of all those tests is to map out that the disease stage. So if this program is able to get all this information that we feed it in at the end of the day it prescribes a specific stage, this would help us to treat a patient better. To follow up with that, once a stage of the cancer is prescribed then we need more guidance with the specific treatment." - Participant 02, Doctor (Internist)
"...once a nurse or somebody makes an observation and they are entered patient's e-file then probably they can give a siren or can give a warning that you know this temperature you need to act this BP you need to act." - Participant 02, Doctor (Internist)

Because many protocols and guidelines used at the UCI are adopted from developed countries where clinical trials and research are performed, the logic for the CDSS should be according to accepted modifications of such or local consensus because there are observed differences in patients in the two contexts as this participant explains:

"Nutritional issues may vary, immune status may vary, especially when I deal with children. So, if I get the protocol developed in the UK or the US to be used here,... we must modify them based on some small criteria like nutrition status of the patient." -- Participant 07, Pediatrician

4.2. Chemotherapy safety checks: Participants reiterated the complexity and difficulty of chemotherapy management including correct dose calculation, monitoring of ceiling doses, side effects and contra indications such as allergies, co-morbidities that might require modification of doses e.g. HIV/AIDS, malnutrition or organ dysfunction (e.g. renal failure) or cytopenias. One requirement they had of the EMR therefore is for it to assist in checking these to ensure safe chemotherapy use because errors can be detrimental.

"It's really very easy for someone to write chemotherapy twice... but if we had a system that has everyone's name once, ok? and you know these are the people for chemotherapy... a centralized system that has all the patients.. But we have a very erratic recording system that allows for a lot of mistakes." - Participant 01, Junior doctor

"If the patient was treated with doxorubicin the program gives an alert: echo at six months, echo at 1 year, echo every year after treatment..." - Participant 02, Doctor (Internist)

5. Inventory/stock management

The system should keep track of stock of medications and supplies, and update the different users (prescribers, pharmacists, etc). Participants need this functionality because stock management is hard in the current paper-based system, yet it greatly affects patient care. Prescriptions and protocols for treatment are chosen in part basing on available drugs, and stock records are needed to make orders for drugs and other supplies in time to avoid stock-outs.

"...to manage stock and supplies.. monitor drugs in my store rather than going there in the last minute and am putting in a request in the store and they tell me it's out of stock, it got finished, I gave you the last stock last week... then you can be able to prioritize...in terms of ordering... Then the other thing would be the cost attached to that because in most cases we don't know these costs. It may not be that patients are going to pay but it is important to know how much the government is spending on the patients..." -- Participant 05, Pharmacist

6. Meeting the constraints (Fit into context)

Participants mentioned several conditions and pre-requisites that an EMR suitable for their context should meet. These stem mainly from the fact that the UCI is in a low-resource setting, while others are due to the unique workflow of cancer care at the UCI. The following are the key constraints: i) The system should allow for mobility or "computer on wheels" because users need to access the system at the bedside during ward rounds. ii) The system should fit into the technological infrastructure e.g. working with unreliable internet connection and electricity supply (or backups should be in place. As one participant noted, "In 24 hours, there are 10 blackouts in Uganda... Even when you talk about a backup generator, with fuel, this fuel must come from ministry of energy". iii) The system should allow flexibility to accommodate variations and exceptions that are common in healthcare, and to accommodate new requirements that may arise (such incorporation of new medical knowledge). iv) In addition, functionalities such as CDSS should not take over control from the clinician. Authorized users should be able to make modifications to the rules/functionalities.

"...there should be room for clinicians to have control over system... and I think that also still need to be restricted, not every clinician should have... because if that is the case, a nurse can sit somewhere and modify a protocol, a pharmacist can sit somewhere and single handedly modify a protocol. I think the lead clinicians who are authorized by the Institute to actually prescribe should be given the administrator access administrator password to modify some protocols." -- Participant 07, Pediatrician
Discussion

This study aimed at eliciting and analyzing the user requirements for an EMR for the UCI as a cancer hospital in a low-resource setting. A user centered approach was taken where target end users were involved in a FGD and interviews to analyze the workflow, their challenges and wishes and to derive the requirements. The key findings were that cancer care is complex - chronic, multidisciplinary with many care points, and it is faced with many challenges that a usable EMR could help in solving. The target end users thus require an EMR that supports structured and standardized clinical documentation, and in particular, oncology specific documentation such as cancer diagnosis, stage, tumor descriptions and treatment details (such as chemotherapy roadmaps) among other things. Consistent and complete structured data can also be reused for reporting and research which are key tasks/processes for the UCI. Participants also require functionalities to assist/support tasks such as scheduling and coordination of the complex, multi-disciplinary care for cancer patients, CDSS to compliment the limited expertise in oncology and offer automated safety checks as well as taking stock for better planning. The system should however fit into the context and function despite such limitations as unreliable internet and electricity supply, it should be flexible to allow exceptions/variations (such as authorized modifications of the guidelines/protocols on which CDSS is based) and shouldn't take over control or decision making from the clinicians.

On a high level, these requirements are similar to those described for oncology EMR in resource-rich countries possibly because cancer care is quite similar in terms of care coordination, scheduling, documentation needs, etc. However, due to differences in the context, extra constraints and often context specific requirements are posed to the EMR on a detailed level. For instance, technological infrastructure such as internet and electricity supply is not usually considered a key issue in HIT in developed countries yet is it a major constraint in developing settings. Policy and organization of healthcare e.g. reporting requirements to ministry of health or other stakeholders or the procurement process for drugs and supplies (drugs are procured by a central national body on behalf of the hospitals so timely orders from this national medical store are required) also call for EMR features and functionalities at the UCI to serve these needs hence requirements like stock management which might not be emphasized in oncology EMR specifications for other contexts. Moreover, there is more emphasis on CDSS because of the overstretched healthcare workforce and limited number of cancer care experts making automation very crucial if cancer care in this setting is to be improved. Users' experience e.g. prior use of EMR or computer skills which are generally less compared to developed settings might lead to usability challenges when a system is developed, in addition to limiting the opinions of the participants during requirements elicitation. Demographics and identification data can also be different e.g. use of social security numbers or RFID in which was never suggested or not feasible in the context of UCI because such systems aren't in place. Other key requirements specified for oncology EMRs in developed countries that were not mentioned or emphasized included billing features (possibly because the UCI is generally a public facility and healthcare is free), clinical trials support (because the UCI is not involved in many complex clinical trials), and provisions for a patient portal or mechanisms for patients to access their records and contribute to them e.g. through patient reported outcomes (because health and computer literacy levels are relatively low in developing countries and patients probably wouldn't use this feature even if it was available. Besides, there were no patient representatives among the participants).

It is worthwhile noting that some of the challenges and wishes by the participants in this study do not necessarily need a technological solution and are rather organizational. For instance, proper documentation (consistence and completeness) should be organizational culture with or without an EMR. Implementing an EMR before such issues are streamlined might actually lead to failure of the EMR even when it is in fact a good system. Nevertheless, the RE process allows organizations to review their workflows and identify inefficiencies. As an example, a key organizational challenge that came out during the data collection in this study was the fact that there was no triage area and patients would first meet with the records personnel on arrival at the UCI. Because the records personnel aren't clinical, there would be challenges of dealing with patients who came in critical conditions and needed urgent evaluation. This challenge has since been addressed by setting up a triage area at the UCI where new patients report first and a doctor is present to do a quick assessment. It is also common, especially in the initial phases of implementation, that EMR implementation leads to reduced productivity and inefficiency and this might further negatively affect care in the setting of an overstretched healthcare workforce. An analysis of such potential impact on productivity or workflow changes was beyond the scope of this study but needs critical thought during EMR implementation process.

Strengths of the study: Target end users who understand the workflow and intended context of use of the EMR were involved as participants (UCD). In addition, two qualitative methods, FGD and interviews were used, and
participants were from different medical backgrounds representing different perspectives. Moreover, involvement of the target end users also has an advantage of giving them ownership which enhances buy-in.

**Weakness of the study:** Only opinions can be obtained from FGDs or interviews and before these are implemented and tested, one can never be sure of how users will perceive and if they will accept it. More pragmatic and interactive RE techniques such as prototyping need to be applied, and more importantly the process needs to be iterative until usability is achieved.

**Future steps**
This study elicited a comprehensive list of user requirements from the perspective of different stakeholders involved in documentation and care of patients at the UCI without any form of prioritization. Since all the requirements might not be delivered at once, a next step of prioritization is needed e.g. in terms of which requirements, if delivered, provide maximum benefit. The resulting prioritized requirements should then be availed to EMR developers/vendors such as Clinic Master Inf1 to guide customization of their product to fit the UCI's needs. In addition, an analysis is needed of other organizational and human issues such as training, availability of human resource, funding and other support from management, workflow changes, etc which influence (or might be influenced by) implementation of an EMR, as well as engaging potential users beyond the UCI to ensure generalizability.

**Conclusion**
Oncology is a unique medical specialty and cancer care is complex with many challenges in terms of documentation, care coordination and need for support from "expert systems" to ensure patient safety and standardized care based on guidelines, among others. If an EMR is to be used in the context of low resource settings, extra constraints are imposed to it resulting from the resource limitation (limited funding, lack of human resource, technical infrastructure and skills) and differences in organization of healthcare (reimbursement/billing, procurement process for medication, etc). These context-specific constraints and requirements have to be met to ensure usability and acceptability of the EMR, and might require customization of existing EMRs or specifically developing one targeted for this context.

**Acknowledgments:**
- The Swedish Institute (www.st.se) for the scholarship that funded the masters studies for the first author
- Uganda Cancer Institute for funding the logistics of this study including travel from Sweden to Uganda
- The participants who volunteered to take part in this study

**References**

**Contact for correspondences:** Johnblack K. Kabukye on jkabukye@gmail.com
Mechanism-based Pharmacovigilance over the Life Sciences Linked Open Data Cloud

Maulik R. Kamdar, M.Tech., Mark A. Musen, M.D., Ph.D.
Center for Biomedical Informatics Research, Stanford University, CA, USA

Abstract

Adverse drug reactions (ADR) result in significant morbidity and mortality in patients, and a substantial proportion of these ADRs are caused by drug–drug interactions (DDIs). Pharmacovigilance methods are used to detect unanticipated DDIs and ADRs by mining Spontaneous Reporting Systems, such as the US FDA Adverse Event Reporting System (FAERS). However, these methods do not provide mechanistic explanations for the discovered drug–ADR associations in a systematic manner. In this paper, we present a systems pharmacology-based approach to perform mechanism-based pharmacovigilance. We integrate data and knowledge from four different sources using Semantic Web Technologies and Linked Data principles to generate a systems network. We present a network-based Apriori algorithm for association mining in FAERS reports. We evaluate our method against existing pharmacovigilance methods for three different validation sets. Our method has AUROC statistics of 0.7–0.8, similar to current methods, and event-specific thresholds generate AUROC statistics greater than 0.75 for certain ADRs. Finally, we discuss the benefits of using Semantic Web technologies to attain the objectives for mechanism-based pharmacovigilance.

1 Introduction

Pharmacovigilance methods are used to detect unanticipated adverse drug reactions (ADR) that manifest due to the intake of drugs by patients. These ADRs are often not detected during the clinical trials of the corresponding drugs. A majority of these ADRs are caused by polypharmacy, a situation where multiple concomitant drugs are administered to one patient in a short span of time to treat multiple medical conditions.1 These drugs may interact with each other through several different underlying biological mechanisms.2 Drug–drug interactions (DDI) due to polypharmacy are potentially avoidable, if detected early.3 ADRs are the 4th leading cause of death ahead of diabetes, AIDS, and pneumonia.4 ADRs often result in the hospitalization or serious injury of more than 2 million individuals in the United States, with more than 100,000 deaths annually.5 The costs of drug-related morbidity and mortality in the United States alone were estimated to be US$177.4 billion in 2000, and have been rising ever since.6

Pharmacovigilance methods often use data from Spontaneous Reporting Systems, such as the US Food and Drug Administration (FDA) Adverse Event Reporting System (FAERS),7 or electronic medical records.8 These methods have inferred new DDIs and the ADRs that manifest on the account of those interactions (e.g., Vioxx → Heart Attack and Aspirin + Warfarin → Bleeding). However, these studies do not systematically demonstrate how the drugs interact within the biological system of the patient, leading to a particular adverse reaction. Mechanism-based pharmacovigilance can lead to the inference of newer DDIs and ADRs, and can also provide a better understanding of the underlying biological mechanisms behind the DDIs.9 Moreover, this understanding can lead clinicians to prescribe drugs that can treat the same medical conditions in a patient while minimizing the risk of DDIs due to different mechanisms of those drugs. The objectives of mechanism-based pharmacovigilance can be attained through the development of network-based approaches of integrative pharmacology, often termed systems pharmacology.9 These approaches rely on an exhaustive systems network, that possesses knowledge of the drug-induced perturbations of the physiological functions in a biological system as well as knowledge of the underlying biological interactions.

However, the data and knowledge to generate such a network exists in several databases and knowledge bases that may be fragmented across the Web. These sources, if available for download, may: i) use varying schemas to structure the data, ii) use different entity notations (e.g., proteins referenced using HGNC10 or KEGG11 identifiers), and iii) use different formats for storage (e.g., XML, CSV, etc.). An ad hoc integration approach involving downloading and integrating each source independently, and reconciling similar entities, is non-trivial, non-scalable and is often redundant for different tasks. Hence, the objectives of mechanism-based pharmacovigilance are yet to be realized.
1.1 Semantic Web Technologies and Linked Data

The Semantic Web was conceived with the vision that a decentralized, distributed and heterogeneous data space, extending over the traditional Web, can reveal hidden associations that are not directly observable. Semantic Web technologies and linked data principles enable the representation, linking and querying of data and knowledge on the Web. Semantic Web technologies include the W3C standards Resource Description Framework (RDF) and the SPARQL graph query language. Due to the challenges of integrative bioinformatics, biomedical researchers have been the earliest adopters of Semantic Web technologies and linked data principles to create the Life Sciences Linked Open Data (LSLOD) cloud. Biomedical data and knowledge sources are converted to graphs using the RDF model. SPARQL can use specific graph patterns to query these RDF graphs. Several different efforts publish and link biomedical data and knowledge in the LSLOD cloud (e.g., Bio2RDF). Several sources that may be relevant to systems pharmacology, such as PharmGKB and DrugBank, are made available through the LSLOD cloud.

An example of an RDF graph that represents the following information — “Gleevec (Mol. Wt.: 589.25, Half-Life: 18 hours) inhibits PDGFR (platelet-derived growth factor receptor), involved in signal transduction” — is shown in Figure 1a. Here, similar entities (e.g. Gleevec) in two different sources, DrugBank and the Kyoto Encyclopedia of Genes and Genomes (KEGG), are linked together using explicit x-ref (cross-reference) attributes. Attributes and relations stored in these data sources are represented as nodes and edges in the RDF graph. Moreover, as shown in Figure 1b, the following query — “List drugs that have Mol. Wt < 1000 and inhibit proteins involved in signal transduction. Mention their half-life.” — can be executed using SPARQL. The graph expression patterns are derived from the RDF schema that is used to structure the graphs, as well as the x-ref attributes.

It should be noted that these biomedical RDF graphs may be exposed through isolated SPARQL endpoints on the web. Querying multiple isolated SPARQL endpoints simultaneously over the web requires a scalable SPARQL query federation method. Different graph patterns may be used to represent the same relation type. In Figure 1c, the relation Drug hasTarget Protein is represented using different labels (drug-target and target) and different graph patterns in DrugBank and KEGG respectively. In the latter case, as KEGG is a pathway data source, the RDF graph also captures the type and provenance of the interaction between the Drug and Protein. Query Federation methods can transform a given query to source-specific queries and retrieve information from two or more sources simultaneously.
In our previous research, we had developed such a query federation architecture, termed PhLeGrA – Linked Graph Analytics in Pharmacology, over the LSLOD cloud. PhLeGrA uses prior knowledge on such graph patterns to generate a systems pharmacology network by retrieving data from four different biomedical sources. As there is minimal overlap between different sources for drugs and drug–protein relations, we had also demonstrated how query federation over the LSLOD cloud can help systems pharmacology approaches.

In this research, we extend the PhLeGrA architecture, with an improved inference module to detect drug–drug interactions and adverse drug reactions. The module will also assign confidence scores to all possible underlying biological mechanisms for the DDIs. The key contributions of this research can be outlined as follows:

1. We propose and implement a graph analytics method, inspired from the Apriori algorithm for association rule mining, to identify frequent substructures in our systems network, as derived from mining FAERS reports.
2. We compare our method with two baseline methods in pharmacovigilance – the Gamma Poisson Shrinker (GPS) method and the Bayesian Confidence Propagation Neural Network (BCPNN) over three different validation sets.
3. We discuss briefly, the insights obtained from our method on few drug–ADR associations as well as discuss the advantages of using Semantic Web Technologies for mechanism-based pharmacovigilance.

All the results described in this paper, as well as prior research on the PhLeGrA platform, are available online at http://onto-apps.stanford.edu/phlegra/.

2 Related Work

Several methods have been developed to predict DDIs, or predict ADRs that manifest due to concomitant intake of multiple drugs, by mining spontaneous reporting systems such as FAERS or electronic medical records. Harpaz et al.7,22 used the Apriori algorithm to mine the FAERS reports and generate statistically significant association rules between multiple drugs and ADRs (e.g. Aspirin + Warfarin → Bleeding). Iyer et al.8 used electronic health records and generated patient timelines of drug and ADR mentions in the records. Using adjusted disproportionality ratios to identify significant drug–drug–event associations, and a manually-curated gold standard of such associations from Drugs.com and MediSpan, they demonstrated that their approach can be used to complement FAERS mining for pharmacovigilance. Bayesian approaches such as the Multi-Item Gamma Poisson Shrinkage (MGPS) algorithm and the Bayesian Confidence Propagation Neural Network (BCPNN)24, as well as approaches using existing knowledge on drug and ADR similarities, have recently been proposed to deal with reporting bias and confounding factors, observed in Spontaneous Reporting Systems. The performance of these methods are compared by Harpaz, et al.26 However, these methods fail to demonstrate the possible underlying molecular mechanisms behind these associations.

Systems pharmacology methods9,27 have also been explored in the context of drug–ADR association discovery or drug repurposing (use of existing drugs to treat new conditions). These methods generally combine databases and knowledge bases, to generate a systems network, manually without the use of Semantic Web technologies. CauseNet28 combines four biomedical sources into a k-partite network for generating new drug repurposing hypotheses. Berger, et al.29 integrated diseases with the human protein–protein interaction network to understand the systems pharmacology underlying specific forms of drug-induced arrhythmias. While these approaches are similar to our research, our method retrieves data and knowledge from the LSLOD cloud and can generate such systems networks more easily21.

The LSLOD cloud has been utilized to predict new DDIs recently. Tiresias processes drug-related data and knowledge and predicts new DDIs using large-scale similarity matching. Most approaches consider binary drug pairs and not multiple drug interactions, they ignore the underlying molecular mechanisms, and they may not associate the adverse drug reactions with the DDIs. Noor et al.33 constructed a mechanism-based DDI knowledge warehouse by integrating knowledge from multiple sources in the LSLOD cloud at the pharmacokinetic, pharmacodynamic, and pathway interaction level, and developed an inference engine to generate mechanistic explanations for DDIs. However, this method does not rank the mechanistic explanations, is not implemented for pharmacovigilance, and due to the knowledge warehouse, updates in the underlying sources are not captured instantaneously.
3 Materials and Methods

3.1 PhLeGrA network generation

In this section, we summarize the query federation method to extract a \( k \)-partite network from multiple, heterogeneous biomedical data sources available through the Life Sciences Linked Open Data Cloud (LSLOD). The method is described in more detail in Kamdar, et al.\(^{21}\). We integrate four different data sources that are published as SPARQL endpoints by the Bio2RDF project\(^{17}\) (Version 4) in the LSLOD cloud – DrugBank\(^{19}\), PharmGKB\(^{18}\), Kyoto Encyclopedia for Genes and Genomes (KEGG)\(^{11}\) and Comparative Toxicogenomics Database (CTD)\(^{34}\). These four sources contain data and knowledge on drugs, proteins, pathways, phenotypes and their inter-connections (e.g. drug–protein target relations) and have been used in several pharmacological methods previously.

We use a pattern-based query federation method\(^{20,21}\) to query the SPARQL endpoints of these sources simultaneously to generate the \( k \)-partite systems pharmacology network. Specifically, we retrieve four different types of entities — (E1) Drug, (E2) Protein, (E3) Pathway, and (E4) Phenotype (adverse drug reaction). We also retrieve five different types of biological relations — (R1) Drug hasTarget Protein, (R2) Drug hasEnzyme Protein, (R3) Drug hasTransporter Protein, (R4) Protein isPresentIn Pathway, and (R5) Pathway isImplicatedIn Phenotype. The SPARQL graph patterns used to retrieve the entities and relations from the sources are listed at http://onto-apps.stanford.edu/phlegra/about.

The entities and relations, retrieved from the LSLOD cloud, form a \( k \)-partite network — a network whose nodes can be partitioned into \( k \) different independent sets (\( k = 4 \)). We decided on these types of entities and relations to capture the following underlying mechanisms behind drug–drug interactions: a) one drug may inhibit the enzymes that metabolize a second drug to its inactive or active state, c) one drug may inhibit the transporters that decrease the absorption or elimination of a second drug, c) two drugs may target the same protein, leading to varying effects of both drugs, or d) two drugs may target proteins in the same pathway leading to varying effects of both drugs. Hence, here we consider transporters and enzymes to be considered as specialized proteins.

Figure 2: PhLeGrA network generation. a) PhLeGrA query federation method uses the type of entities and relations, as well as prior knowledge on SPARQL graph patterns to query four sources (DrugBank, KEGG, PharmGKB and CTD) in the LSLOD cloud to create a \( k \)-partite network composed of drugs, proteins, pathways and phenotypes. The phenotypes are further arranged using the MESH hierarchy tree. b) A visualization of a small portion of the network, with the Drugs Invega and Viagra, Enzyme CYP3A4 and Phenotype Hypertriglyceridemia highlighted.

To reconcile similar entities in different sources (e.g. drugs present in KEGG and DrugBank referenced using different identifiers), we use the \( x \)-\( ref \) attributes provided by the Bio2RDF project. We reconcile entities to a uniform identifier.
nomenclature scheme using existing terminologies – ATC (Anatomical Therapeutic Chemical Classification) for Drug, HGNC (Hugo Gene Nomenclature Committee) for Protein, KEGG for Pathway, and MESH (Medical Subject Headings) for Phenotype. We further organize the Phenotype MESH identifiers into a hierarchy, using the MESH hierarchy. As Spontaneous Reporting Systems (e.g., FAERS) collect patient reports in which the adverse reactions may be specified at different levels of abstraction, this step helps in aggregation of reports on higher abstract terms. For example, there may be patient reports on both Anaphylaxis and Hypersensitivity, Immediate, where the former Phenotype term is a subclass of the latter term. Hence the k-partite network is coupled with the MESH Phenotype hierarchy. A visualization of such a network is shown in Figure 2b.

3.2 FDA Adverse Event Reporting System

Spontaneous reporting systems are the primary means to conduct post-marketing surveillance of drug products to detect ADRs that were not determined during clinical trials. The US Food and Drug Administration (FDA) collects reports on the adverse drug reactions observed in patients subjected to multiple drugs. The FDA Adverse Event Reporting System (FAERS), a public data portal, publishes these reports after the anonymization of the patient data. We downloaded the FAERS datasets for three years from January 2013 to December 2015. Each dataset is composed of several safety reports. Among many features, each safety report indicates the set of ADRs observed in a patient (e.g., heart attack), and the set of drugs administered to the patient (e.g., Sildenafil). The steps taken to process and align the FAERS records with the Drug and Phenotype nodes in our k-partite network are described previously.

3.3 Frequent Substructure Mining

We extend the method proposed by Harpaz, et al. for statistical mining in FAERS datasets. This method is inspired from the Apriori algorithm to mine association rules (e.g., \{Drug\}_n \rightarrow ADR) in large databases, in an unsupervised, computationally tractable way. The Apriori algorithm prunes the search space of associations, such that if a certain combination of drugs and ADRs is infrequent, then any larger combination that builds upon the smaller infrequent one, will also be infrequent. Certain thresholds can be decided for ignoring these combinations. The Apriori algorithm has also been modified to mine frequent substructures in graphs. We have used this implementation of the Apriori algorithm to work on k-partite networks. Specifically, we can determine the set of FAERS reports that contain any specific (drug, ADR) pair. As shown in Figure 3a, we propagate the set of reports along all the possible shortest, directed, paths that connect the corresponding Drug node to the Phenotype node in the k-partite network. We decided to use only the shortest paths to make the method computationally tractable. Hence, each node in the k-partite network is annotated with the set of reports it may be associated with. We are unaware of the underlying biological mechanisms at this point, so all implicit associations are equally probable.

Generally, Apriori-based methods compute the Support and Confidence statistics for an association rule. Suppose,
\(S(A)\) indicates the number of reports that describe the items in itemset \(A\) (e.g. set of drugs). Then the support for an association rule is simply \(S(A \rightarrow B) = S(A \cup B)\) (i.e. number of reports where the items in itemsets \(A\) and \(B\) cooccur). The confidence of an association rule can be described as \(C(A \rightarrow B) = S(A \cup B)/S(A)\) (i.e. the conditional probability for observing items in itemset \(B\), given items in itemset \(A\)). The space of all possible association rules is pruned by selecting only those itemsets that exhibit a minimum value for the support statistic. In our method, as we have four different types of nodes, the association rules are generated by observing a minimum support at each step, where the direction of adding new itemsets is strictly Drug \(\rightarrow\) Protein \(\rightarrow\) Pathway \(\rightarrow\) Phenotype. Nodes and edges that do not exhibit a support statistic that exceeds a given threshold are automatically pruned from the \(k\)-partite network. To further reduce the number of computations, we only consider Drug \(\rightarrow\) ADR associations such that they have some direct path in the \(k\)-partite network, and we only consider \(\{\text{Drug}\}_2 \rightarrow \text{ADR}\) associations such that the paths that link the two drugs to the ADR have a common intersection point, either at the Protein or the Pathway node in the network. This method and our optimizations are visually explained in Figure 3b. It should be noted that this method can be extended to include multi-drug interactions (\(\{\text{Drug}\}_n \rightarrow \text{ADR}\)).

The confidence statistic is computed as \(C(\text{Drug} \cup \text{Protein} \cup \text{Pathway} \rightarrow \text{Phenotype})\), and is used to rank the different substructures (i.e. different underlying mechanisms), that lead to the manifestation of the ADR given the set of drugs. However, due to the reporting bias in FAERS, this statistic in itself is not sufficient to actually determine if there is any association between the drugs and the ADR, as indicated by Harpaz, et al.\(^7\) Hence, we also compute a Network-based Relative Reporting Ratio (RRR) statistic, that considers the Actual/Expected ratio at each path in the \(k\)-partite network. Hence, each possible substructure has an RRR statistic. RRR is defined as the ratio between an association rule’s observed frequency to a baseline expected frequency under the assumption of independence.

\[
\text{RRR} = \frac{N \times S(A \cup B)}{S(A) \times S(B)}
\]

Here \(A = \text{Drug} \cup \text{Protein} \cup \text{Pathway}\), \(N\) is the total number of FAERS records, and \(B = \text{Phenotype}\). The median value of the RRR, computed for relevant substructures, is used as the statistic to compare our method against other baseline methods. To summarize, Support statistic is used to prune the \(k\)-partite network, Network-based Relative Reporting Ratio is used to determine whether an association between a set of drugs and an ADR exists, and Confidence statistic is used to rank the different underlying mechanisms behind \(\{\text{Drug}\}_n \rightarrow \text{ADR}\) association.

### 3.4 Method Evaluation

We collected three different datasets that consist of manually-curated positive and negative drug–adverse reaction associations. These datasets have been used to validate DDI prediction methods previously. The Observational Medical Outcomes Partnership (OMOP\(^3\)) dataset and the European “Exploring and Understanding Adverse Drug Reactions” project (EU-ADR\(^4\)) dataset consists of single drug–ADR associations. The dataset described in Iyer, et al.\(^8\) consists of drug–drug–ADR associations retrieved from Drugs.com and MediSpan.

**Table 1:** The coverage of different validation datasets used in this study.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Unique Drugs</th>
<th>Unique ADRs</th>
<th>Positive associations</th>
<th>Negative associations</th>
</tr>
</thead>
<tbody>
<tr>
<td>OMOP</td>
<td>155</td>
<td>4</td>
<td>137</td>
<td>158</td>
</tr>
<tr>
<td>EU-ADR</td>
<td>59</td>
<td>9</td>
<td>44</td>
<td>39</td>
</tr>
<tr>
<td>Iyer et al.(^8)</td>
<td>252</td>
<td>9</td>
<td>315</td>
<td>288</td>
</tr>
</tbody>
</table>

Some statistics for these datasets, in terms of positive and negative associations, as well as coverage of drugs and ADRs are shown in Table 1. All the three validation sets were transformed, such that the drugs were referenced using ATC identifiers and ADRs were referenced using MESH identifiers. Some common ADRs across the three datasets include – Gastrointestinal Hemorrhage, Hyperkalemia, Acute Kidney Injury and Drug-induced Liver Injury. Using these validation sets, we compare our method with two baseline methods — the Gamma Poisson Shrinkage (GPS) method and the Bayesian Confidence Propagation Neural Network (BCPNN) method. We used the R package for PharmacoVigilance Signal Detection (PhViD\(^1\)) for the baseline methods.

\(^1\)https://cran.r-project.org/web/packages/PhViD/PhViD.pdf
4 Results

The $k$-partite network that was generated from the PhLeGrA query federation method, consisted of 2,759 drugs (E1), 3,890 phenotypes (E4) organized using the MESH hierarchy, 19,903 genes (E2) and 301 pathways (E3). The network also consisted of 249,001 drug–target relations (R1), 2,062 drug–enzyme relations (R2), 919 drug–transporter relations (R3), 25,480 protein–pathway relations (R4) and 46,300 pathway–phenotype relations (R3). Individual statistics for the different entities and relations extracted from each source were presented previously by Kamdar, et al.21. We used ≈ 3 million FAERS reports for the frequent substructure mining method demonstrated in this research.

The FAERS reports were propagated along our $k$-partite network, with each node in the network annotated with the set of nodes that it may be associated with. We use a Support threshold of 200 to filter out nodes and edges in the $k$-partite network. After applying the support threshold, we were able to decrease our $k$-partite network to include only 7,217 nodes and 89,451 edges. Moreover, as seen from the Figure 4, the number of entities and relations of a specific type (e.g. Drug hasTarget Protein) is reduced drastically for a particular source (e.g. CTD). It can be argued that our method can remove spurious relations in the $k$-partite network that may not be relevant, or may be incorrect.

We compute the Network-based RRR statistic for a given association between a set of drugs and an ADR, given all possible substructures. We compare this statistic with the GPS and BCPNN statistic over the three validation datasets. The Receiver-Operator Characteristic curve for each validation dataset is shown in Figure 5. It can be seen that the area under the curve (AUROC) statistic for each of the three validation sets is almost similar to the baseline methods, and the value is around 0.7–0.8. The AUROC statistic actually exceeds by 0.01–0.02 over the baseline methods for the OMOP and the EU-ADR validation sets. Moreover, as observed by Iyer et al.8, using event-specific thresholds on the statistic can actually generate higher AUROCs for certain ADRs. This is observed in Figure 5d where we obtain an AUROC of almost 0.94 for rhabdomyolysis. Finally, we would like to note that for higher values of Specificity, our sensitivity is considerably less than existing baseline methods. This may be due to the reporting bias in FAERS, which is not tackled in this research. However, using GPS-adjusted Expected values in the Network-based RRR statistic actually alleviated this issue (plot not shown). On a closer inspection of the false positive associations, we observed that these associations have more connecting paths in the $k$-partite network when compared to the true negative associations. Similarly, the false negative associations have fewer connecting paths in the $k$-partite network when compared to the true positive associations. Using a $t$-test, we found that these comparison findings were statistically significant ($p < 0.05$). Hence, the topology of the network has an impact on the association discovery method.
Figure 5: The first three plots include the ROC (Receiver-Operator Characteristic) curves for the three validation datasets used in our study – a) OMOP, b) EU-ADR and c) Drugs.com-MediSpan corpus curated by Iyer, et al. The predictive power of our Network-based RRR method is represented using the solid yellow line, whereas the dotted magenta and the dotted blue line indicate the predictive power of the GPS and BCPNN methods respectively. In d), we demonstrate ADR-specific predictions using the Network-based RRR method.

5 Discussion

Using the confidence statistic that is also computed by our method, we were able to observe some interesting and some known substructures. For example, while it is known that simvastatin may interact with the CYP3A4 inhibitor itraconazole to cause rhabdomyolysis, the corresponding substructure was observed to have a high confidence value in our analysis. Moreover, we found that the drug paliperidone, which is another CYP3A4 inhibitor and is used as a treatment in bipolar disorder, may interact with several other drugs to cause hypertriglyceridemia, hyperprolactinemia and gynecomastia. However, these findings need to be validated by a domain expert in the future. We plan to incorporate this method in the PhLeGrA visualization browser², such that different substructures can be highlighted and ranked with the support and confidence statistics and can provide a better understanding to the domain user.

Currently, our method is limited to substructure discovery from only those parts in the networks where there exist some edges that connect the different entity types (e.g. two drugs may interact with the same proteins, but this knowledge is derived from an existing source). This assumption was used to allow our Apriori algorithm to terminate under a reasonable runtime. However, there may exist drug–protein relations that are not known currently, or may not be stored.

²http://onto-apps.stanford.edu/phlegra/
in the biomedical sources that were integrated to generate the $k$-partite network. We will argue that the benefits of the Semantic Web technologies and Linked Data principles allows us to incorporate multiple data sources in the $k$-partite network, whereas the initial thresholding using the Support statistic can allow us to filter irrelevant edges and nodes in the network. This was observed for the Comparative Toxicogenomics Database that incorporated a huge proportion of noisy edges between drugs and protein targets that were not actual relations.\textsuperscript{21}

As presented here, pattern-based query federation\textsuperscript{21} can bring together pharmacological knowledge existing in isolated, heterogeneous sources without being concerned about the underlying semantics and schema differences. This advantage, when coupled with the network-based Apriori method for association rule mining, can facilitate domain users to obtain mechanistic explanations behind detected DDIs and ADRs, as well as generate new knowledge on underlying biological mechanisms (frequently observed substructures). Such systems pharmacology networks, as previously described\textsuperscript{21} are easier to generate using Semantic Web technologies and query federation methods.

6 Conclusion

In this research, we have demonstrated a method for mechanism-based pharmacovigilance from Spontaneous Reporting Systems, such as the FAERS datasets. While our method has equivalent, if not better, performance compared to existing state-of-the-art methods in pharmacovigilance, it can also be used to provide a mechanistic understanding behind the drug–drug interactions and the adverse reactions that manifest on the account of those DDIs. Moreover, it can enable biomedical researchers to obtain newer knowledge on molecular mechanisms that may be relevant in pharmacovigilance, or may be spurious in a particular database. We use Semantic Web technologies to easily generate a systems pharmacology network, and to the best of our knowledge this is the first approach to provide explanations of underlying biological mechanisms using a ranking scheme.

Acknowledgments

The authors acknowledge Rainer Winnenburg, Erik van Mulligen and Juan Banda for providing the OMOP, EU-ADR and Drugs.com-MediSpan datasets respectively. The authors also acknowledge Michel Dumontier for his help using Bio2RDF linked data. This work is supported by Grant HG004028 from the US National Institutes of Health.

References

Initializing and Growing a Database of Health Information Technology (HIT) Events by Using TF-IDF and Biterm Topic Modeling

Hong Kang, Zhiguo Yu, Yang Gong
School of Biomedical Informatics, the University of Texas Health Science Center at Houston, Houston, TX, USA

Abstract

Health information technology (HIT) events were listed in the top 10 technology-related hazards since one in six patient safety events (PSE) is related to HIT. Although it becomes a common sense that event reporting is an effective way to accumulate typical cases for learning, the lack of HIT event databases remains a challenge. Aiming to retrieve HIT events from millions of event reports related to medical devices in FDA Manufacturer and User Facility Device Experience (MAUDE) database, we proposed a novel identification strategy composed of a structured data-based filter and an unstructured data-based classifier using both TF-IDF and biterm topic. A dataset with 97% HIT events was retrieved from the raw database of 2015 FDA MAUDE, which contains approximately 0.4~0.9% HIT events. This strategy holds promise of initializing and growing an HIT database to meet the challenges of collecting, analyzing, sharing, and learning from HIT events at an aggregated level.

Introduction

Patient safety event (PSE), defined as any process, act of omission, or commission that results in hazardous healthcare conditions and/or unintended harm to the patient 1, is the third leading cause of death in the United States2-4. PSEs are complex and difficult to control because they are related to healthcare systems, operations, drug administration, or any clinical aspect of patient care 5. Due to the wide application of health information technology (HIT) in clinical settings, HIT events become a key component of PSEs. HIT includes hardware or software that is used to electronically create, maintain, analyze, store, receive (information), or otherwise aid in the diagnosis, cure, mitigation, treatment, or prevention of disease, and that is not an integral part of an implantable device or medical equipment 6. The positive impacts of HIT 7, 8 include cost savings and improved patient outcomes 9, 10, decreased occurrence of medication errors 11, and improved healthcare process measures across diverse settings 12. However, if HIT is poorly designed or implemented, it poses a risk to patient safety 13-15. For instance, an anesthesiologist did not know his patient had taken oxycodone because of the lack of interoperability between the office-based medical record platform and the inpatient system; the patient became somnolent upon receiving morphine after the oxycodone 16. HIT events were listed in the top 10 technology-related hazards identified by the Emergency Care Research Institute among a range of common problems in 2015 17. One in six PSEs is related to HIT, such as medical devices, EHRs, and CPOE 18. Therefore, HIT related events pose a major threat and barrier toward a safer healthcare system, and the large number of HIT related events must be addressed to reduce patient harm.

Safety event reporting has been proven effective in many high-risk industries 19-21, for improving safety and enhancing organizational learning from errors. Healthcare systems have adopted event reporting since the 1999 Institute of Medicine (IOM) report 22 that greatly raised the public awareness of patient safety issues. Through collecting reports of adverse events and near misses in healthcare, reporting systems would enable safety specialists to analyze events, identify underlying factors, and generate actionable knowledge to mitigate risks 23. In the U.S., the IOM recommended using patient safety reporting systems 24, 25 to determine why patients are sometimes harmed during medical care. The Agency for Healthcare Research and Quality (AHRQ) created the Common Formats (CF), common definitions and reporting formats 21 to help healthcare providers uniformly report PSEs.

However, collecting data on HIT-related PSEs for learning purposes is challenging owing to the lack of HIT reporting forms or platforms. Although CF forms are commonly used in Patient Safety Organizations (PSO), the CF form containing HIT event categories is defined at a very high level and embedded with the category of device or medical/surgical supply. This makes it difficult for reporters to recognize the proper categories when reporting, and as a result, reporters often leave fields blank rather than responding to the prepopulated questions. In the 2015 annual report of a PSO institute, the Missouri Center for Patient Safety, only one PSE was identified in the original report as an HIT event. Therefore, the scarcity of HIT event-exclusive databases and event reporting systems indicates the challenge of identifying the HIT events from existing resources.
EHR seems to be a potential resource to extract HIT events because EHRs are real-time, patient-centered records that make information available instantly and securely to authorized users. While an EHR does contain the medical and treatment histories of patients, an EHR system is built to go beyond standard clinical data collected in a provider’s office and can be inclusive of a broader view of a patient’s care26. However, EHR itself belongs to HIT, which may hide a large amount of EHR related events that happened during data entry and data transfer. In addition, there is no structured field that could help extract HIT events in an EHR that is not designed for such a purpose.

The U.S. FDA Manufacturer and User Facility Device Experience (MAUDE) database27 is a rich and publicly accessible resource with the potential of extracting HIT events. Different from EHRs, the FDA MAUDE database focuses on the reports of events involving medical devices, voluntary reports of medical device malfunction, and reports of problems leading to serious injury and death since June 199328. The database houses medical device reporting submitted to the FDA by mandatory reporters (manufacturers, importers and device user facilities) and voluntary reporters (healthcare professionals, patients and consumers). The FDA MAUDE is updated weekly and searchable online. As of February 2017, MAUDE had more than 6 million reports. The challenge of identifying HIT events from the FDA MAUDE is that only 0.1% of reports are related to HIT, and are mixed with other reports regarding equipment failures and hazards29. Due to the enormous size of the FDA MAUDE and the small percentage of HIT events, directly identifying and extracting all HIT events from the FDA MAUDE is almost impossible using a straightforward strategy. The classification of data with imbalanced class distribution has encountered a significant drawback of the performance attainable by most standard classifier learning algorithms30. In addition, the rapid evolvement of topics and free text regarding HIT events presents another challenge for distinguishing these events from others. The initialization of HIT event database based on the FDA MAUDE database requires an up-to-date, efficient, and effective strategy.

A MAUDE report is composed of structured fields (device and patient information) and unstructured fields (textual information). In the preliminary study, we developed an HIT filter based on the generic name and manufacturer name, two structured fields of the 2015 MAUDE database, enabling us to screen the FDA MAUDE and consequently create a report subset with more than 50% HIT events31. We sampled and reviewed the filtered reports which resulted in an estimate that 0.4~0.9% FDA MAUDE reports were HIT events. Although this ratio has been significantly enhanced, it is still far away from establishing an HIT dataset. Fortunately, a classifier based on unstructured field may be feasible to further identify HIT events, since the information from unstructured data provides an integrated and comprehensive view. In addition, the ratio of the HIT events in filtered data is perfect for classifier training and evaluation.

In this study, six popular classification algorithms were compared on 2015 MAUDE database by using term frequency-inverse document frequency (TF-IDF)32 as the feature of each single word in the unstructured fields. Then bitem model (BTM)33, a word co-occurrence based topic model that learns information from word combinations, was applied to further improve the best classifier of the six. After discussing the balance of precision and recall, according to the learning requirement, a final model that can provide a dataset with more than 97% HIT events was proposed. This model offers the probability of initializing and growing an HIT database from the FDA MAUDE and other potential resources. The outcome will be a timely reflection of the evolvement of HIT events and will be helpful for enriching HIT knowledge and better using the historic reports toward an overall understanding and analysis of the characteristics, occurrence, observation, and description of HIT events. In addition, this project provides an intelligent strategy to connect structured data with unstructured data and holds promise in triggering a revolution of data management and analysis in healthcare and other industries.

**Methods**

**Improve the filter for the structured fields of FDA MAUDE**

Most of the 45 structured fields of the FDA MAUDE database are either left blank by reporters or are of little use for the purpose of identifying HIT related events. The only two fields have the greatest potential in identifying HIT related events are **generic name** and **manufacturer name**. Therefore, a filter based on these two fields was established in our preliminary study31.

We started with a set of common computer hardware and software related keywords that had been previously identified34. The starting keyword list was expanded by the addition of several generic terms such as “software,” “program,” and “hardware” and several more modern terms such as “electronic medical record” and “portal technology.” Then all of the generic names from the FDA MAUDE database starting from Jan 2010 to Dec 2015 were extracted, yielding a total of 60,000 unique generic names. Next, through partial keyword matching of the generic names to the keywords list, a subset of generic names appearing HIT-related was extracted. The subset was further
analyzed by domain experts to determine which generic names were actually linked to HIT reports by using a small portion of the 2015 FDA MAUDE database. A similar approach was utilized in selecting a list of manufacturers for the filter. We started with a list of seven popular HIT manufacturers\(^{34}\) and then added 347 manufacturers of HIT software. The reports from the 2015 FDA MAUDE database related to those manufacturers were extracted and checked to determine the most relevant manufacturers.

To evaluate and improve the HIT filter, the filter was first applied to the 2015 FDA MAUDE database. Then a subset (10\%) of the reports screened by the filter was manually reviewed by two domain experts. The experts labeled each report with one of three labels: HIT, Not HIT, or Unsure. The reports with disagreements among the reviewers were resolved through group discussion. During the review, our team narrowed the HIT definition to identify the most clinically relevant and consequential HIT devices. Under our current understanding, an HIT device is any device that utilizes both hardware and software to facilitate health information exchange in order to aid in the diagnosis, treatment, or prevention of disease. Using this definition, priority is given to HIT systems that focus on information exchange such as electronic health records, computerized physician order entry, and picture archiving communications systems. Implantable devices, glucose monitors, defibrillators, and similar devices are excluded under this definition, as they do not actively facilitate health information exchange.

**Compare six classification algorithms based on the unstructured fields of FDA MAUDE**

Six popular machine learning classifiers including logistic regression, support vector machine (SVM), naïve Bayes, decision tree, JRip rules, and random forest were constructed using the unstructured data of the reviewed reports. Weka 3.8.1\(^{35}\) was applied for training and validating the classifiers. The unstructured data of 2015 FDA MAUDE was filtered by removing the words in the Rainbow stoplist\(^{36}\). Then each report in the labeled training set was treated as a vector of words and was weighted by a TF-IDF technique. The effectiveness of this word-based information retrieval approach has been proven effective in searches of patient records, which are often corrupted by misspelled words and conventional graphs or abbreviations\(^{37,38}\). Eq.1 shows how to calculate the weight of term \(x\) within document \(y\) using TF-IDF (\(tf_{x,y}\) is the frequency of term \(x\) in document \(y\), \(df_x\) is the number of documents containing term \(x\), \(N\) is the total number of documents). The six classifiers were evaluated using leave-one-out cross-validation (LOOCV) and their performances were weighted based on both their F-scores and ROC areas.

\[
w_{x,y} = tf_{x,y} \times \log \left( \frac{N}{df_x} \right)
\]

(Eq.1)

**Improve HIT classifier by using BTM**

TF-IDF only considers the information from a single word and its distribution across the whole corpus, but cannot capture the semantics information among the dataset. Topic models\(^{39}\) are a class of statistical machine learning algorithms that extract the semantic themes (topics) from a corpus of documents. These topics generated based on the document-level word co-occurrence patterns describe the thematic composition of each document and can thus capture the semantic similarity between topics and documents. Moreover, those generated topics have been widely used as features to improve document classification and information retrieval performance\(^{40,41}\).

The BTM\(^{33}\) that we used in this work to improve the HIT classifier is specifically designed for short text documents. In general, an HIT report is too short to provide enough word counts for conventional topic models, like PLSA\(^{42}\) and LDA\(^{43}\), to know how words are related, comparing to lengthy documents. Also, the limited contexts make it very difficult for topic models to identify the sense of ambiguous words in short documents. BTM overcomes those short text limitations by assuming that all the biterms (i.e. unordered co-occurring word pairs) are generated by a corpus level topic distribution to benefit from the global word co-occurring patterns. Informally, the “generative story” for BTM is as follows. Firstly, each topic’s word distribution is drawn over the vocabulary of the whole corpus. Then a corpus-level topic distribution is drawn to describe this dataset. To generate each biterm in the biterm set extracted from the whole corpus, one draws a topic from the corpus-level topic distribution and subsequently selects two words from this distribution across all vocabulary of the whole corpus corresponding to this topic. BTM uses this generative process to model the co-occurrence of a pair of words over the whole corpus to reveal the topics than just the occurrence of a single word, and then enhance the learning of topics. Specifically, it estimates the parameters that define the topic mixture over the whole corpus and the conditional probability of each word given each topic. The topic distribution of each document can then be naturally derived based on the learned model. Parameter estimation is done via Gibbs sampling approach.

The number of topics generated by BTM must be pre-specified. Determining the “right” number of topics for different data sets remains a challenge. When the number of topics increases, redundant and nonsense topics may be generated.
In this study, we conducted experiments with various numbers of topics ranging from 5 to 100. The performance in the classification tasks was quite consistent after a particular topic number threshold. Hence, it is likely that a BTM with a relatively large number of topics will capture all the key themes over this corpus and in further to improve the document classification task.

Results

HIT event filter based on structured data

Based on the analysis of pre-viewed FDA MAUDE reports and the discussion with two healthcare professionals, 58 keywords from generic names (39 software and 19 hardware keywords respectively) and 16 keywords from manufacturer names were determined to compose the filter for HIT related reports, as shown in Table 1. The filter was first applied to the 2015 FDA MAUDE database including 860,915 reports. 4871 reports (2479 software and 2392 hardware reports) were initially found. 490 reports (10%) were randomly selected according to the keyword distribution for expert review and labeling. 289 reports were identified as HIT related by experts, which means the filter can generate a report subset from original 2015 MAUDE database with about 50-60% HIT related reports. This proportion is much more practicable in terms of classifier training than that on the entire MAUDE database (0.4~0.9%)²⁹.

Table 1. Keywords of the HIT filter in alphabetical order

<table>
<thead>
<tr>
<th>Generic Names</th>
<th>Manufacturer Names</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Hardware related:</strong></td>
<td></td>
</tr>
<tr>
<td>Anesthesia Monitor,</td>
<td>Telemetry Monitor,</td>
</tr>
<tr>
<td>Apnea Monitor,</td>
<td>Telemetry Transmitter,</td>
</tr>
<tr>
<td>Arterial Monitor,</td>
<td>Vital Sign Monitor,</td>
</tr>
<tr>
<td>Atlas Monitor,</td>
<td>Workstation</td>
</tr>
<tr>
<td>Blood Pressure Monitor,</td>
<td></td>
</tr>
<tr>
<td>Central Monitor,</td>
<td></td>
</tr>
<tr>
<td>Computer,</td>
<td></td>
</tr>
<tr>
<td>Console Monitor,</td>
<td></td>
</tr>
<tr>
<td>Drug Screen,</td>
<td></td>
</tr>
<tr>
<td>Fetal Monitor,</td>
<td></td>
</tr>
<tr>
<td>Patient Monitor,</td>
<td></td>
</tr>
<tr>
<td>Physiological Monitor,</td>
<td></td>
</tr>
<tr>
<td>Pressure Monitor,</td>
<td></td>
</tr>
<tr>
<td>PT Monitor,</td>
<td></td>
</tr>
<tr>
<td>Safety Monitor,</td>
<td></td>
</tr>
<tr>
<td><strong>Software related:</strong></td>
<td></td>
</tr>
<tr>
<td>ADC,</td>
<td>Management System,</td>
</tr>
<tr>
<td>Alert,</td>
<td>Monitoring System,</td>
</tr>
<tr>
<td>Automated Dispensing Cabinet,</td>
<td>Network,</td>
</tr>
<tr>
<td>Communication Device,</td>
<td>Order Entry,</td>
</tr>
<tr>
<td>Communication System,</td>
<td>PACS,</td>
</tr>
<tr>
<td>CPOE,</td>
<td>Electronic Health,</td>
</tr>
<tr>
<td>Data Backup,</td>
<td>Picture Archiving,</td>
</tr>
<tr>
<td>Database,</td>
<td>Portal,</td>
</tr>
<tr>
<td>Decision Support,</td>
<td>Powerchart,</td>
</tr>
<tr>
<td>Digital,</td>
<td>Program,</td>
</tr>
<tr>
<td>Dispensing System,</td>
<td>Soarian,</td>
</tr>
<tr>
<td>Dose Suggestion,</td>
<td>Telemetry,</td>
</tr>
<tr>
<td>Downloader,</td>
<td>Trima Accel Platelet,</td>
</tr>
<tr>
<td>Drug Suggestion,</td>
<td>Web</td>
</tr>
<tr>
<td>EHR,</td>
<td></td>
</tr>
<tr>
<td>Electronic Medical,</td>
<td></td>
</tr>
<tr>
<td>Electronic Patient,</td>
<td></td>
</tr>
<tr>
<td>EMR,</td>
<td></td>
</tr>
<tr>
<td>ICT,</td>
<td></td>
</tr>
<tr>
<td>Imaging System,</td>
<td></td>
</tr>
<tr>
<td>Information System,</td>
<td></td>
</tr>
<tr>
<td>Internet,</td>
<td></td>
</tr>
<tr>
<td>Invision,</td>
<td></td>
</tr>
<tr>
<td>LIS,</td>
<td></td>
</tr>
</tbody>
</table>

HIT Classifiers using TF-IDF model

The manually labeled reports (289 HIT and 376 non-HIT reports) from 2015 MAUDE database were applied as the training set of the classifiers. 1,541 words extracted from the unstructured fields were fixed as the feature set after removing stopwords. TF-IDF was calculated for each word in each report. A comparison was made among the classifiers trained using six popular classification algorithms: logistic regression, random forest, naïve Bayes, SVM, decision tree (J48), and JRip rules. As shown in Table 2 and Figure 1, random forest has the best performance in terms of accuracy, ROC area, and F-score.

Table 2. Performances of the six HIT classifiers

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
<th>F-score</th>
<th>ROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random Forest</td>
<td>0.809</td>
<td>0.807</td>
<td>0.900</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>0.802</td>
<td>0.801</td>
<td>0.871</td>
</tr>
<tr>
<td>Naïve Bayes</td>
<td>0.786</td>
<td>0.786</td>
<td>0.853</td>
</tr>
<tr>
<td>SVM</td>
<td>0.792</td>
<td>0.792</td>
<td>0.787</td>
</tr>
<tr>
<td>Decision Tree (J48)</td>
<td>0.737</td>
<td>0.737</td>
<td>0.718</td>
</tr>
<tr>
<td>JRip Rules</td>
<td>0.716</td>
<td>0.715</td>
<td>0.716</td>
</tr>
</tbody>
</table>
HIT Classifiers using BTM

BTM was implemented to analyze the words of the unstructured fields of the training set and to discover the topics without any prior annotations or labeling of the reports. This statistical model reflects the intuition that narrative reports exhibit multiple topics. Each report exhibits the topics in different proportions; each word in a report is drawn from one of the topics, where the selected topic is chosen from the per-document distribution over topics. All the reports in the collection share the same set of topics, but each report exhibits those topics in different proportions. We applied the topics as the feature set and the weights of the topics as the features’ values. To figure out the best topic number, BTM was run for 20 times by setting topic numbers from 5 to 100 with an interval 5. Random forest classifiers were also trained for 20 times as the topic number changed. As shown in Figure 2, most BTM-based classifiers have higher F-scores than the TF-IDF-based classifier, while the ROC areas are not improved by BTM. The BTM-based classifiers with the best F-score and ROC were obtained when setting the topic number at 70 and 80. Therefore, we combined the features of BTM and TF-IDF at each of the two topic numbers and trained two additional classifiers respectively (the dots in Figure 2). The classifier combining TF-IDF features with 70 topics from BTM has the best performance with ROC area 0.920 and F-score 0.834, which are significant improvements comparing those of the original TF-IDF-based classifier.

A dataset with 97% HIT events

The original purpose of building the HIT filter and classifiers was not to identify all HIT related reports from the FDA MAUDE. Instead, we need an HIT exclusive database or a report dataset with a large proportion of HIT related events. In other word, we can tolerate the loss of recall to obtain a higher precision. Random forests build an ensemble of tree classification predictors using bagging. Each node of the trees only considers a small subset of features for the split. The classification is done by voting, which means a probability is calculated based on the voting result of all predictors.
Figure 3 shows the changes of true positive and false positive as the threshold of the probability changes. Based on the case review, we found that one third of the HIT related events are typical enough for learning purposes. Therefore, if setting the threshold at 0.7, we could get a high precision 0.97 and adequate true positive cases. By using this strategy, a report dataset with 97% HIT related events was derived from the original 2015 FDA MAUDE, which contains up to 0.9% HIT related events.

**Figure 3.** The effects of random forest probability

**Discussion**

**Initialize and grow an HIT database**

Learning directly from the AHRQ CF and FDA MAUDE is challenging since the two reporting forms are not HIT exclusive. This study provides an effective way to extract HIT related events from the FDA MAUDE and initialize an HIT event database. To establish and grow the database, meanwhile, further improve the HIT filter and classifier, we will apply the HIT event filter and classifier on the 2014 dataset. Only the cases that are identified as HIT will be manually reviewed. All the reviewed cases will be added to the training set, which will consequently help build a new classifier. The same process will, in turn, be retrospectively applied on the datasets from 2013, 2012, 2011, 2010, and beyond. At the end of each iteration, the four training methods will be re-evaluated based on the corresponding manual review, and the best method will be used in the classifier. The classifier is expected to be improved as more labeled cases are included. As the manually reviewed HIT events are accumulated, an HIT event database will be established and keep growing.

In our preliminary study, we prototyped a user-centered PSE reporting system based on a PSE knowledge base, which includes PSE reports, solutions, and their connections. Patient falls, another important PSE subtype, were applied to test the system. Users can either choose an existing fall case or report a new case, then the system retrieves similar cases and customized solutions based on the query and the reporter’s role (e.g., manager, clinician, staff, patient). The user preference may be diverse for different purposes. The system also allows the user to click the feedback button to indicate their preferences to a certain similar case or solution. All feedback will be returned to the algorithm implementation step in order to update the weights of similarity matrices and dynamically upgrade the system performance. This mechanism, similar to the ranking strategy of the Google search engine, will gradually stabilize the similarity matrices, making them more convincing as the feedback increases. The HIT identification strategy together with the TF-IDF and BTM features will be incorporated in this system to grow the database.

**Explore contributing factors and connections of HIT events for shared learning**

Intuitively, finding a way to compare two relevant HIT cases will be beneficial for learning from previous cases. The FDA MAUDE database is rich, broad and unique because it is a collection of errors of many HIT products from prevailing vendors such as Epic, Cerner, GE Healthcare, Allscripts, and McKesson; the connections among the reported errors will be helpful for their users and developers alike. The benefit once perceived would attract more HIT vendors and users to join the reporting and promote shared learning. The BTM can map the HIT events to the topic space that each event has an HIT topic rank according to the relationship between this event and each topic. We can observe how those topics changed over time and how they are connected to each other. Rather than finding reports
through keyword searches alone, we can find a small group of central topics as contributing factors and then examine
the reports related to the topics. More importantly, the topics can help compare and measure the vector-based similarity
between HIT reports, which could provide more targeted knowledge support to the reported cases in comparison with
similar or relevant cases. We can also extract the common characteristics of the HIT cases within a certain similarity
range, which will help us track the changes of HIT events and further improve the HIT identification strategy and
benefit healthcare professionals for shared learning.

Reduce the human labor

Manually reviewing all cases in the FDA MAUDE database was simply infeasible and that machine learning was
likely to be the only viable approach. Traditionally, the two paradigms of machine learning have been supervised (all
labeled data) and unsupervised learning (all unlabeled data). However, recently, much attention has been placed on
semi-supervised learning for its ability to utilize only a small amount of labeled cases combined with a large amount
of unlabeled cases to improve classification accuracy. In the case of HIT, this approach seemed to be well-suited as
the cost of labeling narrative text by manual review is quite high, while the cost of obtaining unlabeled reports is
minuscule in comparison. One of the simplest methods within semi-supervised learning is self-training. In self-
training, predictions made with high confidence by the classifier are added back to the labeled data, so that the
classifier keeps being updated and improved after each iteration. This approach seems to a viable method to extract
and classify HIT reports from large databases such as the FDA MAUDE.

Connecting structured and unstructured data is essential for patient safety data retrieval

The variety of reporting formats is the primary challenge to improving quality and connection of PSE reports. Our
preliminary work suggests that communication among PSEs can be established if we extract all the necessary
information and properly annotate it to the same hierarchical feature structure. However, when the quality of the
structured features is unsatisfactory (e.g., HIT related reports in AHRQ CF), we need to find more information in the
unstructured data (narrative reports) to ensure a comprehensive view. An analogy can be made with the U.S. National
Library of Medicine’s PubMed literature filtering, in which a controlled vocabulary, Medical Subject Headings (MeSH) was
developed to index MEDLINE articles. MeSH has been successfully used to improve PubMed query results, information retrieval, document clustering, and query refinement in “downstream” applications that use
PubMed abstracts. Similarly, a MeSH-like patient safety vocabulary holds promise in connecting structured and
unstructured features and in becoming the framework for the PSE knowledge base, the collection of PSE reports,
solutions, and the potential connections among them. The TF-IDF and BTM features in this study have the potential
to explore the “MeSH” in patient safety domain and standardize PSE reporting and management.

Benefits of shared learning and reporting systems

While all technology is fallible, technology-induced errors in healthcare may have far more serious repercussions than
those in other fields. Even with extensive testing, HIT may be especially vulnerable to failure due to the fast pace and
complexity of the healthcare system. As an example, many EHR systems perform well in testing when used by only
one user of each user category in an office setting. However, when released to the public, the EHR systems may
malfunction as a group of physicians from all areas of medicine may use them in a variety of settings. Manufacturers
of technology often do not have the resources to test their products on such a large scale and must instead rely on user
generated reports to inform them of possible defects with their products. Perhaps, most well-known, Microsoft’s error
reporting system (codenamed Watson) has allowed the company to collect and learn from billions of error reports all
across the world. Integrated into Windows, Watson automatically collects information on program crashes and sends
it to Microsoft. Microsoft analyzes each crash report, and if available, sends a solution back to the user. Watson has
played an integral role in helping Microsoft identify errors and has also helped the company prioritize its debugging
efforts.

Implementation of a similar system for HIT events could yield tremendous benefits. With the knowledge gained from
a shared learning system, patient safety experts could analyze the distribution of HIT events and identify common
underlying factors among the reports. Safety experts could then prioritize their efforts to generate actionable solutions
for the most important and pressing patient safety events. While healthcare systems have utilized event reporting in
the last decade to increase awareness of patient safety issues, several barriers still exist that prevent patient safety
reporting systems from reaching their full potential. These include the fact that many reporting systems are still paper
based, which prevents data sharing; many of the user reports may be hard to understand due to ambiguous and vague
language used; and much of the patient safety data is scattered across many reporting systems, which are challenging
to aggregate.

1030
Barriers of voluntary reporting systems

The proposed HIT database was built on the voluntary reports generated by clinicians, patients, or consumers. Such reports are indispensable for preventing future occurrences of these adverse events and help foster a healthcare culture driven toward a safer healthcare system. Unfortunately, according to the US Department of Health and Human Services, 86% of medical errors go unreported. Many barriers still exist that prevent safety reporting systems from reaching their full potential. Perhaps a primary reason safety events go unreported, regardless of the field they occur in, is that people feel their reports will not make a difference. The origins of their feelings are completely understandable. In a well-publicized incident, Toyota was fined 1.2 billion dollars for negligence in addressing known safety issues that caused unintended acceleration in their vehicles. Prior safety reports were not enough to cause the company to fix the safety issue, resulting in a series of tragedies. In one tragedy, Officer Mark Saylor was driving his car with his family when the accelerator became stuck. Unable to free the accelerator, Mr. Saylor called 911 to narrate his terrifying ordeal. However, it was too late and Mr. Saylor and his family died after crashing into an SUV.

In a high stakes field such as healthcare, it is critical that errors are reported and that manufacturers are held accountable for their products. In regards to HIT, the challenge to get healthcare providers to report these events may be especially high. With the massive spending and efforts taken to adopt HIT by hospitals and the government, there may be a positive bias toward reporting the benefits of HIT and a negative bias toward reporting adverse events caused by HIT. To make matters worse, even when safety events are reported, manufacturers may not have enough resources allocated toward diagnosing and fixing those issues. While the staggering pace of technology has driven much innovation in healthcare, the need to take a careful look at the adverse events caused by HIT has never been greater. Doing so may ultimately keep HIT on the right track toward becoming a safe and integral part of the healthcare system.

Limitations

There is no exact figure about the proportion of HIT related reports in the FDA MAUDE, because the proportion was only an estimation based on limited sampling and reviewing. After applying the HIT filter on the 2015 FDA MAUDE, a subset (10%) of the reports of the screened data was selected for manual review to save human labor. Thus, the ratio 50-60% representing the HIT related reports among the filtered data was estimated based on the 10% random chosen reports. There might be a bias when using the subset’s profile to predict the profile of the whole dataset. An effective approach to reduce the bias could be utilizing a semi-supervised learning to grow the HIT database. The classifier can be trained using an iterative approach, and cases that are labeled by the classifier with a high probability (e.g., >0.90) of being HIT related are added directly to the training set. Using this self-training semi-supervised learning approach, it is possible to introduce biases and for early mistakes to become compounded through subsequent rounds of training. Despite several rounds of error analysis can be performed to prevent such an outcome, selection bias toward certain HIT events cannot be completely eliminated.

Conclusion

A strategy to initialize and grow a database for HIT related event reports from the FDA MAUDE database was proposed by retrieving the information from both structured and unstructured fields. The strategy helps us connect up to 0.9% FDA MAUDE reports with HIT events. The creation of this database holds promise in aiding the understanding, characterization, discovery, and reporting of HIT related events.

Acknowledgement

This project is supported by UTHealth Innovation for Cancer Prevention Research Training Program Post-Doctoral Fellowship (Cancer Prevention and Research Institute of Texas grant #RP160015), Agency for Healthcare Research & Quality (1R01HS022895), and University of Texas System Grants Program (#156374).

References


42. Hofmann T. Probabilistic latent semantic indexing. The 22nd annual international ACM SIGIR conference on Research and development in information retrieval; Berkeley, California: ACM; 1999.
44. Hong L, Davison BD. Empirical study of topic modeling in twitter. The first workshop on social media analytics; Washington D.C.: ACM; 2010.
45. Kang H, Gong Y. Design of a user-centered voluntary reporting system for patient safety events. MedInfo 2017 (Accepted).
Overcoming the Maternal Care Crisis: How Can Lessons Learnt in Global Health Informatics Address US Maternal Health Outcomes?

Suranga N Kasthurirathne, BEng1, Burke W. Mamlin, MD2,3, Saptarshi Purkayastha1, PhD1, Theresa Cullen, MD, MS2,3

1Indiana University School of Informatics and Computing, Indianapolis, IN, 2Indiana University School of Medicine, Indianapolis, IN, 3Regenstrief Institute, Center for Biomedical Informatics, Indianapolis, IN

Abstract

Despite unprecedented spending, US maternal outcomes have worsened drastically over the past decade. In comparison, maternal outcomes of many Low and Middle-Income Countries (LMIC) have improved. Lessons learnt by their success may be applicable to the US. We performed a literature review to identify innovations that had met with success across LMIC, and should be considered for adoption in the US. mHealth and patient facing alerts, Telehealth, patient controlled health records, inclusion of patient relationship data in health information systems and positioning empowered community health workers as catalysts of maternal care delivery were identified as innovations worthy of further evaluation. These innovations were categorized into several themes; knowledge, technology, patient/community empowerment, coordination and process change. Tools that place informed and empowered patients and community members at the center of maternal care has greatly improved maternal outcomes, and are suitable to be considered for the US healthcare system.

Keywords: Global health, Maternal health, Clinical Information Systems, Mobile health, Telemedicine

Introduction

US maternal care is currently in jeopardy. At $111 billion a year, US spending on maternal health is twice that of most other high-income countries [1]. US hospital deliveries may cost up to 8,802 USD per birth, compared to an average 2,050 USD in other OECD countries [2]. Despite this spending, a joint report by the World Health Organization (WHO), United Nations Children's Emergency Fund (UNICEF), United Nations Population Fund (UNFPA), the World Bank and the United Nations Population Division states that US Maternal Mortality Rates (MMR) have surged by 136% between the years 1990-2013, from 12 per 100,000 live births in 1990 to 28 per 100,000 live births in 2013 [3]. At over 52,000 cases per year, the risk of severe maternal morbidity (SMM) leading to life-long health problems is even greater [4]. American women are more likely to die during childbirth or related complications than women of any other high-income country [5] or middle-income countries such as Bosnia or Macedonia [6]. The main causes of maternal deaths in the US are non-cardiovascular diseases (15.3%), cardiovascular diseases (14.7%), infection or sepsis (12.7%), hemorrhage (11.3%), cardiomyopathy (10.8%), thrombotic pulmonary embolism (9.0%), and hypertensive disorders (7.6%). The main causes of SMM are maternal age, pre-pregnancy obesity, pre-existing chronic medical conditions, and cesarean delivery [7]. There is also significant discrepancy in outcomes across different populations [8] The MMR among African Americans (40.4 per 100,000 live births) is over three times greater than white parents (12.1 per 100,000 live births). MMR among woman of other races are 16.4 per 100,000 live births [9]. MMR rates of US states with high poverty rates were 77% higher than wealthier states [10].

Research indicates that with the exception of malaria and HIV, the main causes of maternal mortality in low and middle income countries (LMIC) are generally similar to those in the US [11, 12]. However, many of these countries have been able to significantly improve their maternal health outcomes. LMIC such as Timor Leste, Bangladesh, Ethiopia, Uruguay, and India have reported between 78% to 65% reductions in maternal mortality during the same 1990–2013 time period that saw US maternal death rates jump by 136%. [3]. Fortunately, evidence suggests that a majority of US based maternal mortality and morbidities are preventable, and can be addressed via early detection and care [13]. This aligns with the three delay model which recognizes three significant barriers women face in achieving timely and effective medical care; the delay in (a) decision to seek care, (b) reaching care, and (c) receiving adequate healthcare [14]. Informatics based solutions can address the three delays, thereby improving maternal health outcomes. Additionally, changes in healthcare delivery across LMIC during this period have included significant investments in healthcare infrastructure and informatics expertise to help mitigate these barriers. Considerable emphasis has been placed on the use of point of care technology, including mobile phones to deliver actionable information to providers, community health workers, caregivers, and patients, allowing them to address the three delays that cause maternal
mortality [15]. This indicates that such solutions may have contributed toward improving maternal outcomes. Thus, lessons learnt from LMIC may illustrate approaches to improve maternal outcomes in the US. However, not all innovations may be successful due to significant socio-economic and policy differences between the US and LMIC.

We evaluate maternal care based health informatics interventions adopted by LMIC, and determine which of these could help address maternal care challenges faced by the US. We do not seek to perform comprehensive evaluations of care models or correlating process changes brought about by informatics applications, but to recommend which approaches are most suitable to for consideration to address maternal healthcare needs in the US.

Materials and Methods

An introduction to reverse innovation

Traditionally, technological innovation originated, and was consumed by high-income countries. Such innovations were often considered too expensive or unsuited for LMIC needs. Often, product owners avoided expanding to LMIC as they perceived that marketing their solutions across underserved settings would be unfeasible due to high resource requirements. In many cases, innovations from high-income countries were 'whittled down', and produced with limited feature sets for use in low income countries [16]. This environment encouraged innovators from LMIC to develop frugal solutions that were suited for their own needs. However, thanks to qualities of frugality, low operating costs and robustness, innovations meant for LMIC can also be appropriate for adoption elsewhere, particularly in underserved populations in high-income countries. Reverse innovation is the identification of innovations that have already met success in LMIC and adopting these innovations for use in high-income countries [17].

Since its conception, reverse innovation has won widespread acceptance, and adopted by brands such as Microsoft, Nokia, Tata Motors, and Nestle. Products and services created through reverse innovations focus on reducing operating costs and bringing equity with minimal effort, both of which are significantly important for the US health care system. Reverse innovation has the potential to inform a wide host of healthcare challenges [18]. However, the applicability of adopting innovations from LMIC for use in the US healthcare system is relatively unknown, and further work, including implementation science research is necessary to advance the understanding and potential of diffusion of reverse innovations.

Study approach

We searched scientific literature to assess the maternal outcomes challenge in the US. However, searches across OVID medline and Web of Science produced a preponderance of evidence reaching thousands of publications and reports of various quality. To identify the most relevant and accurate literature, we focused on a subset of academic, research and development organizations engaged in improving or monitoring maternal outcomes in the US as sources of information. We searched the websites of these organizations to identify formal reports/publications that discussed major causes of mortality and morbidity. The results of this search were refined using content and discourse analysis methods to identify challenges that are most significant to the US. Next, we performed a literature review to identify how LMIC addressed each of these challenges, and what health informatics innovations could be leveraged to do so. By mapping the aforesaid challenges to successful health informatics innovations, we identified those suitable for further evaluation in the US. We elected to perform a literature review rather than a systematic review as a systematic review would be focused on a single research question, while a literature review would allow us to bring together a broad body of information on a given topic to draw our own conclusions [19].

Results

Based on the analysis of well-reputed information sources (Appendix A), we identified five major challenges affecting US maternal health outcomes;

(a) Lack of trained staff: Rural Americans and underserved populations such as American Indians and Alaska natives (AI/AN) suffer from significant shortages of healthcare providers due to (a) existing staff shortages and (b) tendency of physicians to relocate and practice at affluent urban or suburban areas [20, 21].

(b) Limited access to care: Lack of access to care is a significant challenge to African Americans, Latinos and other minority groups [22], at risk populations, immigrants and low-income earners [23].

(c) Limited sense of ownership/empowerment: Placing an informed patient at the center of their own care can address delays in seeking care, and empower the patient to be an active participants in their own healthcare.
Isolation/lack of community support: Lack of community support and feelings of isolation can lead to depression, stress and inability to manage other life threatening behaviors/conditions [24, 25].

Gaps in patient awareness/education: Lack of awareness leads to risky behavior, inability to manage harmful conditions such as pregnancy related complications [26], and contributes to delays in seeking care.

We performed a rigorous literature review across Ovid Medline, Web of science and EMBASE to evaluate health informatics innovations adopted by LMIC to address these challenges (Appendix B). Search results were refined using the literature inclusion/exclusion strategy presented below (figure 1).

The literature review indicated that efforts to improve maternal care delivery for LMIC had led to the development of numerous tools and approaches. Many informatics based innovations focused on delivery of actionable information to reduce the three delays. We also considered innovations that had failed due to technology, financial or policy based limitations. In the case of such failures, we assessed why they had failed, and determined if these innovations would survive, and help improve maternal outcomes in the US. These lessons were condensed into technological and system change innovations that could address one or more of the five aforementioned challenges, and were suitable for further evaluation by US based settings; (a) mHealth and Patient Facing Alerts, (b) Telehealth / Telemedicine (c) Patient controlled health records (d) Relationships within Health Information systems (e) empowered Community health workers (CHW) as catalysts of healthcare delivery (table 1).

Table 1. US maternal outcome challenges and LMIC based innovations solutions that could address them

<table>
<thead>
<tr>
<th>Challenges</th>
<th>mHealth / Patient Facing Alerts</th>
<th>Telehealth / Telemedicine</th>
<th>Patient controlled health records</th>
<th>Relationships within Health Information systems</th>
<th>CHW as catalysts of healthcare delivery</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lack of trained staff</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Limited access to care</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Limited sense of ownership/</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Isolation/lack of community support</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Gaps in patient awareness/education</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>
mHealth and Patient Facing Alerts

The role of actionable information in improving health outcomes is well documented [27, 28]. The rapid penetration of mobile networks across Africa and Asia has made mHealth efforts a strong success. Examples of success in using mHealth – using mobile devices in health service delivery, to enable maternal care in emergency medical support in Bangladesh, Ghana, Uganda and India, point of care support in Pakistan and Indonesia, and health promotion in Tanzania, Ethiopia, Serbia and Thailand [29-31]. Research indicates that standard care with reminders, disease monitoring and management, and education through cell phone voice and short message service has helped improve health outcomes [32]. mHealth applications also enable better communication between health workers and leverages social networks to improve health outcomes [33]. mHealth based approaches have proved useful in helping address the three delays. mHealth improves maternal outcomes through (a) empowering women to contact health services and access information [34] (b) access emergency obstetric care [35] (c) giving patients knowledge of where to go and (d) enabling lesser trained health workers to provide the ‘right health care’. Despite the success of maternal care applications such as text4baby [36], mHealth and patient alerting tools have failed to undergo widespread penetration in the US due to (a) questions on quality of existing health apps [37] and (b) lack of adequate research to keep pace with ever changing technological progress [38]. However, work in LMIC indicates that a consistent messaging system integrated into the health care delivery model with the right information and a pathway to ‘receive the right care’ has positive impacts on maternal outcomes, especially across underserved populations [31, 39].

Telehealth/Telemedicine

Telehealth/Telemedicine plays an integral role in providing medical care across space and time via telecommunication technologies ranging from the telephone to robotics. It is widely perceived as a method to share medical knowledge in settings where trained staff is scarce. Telehealth is not a new concept to the US. It was widely used by the US military [40]. However, the lack of broadband infrastructure and financial, legal and healthcare policy barriers posed a significant challenge to the early advancement of telemedicine [41], leading health policymakers to relegate Telehealth to a secondary role [42]. However, these barriers are coming down. The Affordable Care Act (ACA) and the American Recovery and Reinvestment Act (ARRA) of 2009 has reduced barriers around the utilization of Telehealth [43]. These together with CMS telemedicine reimbursement changes and state level policy changes promise significant potential for the increased adoption of Telehealth systems [44]. Telehealth is widespread across LMIC, especially for maternal and child health delivery [45, 46]. While we are aware of the significant use of telemedicine within the US, we note that LMIC are creating unique uses for telemedicine using ubiquitous mobile platforms to enable low bandwidth solutions focused on improving maternal care by creating awareness and enabling better care delivery. Based on these advances, we propose that Telehealth, using more ubiquitous platforms such as mobile phones, should be given the place of prominence that it deserves within the US healthcare system.

Patient controlled health records

Patient controlled health records were introduced over two decades ago as a means of creating online repositories where patients could collect and control access to their healthcare data [47]. While revolutionary in its time, projects such as Google Health and Microsoft Health Vault that aimed to operationalize this concept were unsuccessful due to failure to win vendor support, low trust and potential risks [48]. There has been a gradual re-emergence of patient portals integrated into the larger health information systems over the past decade. Unfortunately, these do not create longitudinal patient records as they can only access a single health care system. In comparison, LMIC have developed alternate versions of longitudinal patient controlled health records in the form of smart cards or plastic cards that contain an embedded computer chip to store data [49] as well as paper based medical 'books' that are retained by the patient. While Europe has embraced smart cards for health care, the network centric nature of the US healthcare culture hampers the widespread adoption of Smart cards [50].

One purpose of granting control of their own health record to the patient is to ensure that they become the locus on control in managing their own healthcare. To do so, healthcare networks must uniquely identify each patient. This necessity led to the development of national patient identifiers and Master Patient Indexes (MPI) that serve as registers and resolvers of patient identities in many LMIC. MPI's have been successfully implemented across many LMIC [51]. In comparison, the US has chosen to not embrace a universal patient identifier system due to concerns of privacy [52]. However, rejecting a universal patient identifier has led to limitations in interoperability and data sharing. If implemented correctly, patient controlled health records have the potential for improved shared decision-making between the clinician, caregivers and patient. They also enable early recognition of maternal complications and / or warning signs through access and control of patient records, ultimately contributing to the reduction of the three delays.
Relationships within Health Information Systems

Healthcare is a team sport that depends on collaboration between providers, patients, payers, families, health care teams and communities. When a provider identifies a patient’s need, that need might be reported to a different participant who is able to take appropriate action. Cost of healthcare services and lack of trained staff often result in at least some parts of patient care being delegated to members of their immediate family. However, proper communication and collaboration is essential to make such efforts successful. Relationships between health care teams and family members cannot bring value to a patient unless they are easily discoverable within a healthcare system. Many EHR systems lack support to persist structured patient relationship information, a weakness identified and addressed by developers of the OpenMRS platform, an Open Source Medical Record System designed for use across underserved settings. To date, multiple sites in LMIC have adopted this feature for managing patient relationships with success [53]. We perceive that the need to technologically incorporate critical patient, family and provider relationships within health information systems will be increasingly important to the US as it moves toward a more comprehensive quality based care reimbursement mode, and that efforts should be made to capture and manage such data within a patient’s health record.

Empowered Community health workers (CHW) as catalysts of healthcare delivery

The success of adopting health information infrastructure depends on how information derived from these systems is used impact healthcare delivery. While providers play a crucial role in decision-making, their ability to prevent the three delays outside of a healthcare setting is very limited. CHW’s are frontline public health workers with a close understanding of the community they serve. Their links to the community allow them to reach out to patients at their homes and communities, and improve the quality and cultural competence of service delivery. CHW programs are widespread across LMIC, where they are a predominant topic in discussions around primary health-care delivery. Success stories in using CHW’s for improving healthcare are reported from Uganda, Ethiopia and Tanzania [54]. These successes have encouraged the governments of Brazil, Pakistan, and India to make CHWs a cornerstone of scaling up community health delivery, especially for providing maternal care [55].

CHW’s are not an entirely new concept to the US. They have significant potential to improve maternal outcomes in the US, as demonstrated by successful projects driven by the Indian Health Service and Tribally Operated Health Programs such as the Alaska Community Health Aide Program (CHAP) [56]. Unfortunately, many US based CHW initiatives are short-term, grant-funded projects and grassroots volunteer community initiatives [57]. Unlike many developing countries, US health policy administrators have relegated CHW to a secondary role. However, lessons learned from underserved settings offer much promise for the introduction and use of empowered CHW who are capable of reaching out to pregnant women at the grassroots level. CHW also have the potential to be far more cost efficient, and reduce treatment costs via early evaluation and detection, appropriate referrals, and longitudinal support.

Discussion

Enabling better seeking of care, access to care, and receiving the right care at the right time can significantly reduce maternal mortality and morbidity across both high-income and low-income settings. In both LMIC and the US, these factors are hindered by the scarcity of economic, physical and human resources, or the inability to make them available to the right people at the right time. While US health information infrastructure is more than capable of supporting high quality care where ample resources are available, it falls short of meeting the needs of underserved populations such as minorities, low-income communities and rural areas.

Our research identified five informatics based solutions successfully adopted by LMIC countries, and suitable for adoption in the US; (a) mHealth and patient facing alerting (b) Telehealth (c) Patient controlled health records (d) Relationships within Health Information systems and (e) Community health workers as catalysts of healthcare delivery. In examining these innovations, we found that some of these had already been considered by the US in the past, but failed/impeded by inherent differences in the US healthcare system. However, recent changes in US infrastructure, law and policy aimed at reducing healthcare costs and making healthcare services more accessible [58] and a commitment to comprehensive care models are breaking down these barriers, and improving the potential of adopting these innovations in the US. We found that many of the proposed solutions were strongly interrelated, and that successful maternal care innovations often consisted of combinations of the aforementioned innovations. As shown before, CHW networks in underserved settings depended on information received and integrated via mHealth solutions to guide much of their work. Similarly, efficient persistence and retrieval of relationships, entered into medical records at the point of care, also helped CHW networks. Having improved patient ownership of clinical data enables better dissemination and data integration for use in clinical care, with CHWs, clinicians, other health providers.
and family and community members. We also found that, while the five proposed innovations were dependent upon technology, technology was insufficient by itself. Common themes of knowledge, technology, patient and community empowerment, coordination and process change emerged (figure 2).

**Figure 2.** Enabling better maternal care outcomes. Potential innovations and general themes

The findings indicate that the proper use of health information could play a significant role in improving US maternal outcomes, as well as reducing associated healthcare costs. Many of the innovations that we recommended for evaluation in the US could be used for a wide range of healthcare services over and above maternal care delivery. Investigation of the underlying themes show that the key to maternal care is to place informed patients at the center of their own care. Traditionally, US healthcare delivery is known for its rigid health infrastructure, workflows and provider centric care. Clearly, this approach has done little to address worsening US maternal care outcomes. In comparison, successful innovation across LMIC countries seem to focus on creating and empowering informed patients and their communities to act on their own needs. If so, informatics solutions will be expected to play a crucial role in integrating patients and communities into the healthcare delivery process, informing them, and empowering them to take action on their own needs.

It could be argued that the US still suffers from significantly lesser maternal mortality and morbidity than LMIC despite of the 136% rise in maternal outcomes from 1990-2013, and that many LMIC reported significant improvements in maternal outcomes during this same period starting from comparatively high baseline measures. In other OECD countries, the same measures have remained relatively constant with significantly lower costs of care, suggesting that the rise in maternal mortality and morbidity is a significant concern. More so because a majority of these outcomes were preventable, and because health outcomes worsened despite of significant increases in financial spending in the US, which indicates the need to re-think current healthcare focus areas. Additionally, health IT based interventions have demonstrated their ability to reduce preventable maternal mortality and morbidity in LMIC [59, 60], indicating that they may be able to do the same in the US. Furthermore, every challenge identified in table 1 could be addressed via health informatics innovations. Another possible area for debate would be contextual changes between LMIC settings and the US. US based healthcare is far more expensive than LMIC or even other high-income countries. However, the most adversely hit populations in the US are at-risk populations and the poor, who also benefit from compatible state funded health programs. Additionally, maternal care is very well funded, with multiple welfare/aid options for those in need. Therefore, adoptations of LMIC based approaches for the US is feasible. As future steps and further validation of our findings, we propose the evaluation of each of the innovations identified by our study using the Technology Acceptance Model (TAM) [61]. The TAM model presents how target users accept and use a given technology, and thus, will further evaluate the potential of successfully adopting such innovations across the US.

**Conclusion**

Our study identified significant potential for adopting reverse innovations to help address weak maternal care outcomes in the US. Many restrictions that historically hampered these innovations have already been removed, leading to improved chances for their success. Recent US policy and administrative changes aimed at reducing
healthcare costs and creating comprehensive health care delivery models are also making these solutions more realistic. General themes that drive these innovations indicate that the key to better maternal care outcomes is to place patients at the center of their own care by empowering informed patients and their community to act on their own needs. While the role of these factors and themes in enabling improved health outcomes is widely recognized, the importance of a holistic model that includes, but is not limited to informatics is often disregarded. We highlight the intersections of technology and process changes that have been shown to improve care in LMICs. We believe that there is ongoing evaluation and future work in this area that needs to occur, including a rigorous comparison of multi-country implementations, and an evaluation of the return on investment of these identified innovations. However, recognizing that LMIC have developed appropriate technological and process solutions that can benefit underserved settings in high-income countries is critical. Adversity itself can be a springboard for innovation. In many cases, technological solutions developed and lessons learned in LMIC are applicable and beneficial to the high-income countries. It behooves us to identify, evaluate and incorporate appropriate solutions, regardless of their origin, as we continue to seek global health equity.

Appendix A

<table>
<thead>
<tr>
<th>Organization</th>
<th>Type of publications/reports</th>
</tr>
</thead>
<tbody>
<tr>
<td>World Health Organization (WHO)</td>
<td>Various reports that include assessments on US maternal outcomes</td>
</tr>
<tr>
<td>Centers for Disease Control and Prevention (CDC)</td>
<td>Annual and weekly reports, data and statistics, reports on major conditions/illnesses</td>
</tr>
<tr>
<td>US Department of Health and Human Services</td>
<td>National surveys, State and national data, research reports</td>
</tr>
<tr>
<td>Indian Health Service (IHS)</td>
<td>Reports to Congress, various publications</td>
</tr>
<tr>
<td>The American Congress of Obstetricians and Gynecologists (ACOG)</td>
<td>Committee opinions, department publications and endorsed documents, journals, position statements, advisories and bulletins.</td>
</tr>
<tr>
<td>The National Academies of Sciences, Engineering, and Medicine</td>
<td>Various reports/publications from the Institute of Medicine (IOM)</td>
</tr>
<tr>
<td>Agency for Healthcare Research and Quality (AHRQ)</td>
<td>Maternal care reports an analytical data reports</td>
</tr>
</tbody>
</table>

Appendix B

(a) OVID Medline database

<table>
<thead>
<tr>
<th>Expression no.</th>
<th>Searches</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>exp Developing Countries/ or underserved nations.mp.</td>
<td>68199</td>
</tr>
<tr>
<td>2</td>
<td>&quot;low and middle income countries&quot;.mp.</td>
<td>6458</td>
</tr>
<tr>
<td>3</td>
<td>lmic.mp.</td>
<td>756</td>
</tr>
<tr>
<td>4</td>
<td>exp Africa/</td>
<td>221998</td>
</tr>
<tr>
<td>5</td>
<td>exp Asia/</td>
<td>663253</td>
</tr>
<tr>
<td>6</td>
<td>exp Latin America/</td>
<td>9580</td>
</tr>
<tr>
<td>7</td>
<td>electronic medical record.mp. or exp Electronic Health Records/</td>
<td>15919</td>
</tr>
<tr>
<td>8</td>
<td>exp Medical Records Systems, Computerized/</td>
<td>32097</td>
</tr>
<tr>
<td>9</td>
<td>exp Information Systems/</td>
<td>200066</td>
</tr>
<tr>
<td>10</td>
<td>exp Maternal Health Services/ or maternal care.mp.</td>
<td>42930</td>
</tr>
<tr>
<td>11</td>
<td>exp Maternal Health/ or exp Maternal Mortality/ or exp Pregnancy Complications/</td>
<td>390150</td>
</tr>
</tbody>
</table>
Due to the limited number of publications retrieved, we searched for all publications that discussed maternal complications and health information systems, and filtered publications that dealt with LMIC via manual review.

TITLE: (maternal health* OR maternal care* OR maternal mortality* OR maternal morbidity* OR Pregnancy Complications* OR Postpartum Hemorrhage* OR Pre-Eclampsia*) AND TITLE: (electronic health record* OR health information system* OR clinical information system* OR computer*)

References

Modes of De-identification

Mehmet Kayaalp, MD, PhD
Lister Hill National Center for Biomedical Communications
U.S. National Library of Medicine, National Institutes of Health, Bethesda, Maryland

Abstract

De-identification of protected health information is an essential method for protecting patient privacy. Most institutes require de-identification of patient data prior to conducting scientific studies; therefore, it is important for clinical scientists to be cognizant of all modes of de-identification and all services provided by their de-identification tools. In this article, we discuss eight different modes of de-identification that yield de-identified data at different levels of quality. Most of these modes can be used in combination to achieve the best performance.

Introduction

De-identification is a process of detecting identifiers (e.g., personal names and social security numbers) that directly or indirectly point to a person (or entity) and deleting those identifiers from the data. Health information containing personal identifiers that are linked to the subject of the information is defined as protected health information and is protected under the federal law. The deleted personal identifier (e.g., “James Smith”) can be replaced with information describing the type of identifier (e.g., [Personal Name]) or with fake identifiers called pseudonyms (e.g., “John Doe”). Replacing personal identifiers in the clinical data with non-identifying terms minimizes the chance of re-identification of the patient during the use of the data for scientific purposes; hence, it is one of the most essential tools for protecting patient privacy.

According to the Privacy Rule of the Health Insurance Portability and Accountability Act (HIPAA), the de-identification process can be accomplished using two separate pathways: Safe Harbor method and Expert Determination approach (see Figure 1). The Safe Harbor method requires all 18 personal identifiers to be eliminated. The latter approach uses the preservation of certain personal identifiers (usually dates and demographics) combined with an expert’s assurance that these identifiers could not be used to re-identify the patient. Note that the Privacy Rule neither defines the qualifications of the expert or the expertise nor sanctions any set of methods that would yield such a determination.

If dates, ages over 89 years, and/or detailed geographic information (five-digit ZIP code or at the town level of details) are required in the study, Safe Harbor method can be used along with another HIPAA Privacy Rule provision called Limited Data Set. Under this provision, researchers can access those pieces of information if they agree to sign a data use agreement established by the data-providing institute. Compared to the Safer Harbor, the expert determination approach is less definitive, so we do not consider it here. For this article, the term de-identification implies the Safe Harbor method only.

Anonymization is closely related to de-identification. It is sometimes used interchangeably and confused with de-identification. Although both anonymization and de-identification aim to protect the privacy of the subject of information, they are semantically different concepts. As a federal regulation, HIPAA Privacy Rule defines 18 types of personal identifiers. State laws and institutional review board (IRB) regulations that are more stringent than the Privacy Rule may require the removal of an additional set of identifiers such as provider identifiers and specimen slide numbers. However, in all cases, the process of de-identification is defined explicitly; that is, finding specific identifiers and deleting them. Thus, de-identification is a well-circumscribed process with explicit specifications of what needs to be done. The same cannot be said for anonymization. Anonymization is not a method per se but a goal, which may be achieved using different methods and strategies. Anonymization does not imply a standard nor does it specify what needs to be done to attain that goal.

Another crucial difference between anonymization and de-identification pertains to the claim of the outcome. The Department of Health and Human Services (DHHS) “is cognizant of the increasing capabilities and sophistication of

* Although the secondary use of clinical information may go beyond scientific purposes, such as financial analysis, fraud detection and marketing. American Medical Informatics Association. A Taxonomy of Secondary Uses and Re-Uses of Healthcare Data. Policy Meeting, 2007. Our scope is limited to non-commercial scientific use only.
electronic data matching used to link data elements from various sources and from which, therefore, individuals may be identified” (see Federal Register p. 53232). The de-identification process minimizes the risk of re-identification but has no claim to make it impossible. On the other hand, methods for anonymization found in the literature such as aggregations of microdata are almost always applicable to tabular data only and attempt to guarantee a certain level of anonymity usually as a function of aggregation.

Figure 1. Two pathways for de-identification of protected health information as sanctioned by HIPAA Privacy Rule as depicted in the guidelines prepared by the Office of Civil Rights, Department of Health and Human Services

Automatic De-identification

We can de-identify tabular datasets manually if we know which fields contain personal identifiers. Unfortunately, de-identifying narrative text is significantly more challenging than de-identifying tabular data. Narrative text has neither a schema nor enumerated set of fields to guide the process. Narrative report structures tend to differ randomly in location and labeling among different providers. Moreover, the ambiguous nature of natural languages makes the problem of narrative text de-identification especially difficult.

Automatic text de-identification tools can de-identify clinical text with high accuracy comparable to the performance of human annotators. Furthermore, automatic de-identification is fast and inexpensive. Unlike human annotators, computers do not get tired but instead, consistently produce the same level of quality for every document as long as these documents do not differ from each other in style and context.

Modes of Operations

A software application may be used in different configurations, forms or modes, e.g. batch vs. interactive mode, text vs. graphic mode, beginner vs. expert mode. In the literature, there are plenty of de-identification papers about automatic clinical text de-identification systems, but there is little information on how to use them to get the best de-identification results. In this paper, we discuss different modes of de-identification, each of which may yield results at different levels of quality.

Most IRBs require de-identification as a prerequisite for clinical research using retrospective patient data; therefore, it is important for every clinical scientist to be cognizant of all options and services provided by de-identification systems.
Institutional Responsibilities

De-identification plays a crucial part in protecting patient privacy by enabling institutions to share a large volume of clinical information for clinical research while maintaining patient privacy. However, de-identification alone is not sufficient for protecting patient privacy. This is a difficult task that can only be achieved through a close collaboration of all parties in the healthcare system chain who create, handle, store, and transmit health information. Furthermore, institutions have overarching responsibilities to establish the right policies and monitor their compliance through IRBs. Institutions have to ensure that the data leaving their control has been properly de-identified.

As part of this responsibility, data managers in clinical institutions should make the necessary efforts to provide requisite patient and provider information to their de-identification system for achieving the highest de-identification rate possible. For example, an automatic de-identification system capable of accepting patient and provider identifiers as inputs should be provided to the system, if possible, so that institutions are not left with subpar de-identification results. Similarly, a de-identification system that requires annotated clinical reports for training should be provided with the requisite training dataset. To ensure that the training dataset is large enough, the machine learning system should be continuously fed with new sets of annotated reports until the de-identification performance increases become negligible. Furthermore, this process should be repeated in scheduled cycles as clinical practice, use of terminology, reporting styles and reporting physicians tend to change over time.

Different Modes of De-identifying Reports Using Automatic De-identification Systems

In this article, we consider any automatic de-identification application as a black box and remain indifferent to the methods used inside the box. The focus is on the process of de-identification from the perspective of the user. The user needs to provide a set of input to the system, so we are interested in the variety of information that the system is capable of taking as input and how the user can operate a given de-identification system to achieve the best patient privacy protection while preserving the integrity of de-identified data. We discuss eight modes of de-identification (see Table 1).

| A. Repository-wide batch de-identification |
| B. On-demand cohort-specific de-identification |
| C. On-demand de-identification of query results |
| D. De-identification with patient and provider identifiers |
| E. Scientist involved de-identification |
| F. Patient involved de-identification |
| G. Physician involved de-identification |
| H. Online de-identification by honest brokers |

Repository-wide batch de-identification. This is the simplest among all the alternative processing methods; thus, perhaps the most commonly used among institutions. When institutions obtain a de-identification system, some may immediately create a de-identified version of their entire repository, so that researchers with proper credentials can quickly access what they need in a de-identified format.

On-demand cohort-specific de-identification. An alternative to the repository-wide batch de-identification method is providing on-demand de-identification for a defined cohort of patients at the request of the scientist. Unlike the repository-wide batch de-identification approach, the de-identified data is not ready to be provided to the research group, but given the speed of modern systems, automatic de-identification is almost instantaneous. The time difference between these two approaches can be negligible compared to the selection of the cohort reports.

On-demand de-identification of query results. If the de-identification system is attached to a clinical database querying system, the de-identification can also be obtained on the fly. This approach would probably be much faster than the cohort-specific approach, since there would be no need to wait for the availability of the data manager who, in the previous two modes, de-identifies the data and provides the de-identified data to the scientist. In this mode, the data manager is out of the loop.
De-identification with patient and provider identifiers. Research results indicate that providing patient and provider identifiers likely to be in the report yields significantly better de-identification results. Personal identifiers can be provided to a de-identification system in four different ways: report-specific identifiers, cohort-specific identifiers, repository-wide identifiers, or a combination of the above. For provider identifiers, repository-wide mode, purging all provider identifiers listed in the institution’s file is an attractive approach, because each patient can be seen by many of the providers who may be mentioned in the patient’s reports.

Patient identifiers are the most important identifier that need to be detected and eliminated from reports. Each report in an electronic health record system (EHR) is typically tied to many patient identifiers. The de-identification system can take advantage of that information and can detect those identifiers in the text at the highest sensitivity level. For example, if the last name of the patient is Cushing, the de-identification system would not overlook the word by erroneously presuming that it refers to Cushing’s disease.

In certain record types, specifically in HL7 v.2 records, additional personal identifiers, such as the patient’s address, telephone number, medical record number, and next of kin identifiers are also available as part of the PID and NK1 segments.

In some circumstances, providing patient identifiers for each report may not be feasible but the identifiers of the patients retrieved for a specific research cohort could be provided as a single set of input to the de-identification system.

Identifiers of providers who dictate and sign reports are also found in EHRs. Inputting both repository-wide and report-specific provider information may improve de-identification results.

Scientist involved de-identification. Clinical text de-identification systems are not perfect. Because their primary goal is to remove all possible words and other alphanumeric strings that might identify the patient, they tend to favor sensitivity (for identifiers) over specificity, and may inadvertently remove non-identifying health information in the process. Users of the de-identification systems (i.e., scientists, directly or data managers, indirectly) can compensate for the problem by providing clinical terms that they would like to preserve from excessive de-identification. A capable system may automatically increase its sensitivity level when it receives such clinical metadata.

Patient involved de-identification. Many clinical institutions allow patients to access their health information stored in EHRs or personal health record systems. As today’s consumers become more and more suspicious of how their data have been shared by Internet companies without their knowledge, the transparency in the healthcare sector could alleviate such concerns. Institutions can recruit patients who volunteer to help de-identify their health records and/or verify that their personal identifiers have been properly purged from their data. Although it is not in current practice, we suspect some patients would demand this level of transparency and the option of self-verification from their providers in the near future.

Note that some identifying information can be inferred from context of the narrative or through circumstantial information instead of just personal identifiers. For example, “the examination of his injury that he endured during his US championship match today...” We call such information personally identifying context, which is hard for any automatic de-identification system to detect. The patient’s assistance would be very valuable for de-identifying such information.

Physician involved de-identification. Patients’ names and chart numbers get into narrative reports because physicians use patient identifiers when they dictate. In a few special cases, e.g. pathology reporting, a requirement exists to dictate both the patient identifying information and the specimen number into the report to tie that report securely to the specimen and the patient. For most dictated reports, however, recording the patient’s identifiers in the body of the report is not needed because the patient is usually identified to the transcriber prior to the dictated note. But from their habits in conversation with patients or other care providers, many providers often say “Mr. Jones,” or “Fred Jones” instead of simply “the patient.” Medical schools and professional societies should discourage the recording of patient names, or identifiers in the body of the dictation. De-identification systems can highlight names and identifiers in transcribed dictation pre signature, which may help providers to break this habit.

Online de-identification by honest brokers. With the advances on big data, scientists start accessing larger patient cohorts than what is available in their own institutes. As this trend continues, we can expect some of the big data

† Provider identifiers are not necessarily part of the protected information, at least according to HIPAA Privacy Rule, but other laws, regulations and policies may require their elimination through de-identification.
processes, including de-identification, are going to be centralized. At that point, de-identification of clinical data can be provided as an online service by these centers. Small research institutes with less funding may also find this option attractive as they can rely on external expertise for proper de-identification. These centers can act as honest brokers providing expertise and assistance to scientists and clinical institutions for their de-identified data needs.

Discussions

We have introduced eight distinct modes of de-identification. Each of these modes comes with advantages and disadvantages, which we discuss in this section.

Institutes with smaller budgets for de-identification may find repository-wide batch de-identification attractive because of its simplicity, but they may not repeat the de-identification process on the same repository data. Since automatic clinical text de-identification systems improve continuously, the quality of the de-identified data improves over time; therefore, the repository data de-identified a number of years ago may not be on par with the current standards. Relying solely on this mode has the danger of stale data, which contain too many intact personal identifiers and too many false positives in the “de-identified” data according to state-of-the-art. If an institution would like to use this mode alone, the IRB should impose a policy for daily or weekly de-identification of new incoming data and cyclical re-de-identification of the repository data using the latest version of the automatic de-identification software.

On-demand de-identification practices solve the stale data problem since the institution could use the latest technology available to produce up-to-date de-identification. Furthermore, the de-identification output can be tailored according to the scientist’s needs. For example, if certain demographic information is necessary, it can be preserved in the output, producing a limited data set. If the de-identified data is produced via batch mode, the output would be either a fully de-identified set (hence the scientist’s needs could not be addressed) or a particular type of limited data set. In other words, the batch mode would be a generic one-size-fits-all approach.

On the other hand, producing a liberal limited dataset may be a viable, middle ground solution for those institutions, if their IRBs allow it. Due to the complexity of natural languages, no de-identified clinical report should be considered safe to share openly except for trivial cases; thus, even fully de-identified clinical reports should not be shared without a data use agreement, which is imposed by HIPAA Privacy Rule for all limited data sets. Because scientists, including those who would receive fully de-identified reports, should sign into a data use agreement as if they were receiving limited data sets, the institution may develop a policy to produce limited data sets only. Since limited data sets comprise PHI, this approach would not be in accord with the Privacy Rule for entities covered by HIPAA. Privacy Rule establishes the standard called the “minimum necessary” indicating limiting the content of PHI to the necessary minimum. 45 CFR 164.502(b): “When using or disclosing protected health information or when requesting protected health information from another covered entity or business associate, a covered entity or business associate must make reasonable efforts to limit protected health information to the minimum necessary to accomplish the intended purpose of the use, disclosure, or request.”

Institutes that use an on-demand de-identification mode can employ a separate batch mode as well. Before submitting a research protocol to their IRBs, scientists need to analyze existing datasets to understand whether available datasets are suitable to their research needs. A de-identified copy of the entire repository would be beneficial for scientists at this stage of their research. A policy that would require such a research repository would better protect patient privacy.

On-demand de-identification of query results can be done quite easily due to the fast de-identification performance of modern automatic de-identification systems such as NLM-Scrubber, which de-identifies a typical report in a fraction of a second, much faster than the time required for a person to read the same text. The downside of this mode is that it is more difficult to integrate a standalone application to an EHR. On the other hand, if the de-identification system is integrated to the EHR, not only this particular mode but also the other two modes in which physicians and patients are involved, can be put in practice. When all of these modes are used in parallel, the performance of de-identification can be maximized.

Cohort-specific on-demand de-identification requires more time to access the data than the query results de-identification, but some institutions may prefer this approach, since the data manager could provide oversight to the de-identification process and monitor the level of privacy protection.

Properly providing patient and provider information to the de-identification system is necessary for better de-identification if the de-identification system is capable of accepting them. This mode requires extra steps from the institutions and data managers. Those steps might not be taken unless the institution establishes a policy to require these actions and the IRB monitors the compliance.
De-identification may inadvertently eliminate some informative clinical terms (e.g., newly introduced gene names) that mimic personal identifiers. A modern de-identification system should be capable of accepting new lists of words and concepts and set the bar of their removal higher to preserve the integrity of the content. If a repository-wide batch de-identification is adopted, it would be more difficult to introduce such functions.

If a de-identification system is used with the repository-wide batch mode, it would have to employ a single set of sensitivity and specificity levels to maximize the utility of the system. When the designer of a de-identification system (and in capable systems, the user of the system) adjusts the de-identification sensitivity of the system to eliminate all personal identifiers, the number of false positives could increase in tandem. If an on-demand de-identification mode is adopted and the scientists provide their metadata, the system can perform at a higher sensitivity level to eliminate all personal identifiers.

Online de-identification mode has a promising future. Large non-profit research organizations such as NIH or State Cancer Registries can employ the requisite expertise and human capital and provide these services to other research institutes free of charge. The benefits of this mode are that the de-identification can be improved and stringent policies for patient privacy can be employed. Furthermore, the data from a number of institutions can be hosted at these sites or in the cloud; thus, scientists can access much larger cohorts and conduct research on big data. Since these datasets for cancer have already been collected at the state level, cancer registries can take this responsibility if necessary funding can be provided. At some point in the future, this could be achieved at the federal level as well, but governance of the data at the federal level could be more complicated and may require new laws and regulations as well as a lot of convincing before allowing data to be shared across states and institutions.

Note that a number of these modes can be used in parallel. Consider the following as a possible scenario. An institute may use repository-wide batch de-identification so that the scientists at the institution can study the repository data for preliminary research. Since the preliminary study was done on de-identified data, some clinical terms could be deleted by the de-identification process and the scientist might not have retrieved all patient cases. During the study, the scientist can query the original data containing personal identifiers and receive the fully de-identified results. After studying the data and observing the eliminated terms, the scientist can provide a list of terms to preserves the data and may request certain date and age information be preserved. The data manager can repeat the de-identification by using patient and provider information as well as the list of clinical terms as inputs to the system and adjust the system to preserve date and age information. After the data manager’s review of the results, the scientist would receive the resulting limited data set with all pertinent clinical information preserved. If available, the scientist can repeat this process by accessing a clinical data warehouse in the cloud to enhance the patient cohort using big data. This process can be improved further by involving patients and physicians in the de-identification process.

Conclusions

There are many different ways to perform de-identification. Most of these modes of de-identification operations can be combined in order to reach ideal results. The success of the de-identification process depends not only on the competency of the automatic de-identification systems, but also on the competency, dedication, and discipline of the institution that is responsible for de-identification of protected health information. Users of de-identification systems need to learn how to operate the different modes properly, recognize and obtain all available patient and provider identifiers, gather the terminology needs of the scientist, and tailor the de-identification to achieve the highest quality of research data and patient privacy.

Funding

This work was supported by the Intramural Research Program of the National Institutes of Health, National Library of Medicine.

Competing Interests

The author is the principal investigator in NLM-Scrubber project at NIH. He receives royalties from University of Pittsburgh for his contribution to a de-identification project; the resulting product was acquired by a third party, which today is known as De-ID Data Corp. NLM’s Ethics Office reviewed and approved his appointment.
References

Comparison of algorithm advice for post-acute care referral to usual clinical decision-making: examination of 30-day acute healthcare utilization

Susan K. Keim MS, MSN, CRNP1, Kathryn H. Bowles, PhD, RN, FAAN, FACMI 1, 2
1University of Pennsylvania, Philadelphia, PA 2Visiting Nurse Service of New York, New York, NY

Abstract
Objective: Compare patient characteristics and acute healthcare utilization between patients identified as in need of post-acute care (PAC) by the clinical decision support (CDS) algorithm yet were discharged home without services, to those where the CDS and hospital clinicians agreed on no referral.
Methods: Retrospective analysis of hospital administrative and clinical data for 1,366 patients.
Results: 30-day acute healthcare utilization rates are significantly higher for those patients flagged as in need of PAC referral. There are also significant differences in patient characteristics based on referral risk.
Discussion: Clinicians were blinded to the algorithm enabling the comparison of usual care to decision support. Future work will examine the effect of sharing algorithm advice with clinicians on PAC referral rates and utilization.
Conclusion: The CDS algorithm clearly identified patients with high-risk characteristics and those who will go on to utilize acute care resources. Providing CDS to discharge planners may improve patient outcomes.

Background and Significance
Efforts to reduce 30-day inpatient (IP) hospital readmissions continue to garner attention since the 2013 inception of Medicare reimbursement penalties to hospitals with above-average rates for high-volume conditions e.g., heart failure (HF), pneumonia, and acute myocardial infarction (AMI).1, 2 Although readmission rates have declined from 18.1 per 100 patients in 2011 to 17.3 per 100 in 2013,3 concerns remain that many Medicare readmissions are potentially preventable,4, 6 cost $12 billion annually,4 and could be lowered through better discharge planning (DP) processes.7 Furthermore, approximately one in five patients presents to the emergency department (ED) within 30 days of an IP hospitalization,8 and over half of these patients, especially if Medicare beneficiaries, are readmitted for an IP or short-term observation (OBS) stay.8-12 Efforts to mitigate this acute healthcare utilization in the post-discharge period include improved coordination of post-acute care (PAC) services i.e., home healthcare (HHC) or facility-based care (skilled nursing facilities or long-term care institutions).13, 14

PAC services comprise one of the fastest growing sectors for Medicare spending,15 but referral decisions are frequently based on non-standardized DP practices,16 resulting in the potential for misaligned utilization. Over-utilization of PAC services incur significant costs to Medicare,17 and under-utilization can leave patients at risk for poor outcomes including hospital readmission.19 DP is a multifaceted and complex process requiring the collaboration of a multidisciplinary healthcare team i.e., physicians, nurses, social workers physical therapists, in partnership with patients and their caregivers. Due in part to shortened lengths of hospital stays (LOS), clinicians often do not have the time or information necessary to make optimal decisions regarding PAC referrals.20 Despite efforts to begin this process early within the hospital stay, most discharge decisions are rushed and made close to the time of discharge.21 Shortened planning time can impact the ability to effectively arrange for HHC visits or facility placement. In a recent web-based national survey of hospitals enrolled in the Hospital-to-Home (H2H) quality improvement initiative, only 34.6% reported that they formally estimate the risk of readmission for HF and AMI patients,22 and less than one quarter reported arranging home visits for most or all patients with these high-risk conditions after discharge.23

Studies have shown that early24 and coordinated DP is more effective than usual care at reducing IP readmission risk.25 However, discharge plans should be tailored to patient’s needs,26 and consistently employed by the multidisciplinary team.27-29 Clinical Decision Support (CDS) tools can aid in improving and standardizing DP,30, 31 and in identifying the need for PAC services.32, 33 A navisHealth® proprietary discharge referral decision support software (IDENTIFY), analyzes data from a validated risk algorithm developed by Bowles and colleagues,31, 34 and notifies discharge planners of PAC risk early in a patient’s hospital stay. Although recommendations for specific PAC locations were not included in this algorithm, IDENTIFY implementation has resulted in significantly lower 30-day IP readmissions for high-risk older adults.35, 36

An expanded CDS algorithm, developed by Bowles and colleagues,37 not only identifies patients in need of PAC referral, but also recommends specific PAC sites (HHC or facility). This algorithm was created from expert panel analyses of 1,496 nationally representative case studies38 of hospitalized patients age 55 and older from six hospitals in
the Northeast, Midwest and Mid-Atlantic regions of the United States. The expert panel, comprised of clinicians from multiple disciplines e.g., physicians, nurses, social workers, physical therapists, identified those cases in need of PAC referral, and if applicable, the specific location for PAC services. Regression modelling was performed to develop a 2-part automated 29-item (21 are unique variables) risk algorithm. The first part of the algorithm (16 variables) flags the need for PAC referral and the second part (13 variables) recommends the optimal site of care. Validation with a holdout sample produced areas under the curve of 91.5 and 89.7 respectively for each part of the algorithm indicating very good to excellent models. Now also a proprietary technology, this algorithm computes a risk score based on a variety of patient characteristics that are routinely collected during a hospital admission including measures of activities of daily living (ADLs), LOS, number of comorbidities, fall risk assessment, and pressure ulcer risk. This score is available within 24 hours of admission and can aid healthcare providers and discharge planners in proactively identifying high-risk patients therefore allowing time for effective DP.

The overall purpose of this study is to compare patient characteristics and acute healthcare utilization between patients identified in need of post-acute care (PAC) by the clinical decision support (CDS) algorithm, yet were discharged home without services, with those where the CDS and hospital clinicians agreed on no referral for PAC services. The specific aims and hypotheses (H) of this study are:

Aim 1: In patients discharged home without services, compare the patient characteristics of those flagged versus those not flagged by the algorithm for PAC services.

H 1: Patients discharged home without services will have more limitations in patient characteristics if flagged versus not flagged by the algorithm.

Aim 2: In patients discharged home without services, compare the rate of 30-day acute healthcare utilization in those flagged versus not flagged by the algorithm for PAC services.

H 2: In patients discharged home without services, those flagged by the algorithm will have a significantly higher rate of acute healthcare utilization compared to those not flagged for PAC services.

Methods
Design
A secondary analysis is performed using de-identified data acquired from the control phase of a multi-center study of quasi-experimental pre-post design (NR01007674) aimed at measuring the effectiveness of the algorithm on PAC referrals and patient outcomes. During the control phase, clinical and administrative data were collected retrospectively from two suburban mid-Atlantic, teaching hospitals within one health system, totaling 800 IP beds with 48,000 admissions and 132,000 ED visits annually. Data was collected on patients’ index hospitalizations from May 8, 2015 through September 11, 2015 with an additional 30-day follow-up period to measure post-discharge acute healthcare utilization. During the control phase, clinicians were blinded to the algorithm scores, and approaches for usual clinical decision-making were utilized with respect to discharge planning and disposition.

Hospital administrative and admission, discharge and transfer (ADT) databases are utilized to collect the number of comorbidities, employment status, admission/discharge dates, primary diagnoses (ICD-9-CM codes) as well as the acute healthcare utilization variables: 30-day hospital readmission or 30-day ED/OBS visit. Clinical data extracted from the electronic health record (EHR) supply Morse (fall risk) and Braden (pressure ulcer) scores, changes in ADLs, number of prior hospitalizations within six months, level of self-rated health, depression assessment, and additional socio-demographic data (algorithm components).

Sample
Eligible patients were age 55 and older, admitted for an initial IP stay and discharged alive during the study period. Patients were excluded from the initial control cohort if they were admitted for an OBS stay only or were missing greater than seven algorithm variables determined to be key through sensitivity testing in the parent study. Primary diagnoses not relevant to the sample e.g., perinatal codes, and those potentially related to seasonality impacting the intervention phase of the parent study e.g., fractures, were omitted from the sample. In the overall final control cohort of 3,302 patients, those who received a referral from the hospital and therefore discharged with PAC services, were excluded, yielding a final sample of 1,336 patients, for this secondary analysis.

Statistical Analysis
The study sample, which represents those discharged to home without PAC services, was subdivided into two groups (Figure 1): those whom the algorithm flagged for PAC services (n=895), and those who were not flagged (n=441).
These subsets were compared using descriptive statistics with non-parametric (Chi-square, Fisher-Exact, or Kruskal-Wallis) tests as appropriate. Acute healthcare utilization was examined using Kaplan-Meier survival analysis with log rank tests to measure significance between subsets by algorithm flag. Finally, risk for acute healthcare utilization between subsets was calculated using Mantel-Haenszel weighted risk-ratio (RR) estimates and stratified based on socio-demographic and clinical factors.44, 45

Figure 1. Diagram of Sample Derivation

Results
In the 1,336 sample, patients were on average 71.8 years old (11.15 SD, range 55-103), predominantly white (84%), and slightly more than half were male (51%). Greater than 50% had higher than a high school education, 56% were married/partnered, and approximately 60% were retired. They tended to rate their health as average-to-good (78%) and most had not had a prior admission six months before the index hospitalization (70%). Among a number of ICD-9-CM categories, the most common primary diagnoses were classified as diseases of the circulatory (27%), digestive (19%), and respiratory systems (11%) and the average number of comorbid conditions were 3.0 (1.79 SD, range 3-14). The index hospitalization mean LOS was 3.3 days (2.17 SD range 1-21). Nursing staff identified 54% of patients at moderate risk for falling, and 89% at low risk for pressure ulcer development; less than 20% had any decline in ADLs during their admission. Finally, 67% (895/1336) were flagged by the CDS algorithm for PAC referral but were discharged home without services; of these, the algorithm recommended 36% (322/895) for HHC services and 64% (573/895) for facility-level care.

Differences in Patient Characteristics by Algorithm Flag
Hypothesis one was supported: Patients discharged without hospital PAC referral, have significant differences in various patient characteristics based on whether they were flagged versus not flagged by the algorithm for PAC services. Comparative analysis of eight socio-demographic characteristics (Table 1) revealed significant differences based on algorithm flag for PAC referral. Those 895 patients who were flagged for PAC services, tended to be older, single, unemployed/retired, and resided in an assisted living, or some type of, facility prior to index hospitalization (p<0.001). Those 441 who were not flagged were more often male, married, and well-educated (p<0.001).
### Table 1. Comparison of Socio-demographic Characteristics by Algorithm Flag (N=1336)

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Flagged (n=895)</th>
<th>Not Flagged (n=441)</th>
<th>p-value¹</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean [SD (Range)] or n (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age (years)</td>
<td>73.9 [11.2 (55-102)]</td>
<td>67.6 [8.7 (55-95)]</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Gender¹</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>415(46.4)</td>
<td>261(59.2)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Female</td>
<td>479(53.5)</td>
<td>179(40.6)</td>
<td></td>
</tr>
<tr>
<td>Race²</td>
<td></td>
<td></td>
<td>0.01</td>
</tr>
<tr>
<td>White</td>
<td>737(82.4)</td>
<td>383(86.9)</td>
<td></td>
</tr>
<tr>
<td>Black</td>
<td>119(13.3)</td>
<td>52(11.8)</td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>36(4.0)</td>
<td>5(1.1)</td>
<td></td>
</tr>
<tr>
<td>Ethnicity¹</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hispanic/Latino</td>
<td>9(1.0)</td>
<td>11(2.5)</td>
<td>0.04</td>
</tr>
<tr>
<td>Non-Hispanic/Latino</td>
<td>876(97.9)</td>
<td>427(96.8)</td>
<td></td>
</tr>
<tr>
<td>Marital Status²</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Married/Partnered</td>
<td>406(45.4)</td>
<td>344(78.0)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Single/Divorced</td>
<td>221(24.7)</td>
<td>56(12.7)</td>
<td></td>
</tr>
<tr>
<td>Widowed</td>
<td>219(24.7)</td>
<td>23(5.2)</td>
<td></td>
</tr>
<tr>
<td>Education²</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Some High School</td>
<td>481(53.7)</td>
<td>150(34.0)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Post-High School</td>
<td>386(43.1)</td>
<td>286(64.9)</td>
<td></td>
</tr>
<tr>
<td>Employment Status²</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Currently Employed</td>
<td>124(13.9)</td>
<td>207(46.9)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Unemployed</td>
<td>89(10.0)</td>
<td>13(3.0)</td>
<td></td>
</tr>
<tr>
<td>Retired</td>
<td>608(68.0)</td>
<td>195(44.2)</td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>43(4.8)</td>
<td>17(3.9)</td>
<td></td>
</tr>
<tr>
<td>Living Arrangement²</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>House/Apartment</td>
<td>746(83.4)</td>
<td>420(95.2)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Assist Living/Facility</td>
<td>77(8.6)</td>
<td>5(1.1)</td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>47(5.3)</td>
<td>13(3.0)</td>
<td></td>
</tr>
</tbody>
</table>

¹p-values calculated using Chi-square, Fisher-Exact or Kruskal-Wallis Tests
²Unknowns excluded from tests

Table 2 displays the comparison of clinical characteristics and Table 3 the comparison of ADLs, by algorithm referral flag; those patients who were flagged for PAC services were at a significantly higher risk for falls, and had more comorbidities (p<0.001). They rated their health status lower (p<0.001), and had more hospitalizations in the past six months (p<0.001). A higher level of decline in ADLs was observed in those flagged for referral (p<0.001) and although there was no significant difference seen in feelings of hopelessness (p=0.93), slightly more patients noted they had lost interest in those activities usually found interesting to them prior to hospitalization, a symptom of depression (p=0.03). There were no significant differences in primary diagnosis classifications with the highest percentage of index hospitalizations due to diseases of the circulatory system in both groups.

### Table 2. Comparison of clinical characteristics by Algorithm Flag (N=1336)

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Flagged (n=895)</th>
<th>Not Flagged (n=441)</th>
<th>p-value¹</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean [SD (Range)] or n (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hospital LOS (days)</td>
<td>3.4 [2.2 (1-20)]</td>
<td>3.1 [2.1 (1-19)]</td>
<td>0.01</td>
</tr>
<tr>
<td>Fall Risk Score [0-125]</td>
<td>42.5 [18.4 (0-110)]</td>
<td>27.5 [11.2 (0-75)]</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Braden Score [6-23]</td>
<td>20.0 [2.1 (8-23)]</td>
<td>21.3 [1.2 (16-23)]</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Number of Comorbidities</td>
<td>3.27 [1.9 (0-14)]</td>
<td>2.7 [1.6 (1-19)]</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Self-Rated Health²</td>
<td></td>
<td></td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Excellent/Good</td>
<td>421(47.0)</td>
<td>320(72.6)</td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td>300(33.5)</td>
<td>95(21.5)</td>
<td></td>
</tr>
<tr>
<td>Fair/Poor</td>
<td>168(18.8)</td>
<td>26(5.9)</td>
<td></td>
</tr>
<tr>
<td>Hospitalized prior 6 mos.²</td>
<td></td>
<td></td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>None</td>
<td>548(61.2)</td>
<td>368(83.5)</td>
<td></td>
</tr>
<tr>
<td>Once</td>
<td>236(26.4)</td>
<td>62(4.1)</td>
<td></td>
</tr>
<tr>
<td>&gt;Once</td>
<td>91(10.2)</td>
<td>8(1.8)</td>
<td></td>
</tr>
<tr>
<td>Depression</td>
<td></td>
<td></td>
<td>0.03</td>
</tr>
<tr>
<td>Lost Interest</td>
<td>860(96.1)</td>
<td>434(98.4)</td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>18(2.0)</td>
<td>2(0.5)</td>
<td></td>
</tr>
</tbody>
</table>

¹p-values calculated using Chi-square, Fisher-Exact or Kruskal-Wallis Tests
²Unknowns excluded from tests
Table 3. Comparison of ADLs by Algorithm Flag (N=1336)

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Flagged (n=895)</th>
<th>Not Flagged (n=441)</th>
<th>p-value(^1)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ADLs:</strong> Ambulation(^2)</td>
<td>No Change</td>
<td>549(61.3)</td>
<td>432(98.0)</td>
</tr>
<tr>
<td></td>
<td>Decline</td>
<td>259(28.9)</td>
<td>0(0)</td>
</tr>
<tr>
<td></td>
<td>Improvement</td>
<td>84(9.4)</td>
<td>9(2.0)</td>
</tr>
<tr>
<td><strong>Bathing(^2)</strong></td>
<td>No Change</td>
<td>664(74.2)</td>
<td>435(98.6)</td>
</tr>
<tr>
<td></td>
<td>Decline</td>
<td>163(18.2)</td>
<td>1(0.2)</td>
</tr>
<tr>
<td></td>
<td>Improvement</td>
<td>61(6.8)</td>
<td>5(1.1)</td>
</tr>
<tr>
<td><strong>Eating(^3)</strong></td>
<td>No Change</td>
<td>820(91.6)</td>
<td>440(99.8)</td>
</tr>
<tr>
<td></td>
<td>Decline</td>
<td>44(4.9)</td>
<td>0(0)</td>
</tr>
<tr>
<td></td>
<td>Improvement</td>
<td>27(3.0)</td>
<td>1(0.2)</td>
</tr>
<tr>
<td><strong>Toileting(^2)</strong></td>
<td>No Change</td>
<td>607(67.8)</td>
<td>435(98.6)</td>
</tr>
<tr>
<td></td>
<td>Decline</td>
<td>210(23.5)</td>
<td>0(0)</td>
</tr>
<tr>
<td></td>
<td>Improvement</td>
<td>72(8.0)</td>
<td>4(0.9)</td>
</tr>
<tr>
<td><strong>Transfer(^2)</strong></td>
<td>No Change</td>
<td>569(63.6)</td>
<td>440(99.8)</td>
</tr>
<tr>
<td></td>
<td>Decline</td>
<td>240(26.8)</td>
<td>0(0)</td>
</tr>
</tbody>
</table>

\(^1\)p-values calculated using Chi-square, Fisher-Exact or Kruskal-Wallis Tests

\(^2\)Unknowns excluded from tests

Acute Healthcare Utilization

Hypothesis two was supported: In patients who were discharged home without services, those flagged by the algorithm had a significantly higher rate of acute healthcare utilization compared to those not flagged for PAC services.

IP Readmission: In the 1,366 sample, 10.3% (141/1336) of patients whom the hospital discharged home without services were readmitted to the hospital within 30-days of discharge. Of these, 77.3% (109/141) were flagged as needing a referral, and ultimately sustained a 12.2% 30-day IP readmission rate; 22.7% (32/141) were not flagged by the algorithm with a significantly lower readmission rate of 7.3% (p=0.006) (Figure 2). Those patients whom the hospital discharged home without services but were flagged by the algorithm for PAC services, had a 67.8% higher risk of having a 30-day IP readmission than those who were not flagged (RR=1.68, 95% CI=1.15-2.45, p=0.006). Within each patient characteristic, there were no statistically significant differences in relative risk for readmission across strata between those flagged and not flagged for PAC services referral.

![Kaplan-Meier IP Readmission Rate by Algorithm Referral](image)
**Discussion**

Our data show differences in socio-demographic and clinical factors between patients who were flagged versus those not flagged by the algorithm for PAC referral. The majority of patients who were flagged for PAC services were recommended for facility-level care, suggesting that they are amongst those most at-risk for poor outcomes, yet they went home without services. These patients had more overall limitations as they were on average six years older, had poorer self-rated health, and greater chronic illness burden as evidenced by more comorbidities, prior hospitalizations, higher fall risk and greater decline in ADLs during the index hospitalization. Furthermore, they were more often single or widowed which may leave them more vulnerable than those who live with others who can surveil changes in health status. More females were flagged by the algorithm for PAC services but did not receive them; this may be related to discharging clinicians perceiving women as more capable of providing their own caretaking, although older females are more likely to use HHC services than males, due almost entirely to greater health needs. Studies have demonstrated that there is great variability in clinician risk tolerance and decision-making for PAC referrals especially given the lack of standardized DP processes.  

Despite algorithm flag, the hospital did not refer patients who were more frequently unemployed or retired which may be due to insurance type/coverage. Insurance status is an important predictor of hospital disposition and PAC services, with uninsured patients less likely to have full access to services. It is also conceivable that patients refused services; Topaz and colleagues found that 28% of patients refuse PAC services but have twice-higher odds of 30- and 60-day readmissions, compared with those that accept PAC referral.  

When the hospital and algorithm agreed on no need for PAC services, patients were more often well-educated and therefore may be perceived as better able to understand discharge instructions. Moreover, males were not flagged by the algorithm as frequently as their female counterparts; this may be confounded due to 64% of males in the sample being married as compared to females. In a chart review of patients who refused PAC services, the most frequent reason cited by males was that their spouse would serve as caregiver.
Acute healthcare utilization rates were significantly higher for those patients who were flagged by the algorithm but discharged home without services. Those in need of PAC services per the algorithm, had a significantly greater risk of either IP readmissions or ED/OBS visits within 30-days of hospital discharge. Since flagged patients who went home without services were readmitted at a rate of approximately five percentage points higher than those not flagged, there is considerable potential for cost-savings. The average cost of a hospital readmission is $13,800 for Medicare beneficiaries aged 65 and older, which is 5 percent higher than for an index admission in the same population. Potentially preventable readmissions may be avoided by aligning PAC services for those most at risk.

There were no differences, in risk of IP readmission or ED/OBS visit by strata within characteristics, demonstrating the complexity of risk prediction for post-discharge acute healthcare utilization. An intricate interaction of patient factors contribute to these risk scores emphasizing the importance of CDS tools to aid clinicians making DP decisions.

Study limitations include data acquisition from one health system in a suburban region yielding a sample that was homogenous with respect to race/ethnicity. As with any observational study, no conclusions regarding causality are possible; also there is always concern about unobserved bias (no instrumental variable was available to mitigate this bias). There are also limitations in examining same-hospital readmissions which may potentially underestimate rates although this is less likely given that neither hospital within this health system are major referral centers. There are frequently challenges in using clinical data extracted from the EHR including concerns regarding the accuracy of data output, much effort was spent during the parent study to reduce this risk. Lastly, we were unable to measure the extent to which patients refused PAC services rather than hospitals not making referrals.

Conclusion
These results indicate that the algorithm is identifying patients who are likely to have poor discharge outcomes and that there is the potential to improve outcomes by referring them for appropriate PAC services. As observed in this analysis, the mean index admission LOS of three days affords healthcare team members a narrow window in which to assess PAC referral and for discharge planners to actualize these plans. Since the algorithm provides decision support soon after admission, there is an opportunity to optimize these DP processes. Future intervention studies are warranted to examine the effect of sharing algorithm advice with clinicians on PAC referral rates, and acute healthcare utilization in the early post-discharge period.

References
Exploiting Unlabeled Texts with Clustering-based Instance Selection for Medical Relation Classification

Youngjun Kim, MS¹, Ellen Riloff, PhD¹, Stéphane M. Meystre, MD, PhD²
¹School of Computing, University of Utah, Salt Lake City, UT, USA;
²Medical University of South Carolina, Charleston, SC

Abstract

Classifying relations between pairs of medical concepts in clinical texts is a crucial task to acquire empirical evidence relevant to patient care. Due to limited labeled data and extremely unbalanced class distributions, medical relation classification systems struggle to achieve good performance on less common relation types, which capture valuable information that is important to identify. Our research aims to improve relation classification using weakly supervised learning. We present two clustering-based instance selection methods that acquire a diverse and balanced set of additional training instances from unlabeled data. The first method selects one representative instance from each cluster containing only unlabeled data. The second method selects a counterpart for each training instance using clusters containing both labeled and unlabeled data. These new instance selection methods for weakly supervised learning achieve substantial recall gains for the minority relation classes compared to supervised learning, while yielding comparable performance on the majority relation classes.

Introduction

Electronic health record (EHR) systems are becoming more prevalent in the U.S.¹ This growth has resulted in very large quantities of clinical patient data becoming available in electronic format, which holds tremendous potential for benefitting clinical research, quality improvement, and surveillance. A substantial proportion of patient-specific information in the EHR is only found in narrative, unstructured clinical notes.² Natural Language Processing (NLP) enables fast, scalable, and accurate extraction of structured and coded information from these clinical notes.³ As part of this information extraction task, identifying semantic relations between concepts is essential to provide accurate and complete information about the concepts and their meaning. For example, extracting relations between mentions of a medication and mentions of allergy symptoms enables differentiation between situations when a medication causes the symptoms and situations when a medication is prescribed to alleviate symptoms.

Given a pair of medical concepts found in a sentence, a relation classification system must determine the type of relation that exists between the two concepts. Our research focuses on the relation classification task introduced in 2010 for the i2b2 Challenge Shared Tasks⁴. This task involves recognizing eight types of relations between pairs of three types of medical concepts: problems, treatments, and tests.

A key challenge of this task is the extremely skewed class distribution across relation types. For example, five types of relations are defined between problems and treatments plus a no relation category (None), but two of these categories (None and TrAP (treatment administered for problem)) account for 86% of the instances in the i2b2 Challenge data. Four relation types (TrCP (treatment causes problem), TrIP (treatment improves problem), TrWP (treatment worsens problem), and TrNAP (treatment not administered because of problem)) are distributed across the remaining 14% of the data. Each of these “minority” relations appears in just 2-6% of the data. Identifying these minority relations is extremely important from a practical perspective because they hold valuable information. For example, the dominant relations between problems and treatments are TrAP (administration of treatment) and None (no relation at all). In contrast, the minority relations (TrCP, TrIP, TrWP, TrNAP) represent situations where a treatment causes, improves, worsens, or is contraindicated for a problem, so they are arguably more important types of situations to recognize.

The most successful methods used for relation classification include various supervised machine learning algorithms.⁴ Extremely skewed class distributions pose substantial challenges for supervised machine learning (ML) because only a small number of labeled examples are available for training. As a result, ML classifiers can achieve high accuracy for the dominant classes but often perform poorly with the minority classes. Manually annotating more data is not a viable solution because of the high cost of manual annotation by medical experts. Also, because the minority classes are relatively rare, each batch of new annotations would provide only a relatively small number of new examples. There is substantial cost for low reward.
Our research aims to improve relation classification in clinical texts with an emphasis on minority classes by exploiting large amounts of unlabeled clinical texts, which are readily available in abundant quantity. We present two new methods to selectively choose unlabeled instances for self-training in an iterative weakly supervised learning framework. Both methods apply a clustering algorithm to group instances into clusters based on similarity measures. The first method, called Unlabeled Data Prototypes (UDP) Selection, uses clusters containing only unlabeled instances and identifies one “prototype” instance from each cluster to use as additional training data. Intuitively, this method aims to identify the different types of instances that occur in the unlabeled data and selects a representative subset of them. The second method, called Labeled Data Counterparts (LDC) Selection, uses clusters containing both labeled and unlabeled data. For each labeled instance, this method identifies its closest counterpart in the cluster by selecting the unlabeled instance that is most similar to it. Intuitively, this method is designed to acquire a new set of training instances that mimic both the class distribution and semantic content (based on feature similarity) of the original training instances. Our experimental results show that these two instance selection methods produce classifiers that can identify minority relation classes more often and more accurately than traditional supervised learning or self-training.

**Background**

The relation classification task was defined as part of the Fourth i2b2/VA Shared Task Challenge in 2010. Our research involves relation classification for pairs of medical concepts, assuming that the terms corresponding to the two concepts have already been identified. The task is to identify how medical problems relate to treatments, tests, and other medical problems in clinical texts. Many sentences contain multiple pairs of concepts, so the challenge includes identifying which pairs are related, as well as identifying the specific type of relation. This task has been typically cast as a supervised learning problem, where a classifier is trained with manually annotated data. Rink et al. used supervised learning to produce the highest micro-averaged $F_1$ score, 73.7%, for this relation extraction task. Their system utilized external resources including Wikipedia, WordNet, and the General Inquirer lexicon as part of their feature set. To improve recall, they set much lower weights to the pairs of non-related concepts (i.e., negative examples) when training an SVM (Support Vector Machine) classifier.

Previous work has presented micro-averaged $F_1$ scores, which assess performance over all of the positive instances regardless of which class they belong. However, micro-averaging obscures performance differences across the classes. For example, it is often possible for a system to achieve a high micro-averaged $F_1$ score by performing well on the majority class but recognizing few, if any, instances of the minority classes. Our research aims to shed light on the performance differences across relation classes, with the goal of comparing the ability of different methods to recognize the minority classes. So we will present macro-averaged $F_1$ scores in the rest of this manuscript.

The Rink et al. system reached macro-averaged scores of 51.7% recall, 55.8% precision, and 53.7% $F_1$ score (not officially reported in Rink et al. but calculated by taking the average of the reported recall and precision of the different sub-classes). de Bruijn et al. explored effective features also applicable to other clinical NLP tasks. In addition to supervised classification, they applied self-training on the provided unlabeled data. Their approach yielded a 73.1% micro-averaged $F_1$ score. The macro-averaged scores for their submission reached 43.7% recall, 66.8% precision, and 51.2% $F_1$ score. These results were calculated by the authors of this manuscript based on the output of de Bruijn et al.’s system. Their subsequent research using composite-kernel learning improved the accuracy of relation classification with a higher micro-averaged $F_1$ score of 74.2%. As an effort to overcome the class imbalance problem, they used down-sampling of negative examples before training the models. D’Souza and Ng presented an ensemble approach exploiting human-supplied knowledge to set up individual classifiers. Their weighted-voting system outperformed a single classifier using the full set of features exploited by different members. Their best-scoring ensemble system produced 69.6% micro-averaged $F_1$ score. Note that their result is not directly comparable with the works described above because of different training data sizes.

In the biomedical and clinical domains, annotating data is especially expensive because of the need for domain experts. Consequently, most systems are trained with relatively small amounts of labeled text, even though much larger amounts of unlabeled text are readily available. Weakly supervised learning, also called semi-supervised learning, has been shown to benefit from training on both labeled and unlabeled data for other NLP tasks, including document classification, named entity recognition, and noun phrase chunking. As a general framework, starting with a small set of initial labeled data, the learner outputs entities from unlabeled data with assigned entity types. Then, the detected entities are collected as new training instances for subsequent iterations. Iterative bootstrapping methods that use seeding heuristics to produce an initial set of training instances have also been a popular choice. Theilen and Riloff showed that semantic lexicons could be learned with extraction patterns from unlabeled texts by...
bootstrapping algorithms. Rosenberg et al.\textsuperscript{14} used self-training to build object detection models and they pointed out that the choice of the initial seeds has a large effect on performance.

There has also been previous work on the relation classification task exploiting unlabeled data. Zhang\textsuperscript{15} proposed a bootstrapping algorithm using random feature projection. Multiple classifiers were trained with randomly selected features from labeled data and they voted to assign labels to the unlabeled data. Mintz et al.\textsuperscript{16} used Freebase\textsuperscript{17}, a large knowledge database, to train a learner with “distant” supervision. Sun et al.\textsuperscript{18} presented a weakly supervised learning method with large-scale word clustering. They augmented the features derived from the word clusters to compensate for the absence of lexical features in labeled data. Related to medical relations, Wang and Fan\textsuperscript{19} collected training data using a clustering algorithm. To minimize the manual annotations, the most representative instances with the highest average similarity to other members of each cluster were chosen for annotation.

**Materials and Methods**

*Labeled Data Description*

We used the i2b2/VA 2010 Shared Task corpus for our research, which consists of a training set of 349 annotated clinical notes and a test set of 477 annotated clinical notes. This test set contains 45,009 annotated medical concepts with 9,069 relations that occur in the same sentence. Relations were defined as follows\textsuperscript{20}:

Medical problem—treatment (Pr-Tr) relations:
- Treatment *improves* medical problem (TrIP).
- Treatment *worsens* medical problem (TrWP).
- Treatment *causes* medical problem (TrCP).
- Treatment is *administered* for medical problem (TrAP).
- Treatment is *not administered* because of medical problem (TrNAP).
- Relation that does not fit into one of the above defined relationships (NoneTrP).

Medical problem—test (Pr-Te) relations:
- Test *reveals* medical problem (TeRP).
- Test *conducted to investigate* medical problem (TeCP).
- Relation that does not fit into one of the above defined relationships (NoneTeP).

Medical problem—medical problem (Pr-Pr) relations:
- Medical problem *indicates* medical problem (PIP).
- Relation that does not fit into PIP relationship (NonePP).

The test set contains 6,949 Pr-Tr pairs that occur in the same sentence, of which 3,463 are positive examples (participate in a relation) and 3,486 are negative examples (NoneTrP). Pr-Te relations include 3,620 positive and 2,452 negative examples (NoneTeP). Pr-Pr relations include 1,986 positive and 11,190 negative examples (NonePP). As seen in Figure 1, the class distributions across Pr-Tr and Pr-Te relation types are extremely skewed.

![Figure 1. Distribution of treatment (Pr-Tr) and test (Pr-Te) relation types in the test set](Image)
Among Pr-Tr relations, four “minority” classes, TrCP, TrIP, TrWP, TrNAP, are distributed across 14% of the data. Each of these relations appears in just ~2-6% of the data. Among the Pr-Te relations, TeCP is the minority class, accounting for < 10% of the instances. Our goal is to improve relation classification with an emphasis on these minority classes by exploiting large amounts of unlabeled clinical texts. Since there is only one type of Pr-Pr relation (PIP), we focused exclusively on the Pr-Tr and Pr-Te relations in our efforts.

Unlabeled Data Preparation for Weakly Supervised Learning

For this research, we also used texts from the MIMIC II Clinical Database, which contains various types of clinical notes: discharge summaries, nursing progress notes, cardiac catheterization notes, ECG reports, radiology reports, and echocardiography reports. From this data set, we used 26,485 discharge summaries after filtering out notes with insufficient text content (<500 Bytes).

For weakly supervised learning preparation, we had to identify the medical concepts in our unlabeled data and classify the assertion of each medical problem concept. Assertion classification aims to determine the assertions of the problem concepts by assigning one of six categories: present, absent, hypothetical, possible, conditional, or not associated with the patient. For concept extraction, we used our previous system consisting of a stacked learning ensemble. We slightly modified the feature set of the individual classifiers by adding skip-grams and word embedding features. For assertion classification of medical problems, we also added new word embedding features to our assertion classifier and retrained the SVM model. As computed by the i2b2 Challenge evaluation script (class exact match), our stacked ensemble achieved 84.4% recall, 89.1% precision, and 86.7% F1 score for concept extraction on the i2b2 test set. The assertion classifier reached 94.5% micro-averaged F1 score. Using the predicted concepts assigned to the unlabeled data, we created a large set of relation pairs to generate feature vectors for weakly supervised learning and clustering.

We used CLUTO, a data clustering software that has been widely used in various tasks, to create clusters containing both labeled (i2b2 training) and unlabeled data: 517,689 pairs of Pr-Tr relations and 455,272 pairs of Pr-Te relations. The same feature vectors generated for SVM classification were re-used with the clustering algorithm. To determine the number of clusters, we use the root-mean-square standard deviation (RMSSD). RMSSD is a measure of homogeneity within clusters and large RMSSD values indicate that clusters are not homogeneous. We ran a series of clustering processes with different numbers of clusters, K, and calculated the RMSSD for each K. We tried 20 different cluster sizes aimed at having the average number of members per cluster ranging from 40 to 800 (i.e. K = the number of instances × n, n = 1/800, 2/800,…, 19/800, 20/800). When we set n to 1/800 and 20/800 (= 1/40), we expected that on average 800 and 40 members would exist in each cluster, respectively. For each of the Pr-Tr and Pr-Te, we then detected the shift point (also known as the “Knee” point) of its RMSSD curve based on the Satopää et al. method. The cluster sizes of 4,529 and 3,414 were identified as the Knee points for the Pr-Tr and Pr-Te relation clusters respectively. In the following paragraphs, we will describe our supervised classification models and then present the instance selection methods based on clustering unlabeled data.

Supervised Relation Classification

We created three supervised learning classifiers (one for each category of concept pairs: Pr-Tr, Pr-Te, and Pr-Pr) using a rich set of features. We applied the Stanford CoreNLP tool for tokenization, lemmatization, part-of-speech (POS) tagging, and phrase chunking. We trained Support Vector Machine (SVM) classifiers with a linear kernel using the LIBLINEAR (Library for Large Linear Classification) software package. The multi-class SVM classifiers use five types of features associated with a pair of concepts <C1, C2>:

- **Assertion**: For each medical problem concept, we create a feature for the assigned assertion type. Assertion categories are considered in a pre-defined order of precedence (e.g., Possible takes precedence over Absent.)

- **Context**: To compensate for the absence of assertions for treatment and test concepts, we incorporated the ConText algorithm at the sentence level to detect three types of contextual properties for each concept: negation, hypothetical, and historical. We also created a second set of ConText algorithm properties restricted to the six-word context window around C1 and C2 (three words on the left of C1 and three words on the right of C2).

- **Distance**: We created several features to represent the distance between concepts C1 and C2 by counting the number of words, concepts, and phrases (e.g., noun phrases and adjective phrases) between them. The number of concepts appearing before or after the pair was also measured. These features were designed to
help the classifiers distinguish between concept pairs that probably have a relation and distant pairs that probably have no relation between them.

- **Lexical:** Lexical features have been very effective for many NLP tasks. We create lexical features for the words contained in $C_1$ and $C_2$, the head words of $C_1$ and $C_2$, two preceding and two following words of $C_1$ and $C_2$, and the words between the two concepts. Also, we defined features for verbs that precede, follow, or occur between the concepts.

- **Syntactic:** POS tags of two words on the right of $C_1$ and POS tags of two words on the left of $C_2$.

- **Word Embedding:** We used the Word2Vec software to perform $K$-means clustering on the word embeddings. We created 1,000 clusters of semantically related words within the unlabeled data (i.e., MIMIC II Clinical Database) with default parameters. Then, we used the cluster identifier of each word between the two concepts as a feature. We also used the cosine similarity of the word embedding vectors for the heads of $C_1$ and $C_2$.

We randomly selected 200 documents from the training set for development purposes. We tuned LIBLINEAR’s parameters to maximize the micro-averaged F1 score with the held-out development data. After experimenting with different values on the development data, we set the cost parameter $c$ to 0.06 for Pr-Tr, and 0.02 for Pr-Te and Pr-Pr. Also, the weights of negative examples were set to 0.2 for Pr-Tr and Pr-Te and 0.3 for Pr-Pr. The lower the weight for instances with no relation, the higher recall was obtained on held-out data.

Although the classifiers showed good performance under the micro-averaged scoring metrics, performance on the minority classes was weak. As shown earlier, the class distributions are extremely skewed and the minority classes are relatively rare. To reduce the performance gap between the dominant classes and the minority classes, we also experimented with retraining the model by assigning higher weights to the minority classes to increase the importance of minority classes being classified correctly. It did not yield an increase in macro-averaged F1 score and more detailed results will be reported in the results section. To improve performance across the different relation classes, we extend our methods to weakly supervised learning described in the following paragraphs.

**Exploiting Unlabeled Data for Relation Classification**

To take advantage of the large amounts of unlabeled clinical notes that are available, we explored an iterative weakly supervised learning framework. We developed two novel methods for instance selection that are specifically aimed at improving performance on minority classes. Our general framework involves the following steps: (1) a classifier is trained with supervised learning using the labeled training data, (2) the classifier is applied to the unlabeled data so that each unlabeled instance receives a predicted label, (3) a subset of the unlabeled instances is selected and then added to the set of labeled data (using the classifier’s predictions as the labels), and (4) the classifier is retrained using the (larger) set of labeled data. This process repeats until a stopping criterion is met (e.g., for a fixed number of iterations or until no new instances can be labeled).

This paradigm is generally known as self-training, where the most common method for instance selection (step 3) sorts the instances based on the confidence scores produced by the classifier (i.e., confidence in the predicted labels) and then selects the most confidently labeled instances. This traditional self-training approach, however, tends to select instances of the dominant classes much more often than the minority classes because the classifier is more confident in its predictions for the dominant classes.

This issue motivated us to explore new methods for instance selection that try to create a diverse and representative set of new instances from the unlabeled data. Consequently, we developed two new methods for instance selection that first cluster the unlabeled data to identify groups of similar instances. Both methods generate clusters and assign labels to the instances in the same way. First, the labeled and unlabeled instances are combined into a single dataset and the clustering algorithm (described previously) is applied. Once the classifier predicts the label of each unlabeled instance, we consider the instances with a high confidence score as candidates for selection. In each iteration, we sort the instances based on the confidence scores produced by the classifier and an instance can be selected when it is ranked in the top 25% per class.

The first instance selection method, called Unlabeled Data Prototypes (UDP) Selection, selects instances from clusters containing only unlabeled data. We compute the purity of each cluster and identify clusters where the highly confident cluster members have the same positive relation type (i.e., cluster purity $= 1$). We discard clusters with purity $< 1$ because the instances are similar but the classifier’s predictions are inconsistent, so the predictions are
suspect. The most representative instance from each cluster is then selected as additional training data, based on average cosine similarity with other cluster members. We assumed that instances selected from these clusters are different from the training instances and they are also dissimilar to each other. Therefore, they could represent some new type of information found in the unlabeled data. This method is illustrated in Figure 2(a). Green-colored instances represent unlabeled data.

Assuming that unlabeled data will be similar to labeled data when they co-exist in the same cluster, our second method, called Labeled Data Counterparts (LDC) Selection, selects instances from the clusters containing both labeled and unlabeled instances. For each instance labeled with a positive relation type, the unlabeled instance most similar to it in the same cluster is selected. Our intuition is that this approach will acquire new training instances that share features with the original labeled data and maintain the same class distribution. This method is illustrated in Figure 2(b). Red-colored instances represent labeled data and green-colored instances represent unlabeled data. In the next sections, we compare the performance of self-training with confidence-based instance selection against our new UDP and LDC instance selection methods.

Results

We have conducted an extensive set of experiments to evaluate the performance of supervised classifiers and weakly supervised learning with different instance selection methods. We evaluated performance with relation data from the i2b2 Challenge test set. We used the official i2b2 Challenge evaluation script to calculate micro-averaged measures. For macro-averaged measures, we created a new script to obtain average values for each relation type. The macro-averaged $F_1$ score is the harmonic mean of the macro-averaged recall and precision.

Supervised Learning Results

Table 1 shows the results produced with the supervised classifiers, which were trained to optimize for micro-averaged measures. This baseline supervised learning system was trained with the i2b2 training data and achieved micro-averaged scores of 74.9% recall, 73.7% precision, and 74.3% $F_1$ score.

Table 1. Results produced with the supervised classifier.

<table>
<thead>
<tr>
<th>Relation type</th>
<th>Recall</th>
<th>Precision</th>
<th>$F_1$ score</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALL</td>
<td>74.9</td>
<td>73.7</td>
<td>74.3</td>
</tr>
<tr>
<td>Treatment</td>
<td>67.4</td>
<td>68.9</td>
<td>68.2</td>
</tr>
<tr>
<td>TrIP</td>
<td>31.8</td>
<td>63.6</td>
<td>42.4</td>
</tr>
<tr>
<td>TrWP</td>
<td>4.2</td>
<td>42.9</td>
<td>7.6</td>
</tr>
<tr>
<td>TrCP</td>
<td>52.3</td>
<td>59.5</td>
<td>55.6</td>
</tr>
<tr>
<td>TrAP</td>
<td>79.9</td>
<td>71.2</td>
<td>75.3</td>
</tr>
<tr>
<td>TrNAP</td>
<td>25.1</td>
<td>49.5</td>
<td>33.3</td>
</tr>
<tr>
<td>Test</td>
<td>82.9</td>
<td>81.5</td>
<td>82.2</td>
</tr>
<tr>
<td>TeRP</td>
<td>90.3</td>
<td>82.7</td>
<td>86.3</td>
</tr>
<tr>
<td>TeCP</td>
<td>45.1</td>
<td>71.4</td>
<td>55.3</td>
</tr>
<tr>
<td>PIP</td>
<td>73.2</td>
<td>67.9</td>
<td>70.4</td>
</tr>
</tbody>
</table>
Although our supervised classifiers achieve overall performance comparable to state-of-the-art relation classification systems, performance on the minority classes lags far behind the dominant classes. The F₁ score of \( TrWP \) was only 7.6% with a recall of 4.2%. Most of the \( TrWP \) instances were misclassified because of the very low prevalence of their cases. For example, a \( TrWP \) case from the test set, “She has a known diagnosis of myelodysplasia that has become recalcitrant to Procrit”, the medical problem ‘myelodysplasia’, the treatment ‘Procrit’, and possibly a keyword ‘recalcitrant’ never appeared in the training data. Based on macro-averaging, this system reached 50.2% recall, 63.6% precision, and 56.1% F₁ score.

We also experimented with decreasing the weights of negative examples to help increase recall on minority classes. This did not yield an increase in macro-averaged F₁ score because of a substantial drop in precision. Adjusting the importance of different relation types by assigning different weights also did not affect performance very much.

Comparing Supervised Learning and Weakly Supervised Learning Results

We evaluated the performance of self-training with traditional confidence-based instance selection (called Self-training below), and instance selection with our new UDP and LDC methods. We ran all of the weakly supervised learning methods for 20 iterations.

For self-training, we only selected positive examples (pairs of concepts with relations) from the unlabeled data to augment the labeled data. For each iteration, we added \( K \) newly labeled examples, where \( K \) is the number of positive examples in the original training data. Our intention was to be conservative in adding new examples and maintain the importance of labeled data. To keep the class distribution of the labeled data, we imposed that the number of newly labeled examples for each positive class should not exceed the number of examples in the original training data.

Table 2 shows results for each class and macro-averaged F₁ scores for the \( Pr-Tr \) and \( Pr-Te \) relations. For each relation type, the best results appear in boldface. We used paired t-tests to measure statistical significance. Results that are significantly different from the supervised learning results at the 95% significance level are preceded by an asterisk (*). Self-training with confidence-based instance selection produced the best F₁ score on \( TrCP \) and \( TrNAP \) classes. For \( TrWP \) and \( TeCP \), self-training’s performance was significantly different than supervised learning.

<table>
<thead>
<tr>
<th>Relation type</th>
<th>Supervised</th>
<th>Self-training</th>
<th>UDP</th>
<th>LDC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment</td>
<td>46.2</td>
<td>48.0</td>
<td>48.9</td>
<td>49.7</td>
</tr>
<tr>
<td>TrIP</td>
<td>42.4</td>
<td>46.0</td>
<td>*49.3</td>
<td>*47.4</td>
</tr>
<tr>
<td>TrWP</td>
<td>7.6</td>
<td>*16.3</td>
<td>12.3</td>
<td>*19.2</td>
</tr>
<tr>
<td>TrCP</td>
<td>55.6</td>
<td>56.8</td>
<td>55.5</td>
<td>53.1</td>
</tr>
<tr>
<td>TrAP</td>
<td>75.3</td>
<td>75.4</td>
<td>75.8</td>
<td>75.8</td>
</tr>
<tr>
<td>TrNAP</td>
<td>33.3</td>
<td>35.4</td>
<td>33.1</td>
<td>33.6</td>
</tr>
<tr>
<td>Test</td>
<td>72.0</td>
<td>72.6</td>
<td>72.8</td>
<td>73.1</td>
</tr>
<tr>
<td>TeRP</td>
<td>86.3</td>
<td>86.3</td>
<td>86.3</td>
<td>*86.7</td>
</tr>
<tr>
<td>TeCP</td>
<td>55.3</td>
<td>*58.5</td>
<td>*59.2</td>
<td>*59.5</td>
</tr>
</tbody>
</table>

Both the UDP and LDC instance selection methods produced higher macro-averaged F₁ scores than Self-training. The UDP method (third column of Table 2) produced the best F₁ score of 49.3% on the TrIP class. The F₁ scores for TrIP and TeCP were significantly higher than for supervised learning. The LDC method (last column of Table 2) produced the highest F₁ scores on most of the relation classes. It obtained the best macro-averaged F₁ scores for Treatment and Test. For TrIP, TrWP, TeRP, and TeCP, the performance of LDC method was significantly better than supervised learning.

Finally, we tried to combine the UDP and LDC methods. New instances were selected separately by the UDP and LDC methods and then the combined set of instances was added to the labeled data. However, this system produced an F₁ score of 58.3%, so did not outperform the LDC method on its own.
In another set of experiments, we performed ablation testing of the supervised learning system to evaluate the impact of each feature set based on micro-averaged and macro-averaged scores, separately. If some features have more impact for macro-averaged scores than micro-averaged scores, then our hypothesis is that they are especially important features for minority classes. The row header in Table 3 specifies the feature set that has been ablated. The columns named “Impact” in Table 3 present the F1 score difference between the ablated classifier and the complete system. Every feature set contributed to the performance of the supervised classifiers except that syntactic features did not increase macro-averaged F1 score. The macro-averaged F1 score dropped the most when the lexical features were removed. This suggests that exploiting unlabeled data could be especially beneficial for the minority classes by bringing in new lexical features. The F1 scores of TrIP, TrNAP, and TeCP decreased from 42.4%, 33.3%, and 55.3% to 29.4%, 21.9%, and 42.4% respectively without the lexical features.

**Table 3. Features Contribution**

<table>
<thead>
<tr>
<th>Feature</th>
<th>Macro-averaged</th>
<th>Micro-averaged</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>F1 score</td>
<td>Impact</td>
</tr>
<tr>
<td>All</td>
<td>56.1</td>
<td>-</td>
</tr>
<tr>
<td>- Assertion</td>
<td>55.4</td>
<td>-0.7</td>
</tr>
<tr>
<td>- Contextual</td>
<td>55.4</td>
<td>-0.7</td>
</tr>
<tr>
<td>- Distance</td>
<td>55.2</td>
<td>-0.9</td>
</tr>
<tr>
<td>- Lexical</td>
<td>49.0</td>
<td>-7.1</td>
</tr>
<tr>
<td>- Syntactic</td>
<td>56.6</td>
<td>0.5</td>
</tr>
<tr>
<td>- Word embedding</td>
<td>55.8</td>
<td>-0.3</td>
</tr>
</tbody>
</table>

**Analysis**

We carried out an empirical analysis of self-training with confidence-based instance selection to better understand its limitations. After clustering the unlabeled data, we counted the number of instances selected from each cluster during the first iteration. We found that most instances were selected from a small subset of the clusters: about 10% of the clusters provided over 78% of the newly selected unlabeled instances. This shows that selecting instances based only on confidence scores tends to yield a relatively homogenous set of new instances that is low in diversity.

Table 4 displays the Recall, Precision, and F1 results of LDC instance selection along with the total counts of true positives (TP) and the number and percentage of true positive gains (compared to supervised learning) in the rightmost column. The numbers in parentheses in the Recall, Precision, and F1 columns indicate the difference between the supervised classifier and the LDC method. Results significantly different from supervised learning at the 95% significance level are preceded by an asterisk (*).

**Table 4. Results of LDC with comparison to the supervised learning model**

<table>
<thead>
<tr>
<th>Relation type</th>
<th>Recall</th>
<th>Precision</th>
<th>F1 score</th>
<th>True positive</th>
<th>TP Gain (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minority classes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TrIP</td>
<td>*38.9</td>
<td>(+7.1)</td>
<td>60.6</td>
<td>(*-3.0)</td>
<td>77</td>
</tr>
<tr>
<td>TrWP</td>
<td>*11.9</td>
<td>(+7.7)</td>
<td>50.0</td>
<td>(+7.1)</td>
<td>17</td>
</tr>
<tr>
<td>TrCP</td>
<td>*65.1</td>
<td>(+12.8)</td>
<td>*44.9</td>
<td>(-14.6)</td>
<td>289</td>
</tr>
<tr>
<td>TrNAP</td>
<td>23.6</td>
<td>(-1.6)</td>
<td>*58.4</td>
<td>(+9.0)</td>
<td>45</td>
</tr>
<tr>
<td>TeCP</td>
<td>*57.7</td>
<td>(+12.6)</td>
<td>*61.4</td>
<td>(-10.0)</td>
<td>339</td>
</tr>
<tr>
<td>Majority classes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TrAP</td>
<td>*80.8</td>
<td>(+0.9)</td>
<td>71.3</td>
<td>(+0.2)</td>
<td>2,009</td>
</tr>
<tr>
<td>TeRP</td>
<td>*88.5</td>
<td>(-1.8)</td>
<td>*85.0</td>
<td>(+2.4)</td>
<td>2,682</td>
</tr>
</tbody>
</table>
Table 4 shows that most of the minority classes benefitted substantially from the LDC method. The largest percentage gain came for TRWP where LDC correctly identified 17 instances but the supervised learner only produced six true positives, resulting in a gain of 11 (183.3%). The majority classes also achieved slightly higher F1 scores. The LDC method appears to be an effective way to improve recall on minority relation classes while maintaining good performance on the majority classes.

Conclusion

We showed that clustering-based instance selection from unlabeled text data could improve performance on minority classes for relation type classification between medical concepts. Experimental results show that our clustering-based methods outperformed supervised classification and traditional self-training from unlabeled texts. We believe that this approach offers a more robust solution for classification problems when the data has a highly skewed class distribution, acquiring manual annotations is expensive, but large quantities of unannotated text data are available.

Acknowledgments

The de-identified clinical records used in this research were provided by the i2b2 National Center for Biomedical Computing funded by U54LM008748 and were originally prepared for the NLP Shared Tasks organized by Dr. Özlem Uzuner and colleagues.

References


De-identification of psychiatric notes is a special case of named entity recognition. Supervised machine-learning (ML) algorithms have achieved promising results for this task. However, ML-based de-identification systems often require annotating a large number of clinical notes of interest, which is costly. Domain adaptation (DA) is a technology that enables learning from annotated datasets from different sources, thereby reducing annotation cost required for ML training in the target domain. In this study, we investigate the use of DA methods for de-identification of psychiatric notes. Three state-of-the-art DA methods: instance pruning, instance weighting, and feature augmentation are applied to three source corpora of annotated hospital discharge summaries, outpatient notes, and a mixture of different note types written for diabetic patients. Our results show that DA can increase de-identification performance over the baselines, indicating that it can effectively reduce annotation cost for the target psychiatric notes. Feature augmentation is shown to increase performance the most among the three DA methods. Performance variation among the different types of clinical notes is also observed, showing that a mixture of different types of notes brings the biggest increase in performance.

INTRODUCTION

Clinical narratives contain detailed information of patients and are a valuable data source for clinical research. However, they often contain protected health information (PHI) such as names and addresses, that have the potential risk of revealing the patient’s identity. Therefore, de-identification (removing PHI) of clinical documents is often required. As manual review to remove PHI is time-consuming and costly, significant effort has gone into developing automatic de-identification methods.

Early de-identification systems were often rule-based, using hand-coded rules and specialized dictionaries to identify PHI mentions in clinical notes. More recent systems have adopted supervised machine learning (ML) algorithms, treating de-identification as a token classification or a sequence labeling problem. Various ML algorithms have been studied, including conditional random fields, support vector machines, decision trees, hidden Markov models, and recurrent neural networks. Often, the ML-based systems are augmented with rule-based parts, forming hybrid approaches. Such ML-based de-identification systems have shown high performance, similar to annotation quality done by humans.

Community challenges on de-identification have also been held. In the 2006 i2b2 challenge Task 1, a set of discharge summaries were provided for training and testing de-identification systems. The discharge summaries were annotated with eight PHI categories, which comprised a subset of the eighteen categories listed by HIPAA (the Health Insurance Portability and Accountability Act). The Track 1 of the 2014 i2b2/UTHealth shared task also addressed a de-identification task where PHIs were to be removed in clinical notes of diabetes patients. Unlike the challenge in 2006, a stricter set of PHI categories was used, defining seven main PHI categories and 30 sub-categories. The most recent community challenge on de-identification is the 2016 CEGS N-GRID Track 1, where a set of psychiatric notes were provided for system development. In these challenges, ML-based systems have shown the best performances. We participated in the 2016 challenge, and developed an ML-based hybrid system that achieved second best performance.

However, ML-based systems depend heavily on large amounts of training data that is highly similar to the test data. This is problematic for clinical text, which consists of diverse types of clinical documents with different writing styles across institutions. For instance, a de-identification system trained on discharge summaries may show much lower performance on psychiatric notes, or a system trained on discharge summaries from one hospital may show much lower performance on discharge summaries from a different hospital. To achieve optimal performance, it is
ideal to annotate individual training sets for each type of clinical note. However, manual annotation of clinical text is expensive and complicated\cite{riedel_metaseq_2015}, thus making such extensive annotation approaches unrealistic and prohibitive for wide adoption of ML-based de-identification systems. Therefore, methods that can leverage existing annotated corpora to quickly build a de-identification system for any target domain of interest are highly desirable.

Here we propose to use domain adaptation (DA) technologies to address this problem. Domain adaptation maximizes the use of existing data (source) for the data of interest (target) by learning useful aspects of source data and largely ignoring the source aspects that cannot contribute to the model. Through domain adaptation, ML-based de-identification systems can efficiently leverage existing de-identification corpora from other sources to quickly build high-performance models using less annotated samples from the target domain, thus reducing the annotation cost.

Domain adaptation has been applied for various natural language processing tasks in the biomedical domain. Dahlmeier and Ng\cite{dahlmeier_evaluating_2016} evaluated three domain adaptation methods for semantic role labeling (SRL) task in biomedical literature. A general domain SRL corpus was used as the source, and a biomedical domain SRL corpus was used as the target. Their result shows that domain adaptation can leverage existing general domain SRL resource and greatly improve the performance of SRL in the biomedical domain. Zhang et al.\cite{zhang_domainless_2018} also conducted a similar set of experiments for SRL of clinical notes. An SRL corpus of biomedical literature as well as general domain corpora were used as sources. Their result also shows that domain adaptation can boost the SRL performance through utilization of existing out-of-domain data. Ramesh et al.\cite{ramesh_domain_2018} applied domain adaptation for automatic discourse connective detection in biomedical text. Here, a general domain corpus was used as the source. They also showed that combination of different domain adaptation methods can further improve the performance.

In this study, we propose to investigate how to leverage three existing corpora (source domains) to improve de-identification in psychiatric notes (the target domain). We implement three domain adaptation algorithms: instance pruning, instance weighting, and feature augmentation for this task. The de-identification system that we developed for 2016 challenge is utilized for this study\cite{riedel_metaseq_2015}. Our study shows that domain adaptation can boost the de-identification performance when it is compared with baselines that use only the psychiatric notes or simply combine annotated data from target and source, indicating that it can effectively reduce annotation cost. Moreover, we also find that feature augmentation performed better than the other two DA methods and existing data from different source domains contribute differently to de-identification of psychiatric notes. To the best of our knowledge, this is the first detailed study on domain adaptation for automatic de-identification of clinical text.

METHODS

Datasets

Table 1. Statistics of the source and the target corpora. The sizes of original datasets, as provided by the challenges or the institutes, are compared to the sizes of sampled subsets used in this study.

<table>
<thead>
<tr>
<th>Description</th>
<th>Outpatient notes</th>
<th>Discharge summaries</th>
<th>Mixture notes</th>
<th>Psychiatric notes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Source</td>
<td>Source</td>
<td>Source</td>
<td>Target (training)</td>
</tr>
<tr>
<td># Document in original dataset</td>
<td>325</td>
<td>889</td>
<td>1,304</td>
<td>600</td>
</tr>
<tr>
<td># Documents used in this study</td>
<td>325</td>
<td>604</td>
<td>470</td>
<td>600</td>
</tr>
<tr>
<td># PHI mentions used in this study</td>
<td>10,454</td>
<td>13,193</td>
<td>8,863</td>
<td>11,937</td>
</tr>
<tr>
<td># Tokens used in this study</td>
<td>413,657</td>
<td>413,841</td>
<td>414,944</td>
<td>1,430,390</td>
</tr>
</tbody>
</table>

As target data, a corpus of psychiatric notes provided by the 2016 CEGS N-GRID shared task Track 1\cite{riedel_metaseq_2015} is used (Psychiatric notes). The corpus is divided into a training set, a development set, and a validation set, for training,
parameter optimization, and measurement of final performance, respectively. Three different source corpora are used: 1) discharge summaries from the 2006 i2b2 de-identification challenge Task 1\textsuperscript{17} (Discharge summaries), 2) outpatient notes from the University of Texas Health Science Center at Houston (UTHealth) with manually annotated PHIs (Outpatient notes), 3) diabetes patients’ notes from the 2014 i2b2/UTHealth shared task on de-identification (Track 1)\textsuperscript{15}, consisting of a mixture of different types of clinical notes such as admission notes, emergency visit notes, and discharge summaries (Mixture notes). These three source corpora have different samples sizes (Table 1), which may affect the performance. To compare them fairly, we decided to use equal sample size for each source corpus. As outpatient notes from UTHealth is the smallest corpus, we reduced the sizes of the other two source corpora to contain similar numbers of tokens as that of outpatient corpus, by randomly sampling notes from the original corpora. Table 1 shows the statistic of the four corpora used in this study. As there are slight differences among annotation guidelines for these four corpora, we define a common set of PHIs for all four datasets that consists of eight PHI categories, or Date, Doctor, Phone, Patient, Location, ID, Hospital, and Age.

**ML-based de-identification system**

A modified version of the de-identification system that was developed for the 2016 CEGS N-GRID challenge\textsuperscript{18} is used in this study. The system employs a single conditional random fields (CRF) model. CRF is a standard model widely used in many state-of-the-art de-identification systems\textsuperscript{7,9,21-23}. Our system first pre-processes the notes using the CLAMP tokenizer (http://clamp.uth.edu), OpenNLP POS tagger (http://opennlp.sourceforge.net), and a dictionary-based section parser that uses a dictionary of standard section names in clinical notes (a modified version of CLAMP section parser). After pre-processing, a token-based CRF tagger is employed to identify PHI mentions using the BIO tagging scheme. CRFSuite (http://www.chokkan.org/software/crfsuite/) is used as the implementation of the CRF algorithm. Features shown to be effective in 2016 challenge are employed, excluding those that are applicable only to psychiatric notes, or relevant only to PHI categories not targeted in this work\textsuperscript{25}. Table 2 shows the features.

**Table 2. Features used for our ML-based de-identification system**

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word shape</td>
<td>Orthographic forms of tokens (by substituting uppercase letters, lowercase</td>
</tr>
<tr>
<td></td>
<td>letters, and numbers with ‘A’, ‘a’, and ‘#’, respectively)</td>
</tr>
<tr>
<td>Surface N-grams</td>
<td>Token, POS, word shape N-grams</td>
</tr>
<tr>
<td>Prefix/suffix</td>
<td>Prefix and suffix N-grams</td>
</tr>
<tr>
<td>Token regex</td>
<td>Token-level regular expression matching results</td>
</tr>
<tr>
<td>Sentence info.</td>
<td>Sentence length and shape (i.e., ending with enumeration indicator or not)</td>
</tr>
<tr>
<td>Section info.</td>
<td>Section name</td>
</tr>
<tr>
<td>Dictionary matching</td>
<td>Matching results to a dictionary of frequent PHI terms</td>
</tr>
<tr>
<td>Word representations</td>
<td>Brown clusters\textsuperscript{39}, random indexings\textsuperscript{30},</td>
</tr>
<tr>
<td></td>
<td>word2vec embeddings\textsuperscript{31}, and GloVe embeddings\textsuperscript{32}</td>
</tr>
<tr>
<td>General domain NER</td>
<td>Outputs of Stanford NER\textsuperscript{31}</td>
</tr>
<tr>
<td>Semantic role labeling</td>
<td>Outputs of SENNA semantic role labeling\textsuperscript{34}</td>
</tr>
</tbody>
</table>

**Domain adaptation methods**

Three state-of-the-art domain adaptation methods are implemented. The domain adaptation methods are described below. Note that we focus on supervised domain adaptation methods that require small amount of labeled (annotated) target data for training, as opposed to unsupervised domain adaptation methods that utilize unlabeled target data.

**Instance weighting:** Instance weighting\textsuperscript{36} assigns higher weights to instances from target data than to instances from source data. The weighted instances from both the target and the source are combined into a single training dataset to train a classifier. The assumption is that the data distribution in the source domain does not match well
to the data distribution in the target domain, and the weights guide the ML algorithms to learn more from the target domain than from the source domain.

**Instance pruning:** Instance pruning\(^35\) first trains a classifier using only the labeled target data, and uses the classifier to predict labels for the source data. Then, it selects top \(k\) instances that are wrongly predicted from the source data ranked by the prediction confidence, and removes the \(k\) instances from the source data. The modified source data is then directly combined with the target training data to train the final classifier. The intuition is that the top \(k\) wrongly classified instances are very different from the target data.

**Feature augmentation:** A feature augmentation algorithm named EasyAdapt\(^36\) is employed. The algorithm augments features from the target and the source data with general version features. Formally, given a target feature vector \(X_t\) and a source feature vector \(X_s\), EasyAdapt generates augmented features \(EA(X_t)\) and \(EA(X_s)\) described as follows:

\[
EA(X_t) = <X_s, X_s, 0 > \\
EA(X_s) = <X_t, 0, X_t >
\]

where 0 is a zero vector of length \(|X|\). As a result, three versions of feature sets are produced: general, source-specific, and target-specific. The general features are expected to get higher weights for instances that are common for both target and source. On the other hand, the target-specific or the source-specific features are expected to gain weights for unique instances for target or source.

**Experimental setup**

In order to validate our CRF-based de-identification system, in-domain performance of the system is measured. The original datasets of the three sources (without subset sampling) are used to perform 10-fold cross validation of the system on source domains. In-domain performance on the target psychiatric notes is measured by training the system on the training set and evaluating on the validation set, to provide a baseline before any DA method is applied.

The three domain adaptation algorithms are evaluated on each of the three sources. The algorithms are compared to three baselines: “target only” (TO), “source only” (SO), and “source & target” (S&T). For TO, no source data is used and the CRF is trained on the psychiatric notes training set only. For SO, only the source data is used for training, without any labeled target data. Finally, for S&T, the source data and the target data are simply merged together without applying any domain adaptation method.

In order to determine the extent of savings conferred by the domain adaptation methods in terms of annotations of the target data, performance of domain adaptation methods as well as the TO and the S&T baselines are reported at different amounts of labeled target data – with an increase of 10% of psychiatric training data. Learning curves that plot the amount of labeled target data and the performances of the de-identification systems (F-score) are reported.

For all the experiments, the psychiatric notes training set is used as the labeled target data, the validation set is used for measuring the performance, and the development set is used for parameter selection, including \(k\) for instance pruning (top \(k\) instances to remove), and \(w\) for instance weighting (weight \(w<1\) to be assigned to the source instances).

**Evaluation**

For evaluation, we follow the method used in previous de-identification challenges\(^15,17\). Micro-averaged precision, recall, and F-score are reported. System outputs are considered to be correct only when both the type and the character offsets match the gold standard. Statistical significance is determined by approximate randomization test\(^37\) using \(N=9999\) and \(\alpha=0.1\).

<table>
<thead>
<tr>
<th>Source</th>
<th>Target only</th>
<th>DA method</th>
<th>P</th>
<th>R</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Outpatient notes</td>
<td>Target only</td>
<td></td>
<td>98.11</td>
<td>96.72</td>
<td>97.40</td>
</tr>
<tr>
<td></td>
<td>Source only</td>
<td></td>
<td>98.69</td>
<td>97.46</td>
<td>98.07</td>
</tr>
<tr>
<td>Discharge summaries</td>
<td>Target only</td>
<td></td>
<td>96.84</td>
<td>92.54</td>
<td>94.64</td>
</tr>
<tr>
<td>Mixture notes</td>
<td>Target only</td>
<td></td>
<td>93.17</td>
<td>86.82</td>
<td>89.88</td>
</tr>
<tr>
<td>Psychiatric notes</td>
<td>Target only</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4. Performance with and without domain adaptation using outpatient notes, discharge summaries, and mixture notes. Statistical significance is marked with * when the performance is higher than the target only baseline, and with ◊ when higher than the source & target baseline. Best F-score for each source is marked in bold.

<table>
<thead>
<tr>
<th>Source</th>
<th>DA method</th>
<th>P</th>
<th>R</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Outpatient notes</td>
<td>Target only</td>
<td>93.17</td>
<td>86.82</td>
<td>89.88</td>
</tr>
<tr>
<td></td>
<td>Source only</td>
<td>92.94</td>
<td>43.77</td>
<td>59.51</td>
</tr>
<tr>
<td></td>
<td>Source &amp; target</td>
<td>93.64</td>
<td>85.89</td>
<td>89.60</td>
</tr>
<tr>
<td></td>
<td>Instance pruning</td>
<td>93.49</td>
<td>86.61</td>
<td>89.92◊</td>
</tr>
<tr>
<td></td>
<td>Instance weighting</td>
<td>93.14</td>
<td>86.73</td>
<td>89.82◊</td>
</tr>
<tr>
<td></td>
<td>Feature augmentation</td>
<td>93.30</td>
<td>87.24</td>
<td>90.17*◊</td>
</tr>
<tr>
<td>Discharge summaries</td>
<td>Target only</td>
<td>70.54</td>
<td>56.18</td>
<td>62.54</td>
</tr>
<tr>
<td></td>
<td>Source only</td>
<td>93.22</td>
<td>85.89</td>
<td>89.40</td>
</tr>
<tr>
<td></td>
<td>Source &amp; target</td>
<td>93.18</td>
<td>86.29</td>
<td>89.60</td>
</tr>
<tr>
<td></td>
<td>Instance pruning</td>
<td>92.45</td>
<td>86.24</td>
<td>89.24</td>
</tr>
<tr>
<td></td>
<td>Instance weighting</td>
<td>93.32</td>
<td>87.26</td>
<td>90.19*◊</td>
</tr>
<tr>
<td>Mixture notes</td>
<td>Target only</td>
<td>82.82</td>
<td>74.26</td>
<td>78.31</td>
</tr>
<tr>
<td></td>
<td>Source only</td>
<td>93.37</td>
<td>86.96</td>
<td>90.05</td>
</tr>
<tr>
<td></td>
<td>Source &amp; target</td>
<td>93.32</td>
<td>86.98</td>
<td>90.04</td>
</tr>
<tr>
<td></td>
<td>Instance pruning</td>
<td>92.96</td>
<td>87.63</td>
<td>90.22</td>
</tr>
<tr>
<td></td>
<td>Instance weighting</td>
<td>93.46</td>
<td>87.53</td>
<td>90.40*</td>
</tr>
</tbody>
</table>

Table 5. Performance comparison of feature augmentation (FA) methods that use the three sources to the target only baseline, per each PHI category. F-scores are shown. The amount of performance change by feature augmentation over the target only baseline is shown inside the parentheses.

<table>
<thead>
<tr>
<th>PHI category</th>
<th>Target only</th>
<th>FA w/ outpatient notes</th>
<th>FA w/ discharge summaries</th>
<th>FA w/ mixture notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>96.37</td>
<td>96.61 (+0.24)</td>
<td>96.72 (+0.35)</td>
<td>96.85 (+0.48)</td>
</tr>
<tr>
<td>Date</td>
<td>94.63</td>
<td>94.79 (+0.16)</td>
<td>95.00 (+0.37)</td>
<td>94.89 (+0.26)</td>
</tr>
<tr>
<td>Doctor</td>
<td>80.33</td>
<td>81.10 (+0.77)</td>
<td>80.8 (+0.47)</td>
<td>81.53 (+1.20)</td>
</tr>
<tr>
<td>Hospital</td>
<td>58.82</td>
<td>62.86 (+4.04)</td>
<td>52.94 (-5.88)</td>
<td>64.86 (+6.04)</td>
</tr>
<tr>
<td>ID</td>
<td>83.95</td>
<td>84.20 (+0.25)</td>
<td>84.34 (+0.39)</td>
<td>84.48 (+0.53)</td>
</tr>
<tr>
<td>Location</td>
<td>82.37</td>
<td>82.47 (+0.10)</td>
<td>82.45 (+0.08)</td>
<td>82.47 (+0.10)</td>
</tr>
<tr>
<td>Patients</td>
<td>96.59</td>
<td>96.59 (+0.00)</td>
<td>96.05 (-0.54)</td>
<td>96.59 (+0.00)</td>
</tr>
<tr>
<td>Overall</td>
<td>89.88</td>
<td>90.17 (+0.29)</td>
<td>90.19 (+0.31)</td>
<td>90.40 (+0.52)</td>
</tr>
</tbody>
</table>
Results

In-domain de-identification performance

Table 3 shows the performance of CRF-based de-identification systems when trained and tested on data from the same domain. For outpatient notes, discharge summaries, and mixture notes, the system shows performance comparable to the results from previous studies (e.g., F-score 93.23 for HIPPA-PHI categories by the third best system in the 2014 i2b2 challenge).15

De-identification performance using domain adaptation

Table 4 shows the de-identification performance with domain adaptation methods using the three sources. When trained only on the psychiatric notes training set (i.e., the TO baseline), the system shows an F-score of 89.88. When only the source corpora are used (i.e., the SO baseline), the performance dropped down to 59.51 (outpatient notes). Simply merging source and target (i.e., the S&T baseline) improves the performance in the case of mixture notes (compared to TO), but is worse than TO for outpatient notes and discharge summaries. DA methods have varied performance: but feature augmentation method shows statistically significant performance increase over both the S&T baseline (with outpatient notes or discharge summaries as source) and the TO baseline (for all three sources), achieving best F-scores for all three sources. Overall, the highest F-score 90.40 is achieved when feature augmentation is applied using mixture notes as source.

Table 4 also shows that the three sources exhibit quite different performance, even though they are equal-sized. With the SO baseline, mixture notes shows the highest performance, followed by discharge summaries, and then outpatient notes (statistically significant difference by approximate randomization test with N=9999 and α=0.1). This performance gap among equal-sized sources indicates that mixture notes may be the most similar to the psychiatric notes and outpatient notes may be the most different. The performance difference between outpatient notes and discharge summaries disappears when S&T or feature augmentation domain adaptation is applied. However, mixture notes still has significantly higher performance than outpatient notes or discharge summaries even when S&T or feature augmentation is applied.

Table 5 compares the performance of the TO baseline and feature augmentation for each PHI category. When outpatient notes or mixture notes is used as source, feature augmentation improves performance for all PHI categories. In particular, ID category shows the biggest improvement over the TO baseline (4.04 increase in the F-score with outpatient notes, and 6.04 with mixture notes). Note that ID is one of the sparsest categories in the psychiatric notes training set (0.58% of all the PHI mentions). Discharge summaries shows somewhat different result than outpatient notes and mixture notes. While most of the PHI categories show performance increase, ID and Phone show performance drop. Lastly, note that the F-score for Age category is 0.00 regardless of which domain adaptation method is applied or which source is used. We conjecture that this is due to the severe sparsity of Age PHI mentions even after addition of source data through domain adaptation; Age constitutes less than 0.1% of all PHI mentions in all the four corpora.

Figure 1 compares de-identification performance of feature augmentation with TO and S&T, for the three sources, with increasing amounts of target training data. When outpatient notes or discharge summaries is used as source, S&T outperforms TO when the amount of target training data is smaller (30% to 40% of the whole target training dataset), but S&T shows lower performance than TO as the amount of target training data increases. When mixture notes is used as source, S&T shows higher performance than the TO baseline regardless of the amount of target training data, but the performance of S&T converges to that of TO as the amount of target training data approaches to 100%. Feature augmentation improves performance over both the TO and S&T baselines regardless of the amount of target training data, when outpatient notes or discharge summaries is used as source. Interestingly, with mixture notes as source, S&T shows higher performance than feature augmentation when less than or equal to 60% of the target training data is used. However, feature augmentation starts to show better performance than S&T when more than 60% of target training data is used. According to Figure 1, if we want to achieve an F-score of 89 for the target domain, it will require annotating 80% of training samples when only target domain data is used. However, if we apply feature augmentation to existing annotated sources such as discharge summaries, it only requires approximately 65% of annotated training samples from the target domain, indicating a 18.8% saving on annotation effort.
Figure 1. Learning curves of the de-identification systems for TO, S&T, and feature augmentation (FA) for three sources: outpatient notes, discharge summaries, and mixture notes. X-axis denotes the percentage of labeled target training data, and Y-axis denotes de-identification F-score.
DISCUSSION

In this study, we proposed to use domain adaptation techniques to maximize the use of existing annotated datasets for de-identification of clinical notes. Psychiatric notes, an important but understudied type of notes, are used as the target. Three existing corpora, each consisting of discharge summaries, outpatient notes, and a mixture of various types of notes written for diabetes patients, are used as sources. Three state-of-the-art domain adaptation methods, instance weighting, instance pruning, and feature augmentation, are tested. It is shown that feature augmentation can increase de-identification F-score over the TO and the S&T baselines, indicating the potential of reducing annotation cost for building ML-based de-identification systems. We also find that the performance increase by domain adaptation could depend on which type of clinical notes is used as source.

In previous work that employs domain adaptation for natural language processing (NLP) tasks in the biomedical domain, general domain datasets or biomedical literature were used as sources. In contrast, in this work, clinical notes of other types than psychiatric notes are used as sources, which might be regarded as being “in the same domain” as the psychiatric notes. However, in most of the cases, simply combining annotated notes of different types does not show any improvement over the TO baseline. On the other hand, feature augmentation domain adaptation method is shown to be able to improve over both the TO and the S&T baseline. This indicates that different note types should indeed be considered as separate domains for the de-identification task.

Based on the results of our experiments, the best domain adaptation algorithm for de-identification task among the three algorithms is shown to be feature augmentation. While instance pruning and instance weighting do not show performance improvement over the TO baseline, feature augmentation achieves statistically significant performance improvement over the TO baseline regardless of the type of source notes. Such performance difference among the domain adaptation algorithms may come from the fact that feature augmentation allows adaptation per feature basis, as opposed to the other two methods that perform adaptation per instance basis. However, previous work that evaluates similar sets of domain adaptation algorithms for different biomedical NLP tasks reports different results than this work. Instance pruning shows the best performance for semantic role labeling (SRL) of biomedical literature, feature augmentation for SRL of clinical notes, and instance weighting for discourse connective detection in biomedical literature. Thus, it seems that the performance of domain adaptation algorithms varies depending on the characteristics of both the source and target datasets and the NLP task at hand.

It is observed that different types of notes contribute differently for de-identification of psychiatric notes. In this task, mixture notes show the biggest performance improvement through domain adaptation, achieving significantly better performance than outpatient notes or discharge summaries. Moreover, when target data size is small, simply merging mixture notes and target data works better than domain adaptation methods. This finding is interesting and requires further investigation to provide more insights. One notable fact is that psychiatric notes and mixture notes are both from the same organization, the Partners Healthcare System; while discharge summaries and outpatient notes contain documents from other institutions than Partners. We leave further investigation on domain similarities among different types of clinical notes as future work.

The size of the target corpus used in this study is much larger than those of the source corpora (in terms of number of tokens). This is different from the typical use cases of domain adaptation, where the size of source is much larger than the size of target. While it is interesting to see that domain adaptation can increase de-identification performance even when source sample size is much smaller than the target size, we expect a larger size of source data to achieve even higher performance increase through domain adaptation. In fact, an additional experiment using feature augmentation with a combined source of all existing datasets shows F-score 90.67, the highest among all the experiments done in this study. However, the effect of much larger source to de-identification of psychiatrics notes through domain adaptation is still to be investigated.

We compared de-identification errors produced by systems with domain adaptation to errors produced by the TO baseline system. While less number of errors is produced when domain adaptation is used, the distribution of the error types is not changed much. For instance, with or without domain adaptation, acronyms of hospital names such as “MBH” or “SAH” is one of the major sources of errors. For future work, we plan to test combinations of different domain adaptation algorithms as well as to incorporate domain similarity between target and source into the domain adaptation processes.
CONCLUSION

In this paper, we investigated the use of domain adaptation methods for de-identification of psychiatric notes. Three state-of-the-art domain adaptation methods were evaluated and three source datasets consisting of discharge summaries, outpatient notes, and a mixture of different types of clinical notes were studied. Our results show that domain adaptation can achieve better performance than simply merging the source and the target data, indicating the potential of domain adaptation to reduce the annotation cost when building automatic de-identification systems for new types of clinical notes.

ACKNOWLEDGEMENT

The authors were supported by NIH grants 5R01LM010681, 1R01GM102282, NU24 CA194215, 1U24CA194215-01A1, and 4R00LM012104. The CEGS N-GRID challenge was supported by NIH grants P50MH106933, 4R13LM011411.

References


A Framework for Data Quality Assessment in Clinical Research Datasets

Kathleen Lee, MPH,1 Nicole Weiskopf, PhD,2 Jyotishman Pathak, PhD1
1Weill Cornell Medicine, New York, NY; 2Oregon Health & Science University, Portland, OR

Abstract
The wide availability of electronic health record (EHR) data for multi-institutional clinical research relies on accurately defined patient cohorts to ensure validity, especially when used in conjunction with open-access research data. There is a growing need to utilize a consensus-driven approach to assess data quality. To achieve this goal, we modified an existing data quality assessment (DQA) framework by re-operationalizing dimensions of quality for a clinical domain of interest - heart failure. We then created an inventory of common phenotype data elements (CPDEs) derived from open-access datasets and evaluated it against the modified DQA framework. We measured our inventory of CPDEs for Conformance, Completeness, and Plausibility. DQA scores were high on Completeness, Value Conformance, and Atemporal and Temporal Plausibility. Our work exhibits a generalizable approach to DQA for clinical research. Future work will 1) map datasets to standard terminologies and 2) create a quantitative DQA tool for research datasets.

Introduction
In recent years, with increasing availability of electronic health record (EHR) data for research, studies frequently integrate EHR data with insurance billing and claims data to study health outcomes and cost-effectiveness.1–3 Large-scale multi-institutional studies, such as pragmatic clinical trials and comparative effectiveness research, rely on accurately defined patient cohorts to ensure that findings are valid. As more and more clinical data derived from EHRs, claims, and other sources are being collected and stored in publicly accessible repositories, such as the Database of Genotypes and Phenotypes (dbGaP)4 and the Biologic Specimen and Data Repository Information Coordinating Center (bioLINCC),5 there is an overwhelming need to harmonize these datasets and assess their quality.6,7

Several studies have looked at quality issues in clinical research data, such as the lack of data standardization, missing or incomplete clinical data, incompatible representations of data types and elements, and identified three primary challenges to evaluating the quality of a research dataset.8–11 First, data quality assessment (DQA) is often subjective and is dependent on the evaluative task or objective, which is particularly problematic because clinical research datasets are often phenotype-specific, requiring a unique patient cohort, or may cater to a specific set of participating medical centers. Moreover, data within the EHRs may be sufficient for clinical purposes, but not for research, which are typically more objective-driven. For example, the clinical concept “History of Cerebrovascular Accident” can be found in the EHR, and would provide actionable information for patient care. A particular research dataset, however, might require the presence of “History of Cerebrovascular Accident Within 3 years of Encounter.” In other words, data quality is context-dependent.12 Second, although there are certain assurance checks that researchers can conduct to ensure data quality, this process is frequently time-consuming and cumbersome, and the results of these assessments may not be meaningful without a thorough understanding of the researcher’s intended goal. Evaluating data missingness, distributions, and accepted values do not entirely paint a full picture of the quality of the dataset as described by the research objective, and manually evaluating quality in this fashion is a resource-intensive task. Finally, there are no consistent evidence-based or community-driven metrics for assessing the quality of research data. Study investigators frequently develop ad-hoc metrics that are specific to the study, and cannot be replicated.13,14

Objective
It is crucial to quantitatively analyze the quality of a dataset to ensure reproducibility in research studies. To encourage implementation of the recommended concepts of quality assessment, rules and conditions are required to
formulate an assessment framework that is related to a task or specific phenotype of interest. This is a more challenging task when evaluating quality in research datasets as they are often task- or domain-specific.

The present study sought to promote the development and utilization of large, multi-institutional research datasets based on existing data sources. Specifically, we inventoried and assessed the quality of common phenotypic data elements (CPDEs) for heart failure research, enabling a reusable framework development process for other researchers looking to use or contribute to this composite dataset. Our first aim is to modify an existing DQA framework by re-operationalizing the definitions of several data quality dimensions. The framework will be dependent on a particular research goal, which in this case are studies identifying novel biomarkers for heart failure diagnosis, prognosis and treatment. The second aim is to create an inventory of CPDEs derived from the research datasets. In this study, we limit our scope to heart failure biomarker studies in the following open-access databases: dbGaP and BioLINCC. The third and final aim is to evaluate the data element inventory using the modified DQA framework. The task-oriented approach will evaluate, based on the necessary data elements for a study design or goal, the Completeness, Conformance, and Plausibility of the CPDE inventory.

Materials and Methods

Prior Work in Quality Assessment Frameworks

Data quality frameworks and harmonized assessment terminologies have been created to evaluate EHR data. Prior research in DQA defined broad dimensions of data quality. However, in a growing field of data quality research, inconsistent definitions make it difficult to uniformly compare results across data sharing partners and institutions. Efforts to harmonize these concepts are necessary to promote interoperability in the field of data quality research. A harmonized and revised DQA terminology framework was developed to encompass quality concepts that have been defined by other researchers. The proposed harmonized framework takes the categories of Conformance, Completeness, and Plausibility and expands on each.

- **Conformance** is defined by whether data values adhered to pre-specified standards or formats. Conformance was separated into three distinct sub-categories: Value Conformance (whether recorded data elements agree with constraint-driven data architectures, such as data models or rules defined in a data dictionary), Relational Conformance (determines if data elements agree with structural constraints of the physical database that stores these values, hinging on the importance of primary key and foreign key interactions within relational databases), and Computational Conformance (focuses on the correctness of the output value of calculations that were made from existing variables, either within the dataset or between datasets). Although Value Conformance can be ascertained at both the data element and data value levels, Computational Conformance can only be properly assessed at the value level. The scope of this study focused only on single datasets, typically in flat files, such as Excel or CSV formats. As a result, Relational Conformance is not applicable at the data element level for this study.

- **Completeness** evaluates data attribute frequency within a dataset without reference to the data values. It does not consider its structure or its plausibility, but instead looks at the absence of data at a specific point in time agreeing with a trusted standard, common expectation, or existing knowledge. This dimension is applicable to both higher-level data element and more granular data value levels of assessment.

- **Plausibility** is defined by whether or not the values of data points are believable when compared to the expected representation of an accepted value range or distribution. Plausibility was separated into Uniqueness Plausibility (values that identify a particular object—person, institution, etc.—are not duplicated), Atemporal Plausibility (data values adhere to common knowledge or are verified by an external source), and Temporal Plausibility (whether time-varying variables also have changing values, based on gold standards or existing knowledge). All sub-categories within Plausibility are applicable to both data element and data value levels.
**DQA Framework Modification**

The study team used a consensus-driven approach to finalize the framework creation to redefine the harmonized DQA terminology from Kahn et al.\textsuperscript{12} to the specific research task. In particular, the concepts of *Conformance*, *Completeness*, and *Plausibility* were operationalized to be specific to heart failure biomarker research. Table 1 includes definitions of the harmonized data quality assessment terms and examples.

**Table 1.** Harmonized Terminology with Examples Drawn from Heart Failure Research Studies.

<table>
<thead>
<tr>
<th>Concept</th>
<th>Applicable Level</th>
<th>Definition</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Value conformance</td>
<td>Data Element</td>
<td>Whether recorded data elements agree with constraint-driven data architectures</td>
<td>The description for the data element, “Body Mass Index (BMI)” is defined by units of kg/m(^2).</td>
</tr>
<tr>
<td>Relational conformance</td>
<td>Data Element*</td>
<td>Whether data elements agree with structural constraints of the physical database that stores these values</td>
<td>The data element for “History of Cerebrovascular Incident” is represented by categorical values, Yes or No. The values are represented as such, and are not in any other form.</td>
</tr>
<tr>
<td>Computational conformance</td>
<td>Data Value</td>
<td>Whether the correctness of the output value of calculations that were made from existing variables, either within the dataset or between datasets</td>
<td>Calculating patient body weight and height would produce the same value as the value represented in the BMI data element.</td>
</tr>
<tr>
<td>Completeness</td>
<td>Data Element</td>
<td>Whether data values or elements are present</td>
<td>The research data elements in the inventory are complete when compared to the aggregated list from literature.</td>
</tr>
<tr>
<td>Uniqueness plausibility</td>
<td>Data Element</td>
<td>Whether values that identify a particular object--person, institution, etc.--are not duplicated</td>
<td>Each data element is not duplicated or represented by another data element within the inventory.</td>
</tr>
<tr>
<td>Atemporal plausibility</td>
<td>Data Element</td>
<td>Whether or not data values adhere to common knowledge or are verified by an external source</td>
<td>Chronic Kidney Disease (CKD) stage 2 criteria of GFR &lt;60 mL/min/1.73 m(^2) for &gt;=3 months is in concordance with existing knowledge and guidelines for of CKD diagnosis.</td>
</tr>
<tr>
<td>Temporal plausibility</td>
<td>Data Element*</td>
<td>Whether time-varying variables also have changing values, based on gold standards or existing knowledge</td>
<td>Follow-up dates are sequentially collected after the study enrollment date.</td>
</tr>
</tbody>
</table>

* Rows marked with an asterisk indicate exceptions to the dimensional applicability at that particular data level. *Relational Conformance* requires a SQL database or relational database structure. While *Relational Conformance* can be applied at the data element level, it requires a specific structure of tables beyond the scope of the study. Similarly, *Temporal Plausibility* requires time-varying data elements, which were not included within the ‘Demographics’ and ‘Medications’ categories in the current CPDE inventory.

In our modified framework, *Value Conformance* measures that assessed the adherence of data values to internal constraints necessitated the evaluation of acceptable ranges for data elements like patient sex, blood pressure, or cholesterol levels. *Completeness* compared the data element inventory to the aggregated data elements within literature.\textsuperscript{17–19} These studies compiled cardiovascular EHR data elements, and in particular, key heart failure data elements, that have research utility and maximum clinical impact. *Plausibility* required the evaluation of data values
specific to cardiovascular disease, and understanding that patients with a heart failure diagnosis should have statistically similar distributions for certain data elements compared to patients without a heart failure diagnosis.

**Conformance** for research data is still defined by whether data values adhered to pre-specified standards or formats. We also separate conformance into three sub-categories. Value conformance in research data would be determining whether elements are represented according to some standard medical terminology or appropriate clinical nomenclature. External standards can be applied as well, such as units of measurement or ranges of accepted values. Relational conformance determines if data elements agree with structural constraints of the physical database that stores these values. Relational conformance deals with how a data model represents reality via metadata descriptions or database rules. **Computational conformance** within research datasets include validation checks that are consistent with EHR validation checks, such as whether body mass index (BMI) calculations for data elements like, “Patient Body Weight” and “Patient Height” should yield the same values for the data element, “Body Mass Index”.

**Completeness** evaluates the presence of data elements within a particular dataset. It does not reference data values and it compares the dataset to an existing standard knowledge or common expectation. Because of the subjective nature of compiling a “complete” research dataset for a given disease domain, we compared the data element inventory against cardiovascular data elements aggregated through literature guidelines. Although these were EHR data elements, they were compiled to create a list of attributes that had research utility in clinical settings. This was in line with our focus for heart failure biomarker research studies.

**Plausibility** is whether or not the values of data points are believable when compared to the expected representation of an accepted value range or distribution. **Uniqueness plausibility** ensures that values are not duplicated or represented by another entity within the dataset. At the research data element level, we ensure that data elements are not duplicated or represented by another data element or attribute. At the data value level, we ensure that each data point is not duplicated within another data element. **Atemporal plausibility** determines whether or not data values adhere to common knowledge or are verified by an external source. This extends to research data elements with metadata descriptions that align with existing knowledge. For example, a data element representing “Chronic Kidney Disease Stage 2” should have metadata descriptions consistent with Chronic Kidney Disease stage 2 criteria, such as appropriate laboratory values for glomerular filtration rate (GFR) and blood urea nitrogen (BUN). Values should be consistent with external standards of acceptable ranges or distributions of values. **Temporal plausibility** is whether time-varying variables also have changing values or follow sequentially, based on gold standards or existing knowledge. For example, do values vary over time as expected, such is the case for spikes in flu diagnosis in emergency room or outpatient visits during flu season? Similarly, recruitment dates into the research study should not come before patients’ dates of birth. A follow-up date should not precede the recruitment date for the study. These values follow for other variables that require follow-up dates within a study.

**Common Phenotype Data Element Inventory Creation**
We conducted a retrospective review of open-access cardiovascular disease datasets in order to identify CPDEs related to heart failure. We focused our literature search on studies identifying biomarkers and risk factors for heart failure diagnosis, prognosis and treatment. Several cardiovascular research studies, particularly focusing on heart failure, from dbGaP and BioLINCC were aggregated to compile an inventory (work was led by co-authors KL and JP) of commonly used phenotype data elements. Our inclusion criteria for studies were: 1) focus on heart failure or congestive heart failure biomarker research; 2) use of clinical data in addition to genotype or sequencing data; and 3) inclusion of demographics, diagnostic test, patient history, physical examination, and medications variables. Authorized access was obtained for these research datasets. Data elements for the inventory were selected based on their relevancy to heart failure biomarker research, research utility in clinical settings, and their presence in EHR systems. **Figure 1** illustrates the process for collating CPDEs from dbGaP and BioLINCC.
The development of our data element inventory consisted of a two-pronged approach. First, we aggregated CPDEs that we found in heart failure biomarker research studies in dbGaP and BioLINCC. In aggregating appropriate CPDEs from the studies, we focused on those that had greater generalizability for heart failure studies. Elements like indices of social support networks, for example, were not included in our inventory, as they were not deemed to be generalizable to most heart failure research studies that utilized EHR data. We followed guidelines from literature that enumerated relevant data elements present in EHR databases that could be repurposed for clinical research. Based on these guidelines, social elements were excluded from the inventory. Second, following similar data element guidelines from literature, clinical variables with high research utility, such as diagnostic tests and medical history, were prioritized over those that are more attuned to specific research designs, such as insurance, government aid sources, and billing zip codes. Of particular usefulness in our aggregation of data elements were a report from the American College of Cardiology Foundation and the American Heart Association Task Force on Clinical Data Standards, which focused on harmonizing existing data standards with newly published ones, and to establish terms that are available in every general purpose EHR, and are extendable and reusable in clinical research, a report from the Data Standards Workgroup of the National Cardiovascular Research Infrastructure Project, which attempted to create or identify and harmonize clinical definitions for a general set of cardiovascular data elements, and a report from the American College of Cardiology and American Heart Association Task Force on Clinical Data Standards reviewed key data elements for heart failure management in clinical research.

**Results**

*Finalized Data Element Inventory for Heart Failure Research*

We created an inventory of 100 data elements from the following studies: the Cardiovascular Health Study (CHS), the Jackson Heart Study, the Framingham Heart Study, the Heart Failure Network (HFN) CARdiorenal Rescue Study in Acute Decompensated Failure (CARRESS), and the Sudden Cardiac Death in Heart Failure Trial (SCD-HeFT). Data elements were selected based on comprehensive coverage of cardiovascular conditions, test results, and clinical presentations that would best reflect the breadth and variety of commonly occurring heart failure biomarker research data elements in clinical research. The final data element inventory is available as an online
Our review of relevant guidelines led to the identification of six categories of frequently utilized and clinically meaningful phenotypic data elements:

- **Demographics**: All research datasets have at minimum a demographics category for patient records. Commonly occurring data elements within this category included the Date of Birth, Sex, Race, and Ethnicity. Four demographic data elements were included in the inventory.

- **Physical Examination or Baseline Observation**: This category can often be fairly detailed depending on the research study. Data elements related to systemic observations of the body and functions include measurements of height, weight, body mass index, blood pressure, and heart rate. We included ten data elements in this category.

- **Diagnostic Tests**: Diagnostic and therapeutic procedures, as well as laboratory tests are included in this category. Units of measurement in their metadata description often accompany these variables. Data elements resulting from electrocardiograms, bypass graft surgeries, echocardiograms, and stress tests are included, as well as laboratory values like cholesterol, glucose, creatinine, and blood urea nitrogen measurements. Twenty-six data elements were included in the inventory.

- **Patient Medical History**: Typical data elements in this category include prior history of disease or diagnoses, prior surgeries or hospitalizations, history of tobacco use, drug use, and alcohol use, and family history of disease. These data elements are often categorical, such as the data element, “Type of stroke”, which takes on values, “Hemorrhagic,” “Nonhemorrhagic,” and “Unknown.” Twenty-seven medical history items were included in the data element inventory.

- **Clinical Diagnoses or Presentation**: This category is limited to the patient’s current status. In research datasets for heart failure biomarker research, patients are often diagnosed with heart failure, or another cardiovascular disease to be included in the study cohort. Inclusion of clinical presentations was considered for this section as EHR systems record current statuses of patients differently. Data elements, such as Myocardial Infarction, Chest Pain (Angina), Heart Failure, Syncope, and the like, were included. Eight patient assessment data elements were included in the inventory.

- **Medications**: Types of medication, such as beta blockers, ACE inhibitors, and statins were included, as well as data elements describing usage, such as Medications Held or Discontinued, Contraindications, and timepoints of usage were added to the inventory. The inventory contained twenty-five medications-related data elements.

### Table 2. Data Element Inventory Framework Assessment by Category

<table>
<thead>
<tr>
<th>Data Element Category</th>
<th>N</th>
<th>Value Conformance</th>
<th>Completeness</th>
<th>Uniqueness plausibility</th>
<th>Atemporal plausibility</th>
<th>Temporal plausibility</th>
</tr>
</thead>
<tbody>
<tr>
<td>Demographics</td>
<td>4</td>
<td>4 (100%)</td>
<td>4 (100%)</td>
<td>4 (100%)</td>
<td>4 (100%)</td>
<td>N/A</td>
</tr>
<tr>
<td>Physical Exam</td>
<td>10</td>
<td>10 (100%)</td>
<td>10 (100%)</td>
<td>10 (100%)</td>
<td>10 (100%)</td>
<td>10 (100%)</td>
</tr>
<tr>
<td>Diagnostic Test</td>
<td>26</td>
<td>26 (100%)</td>
<td>23 (88.4%)</td>
<td>26 (100%)</td>
<td>26 (100%)</td>
<td>23 (88.4%)</td>
</tr>
<tr>
<td>Medical History</td>
<td>27</td>
<td>27 (100%)</td>
<td>27 (100%)</td>
<td>23 (85.2%)</td>
<td>27 (100%)</td>
<td>27 (100%)</td>
</tr>
<tr>
<td>Patient Assessment</td>
<td>8</td>
<td>8 (100%)</td>
<td>8 (100%)</td>
<td>8 (100%)</td>
<td>8 (100%)</td>
<td>8 (100%)</td>
</tr>
<tr>
<td>Medications</td>
<td>25</td>
<td>25 (100%)</td>
<td>25 (100%)</td>
<td>25 (100%)</td>
<td>25 (100%)</td>
<td>N/A</td>
</tr>
</tbody>
</table>
**Assessment of Data Element Inventory Against Modified DQA Framework**

Having redefined the DQA framework and creating an inventory of accepted data elements for comparison, we were able to evaluate the quality of our collected data element inventory. **Table 2** displays the results of the inventory against the framework. The inventory was assessed at the data element variable level. For some DQA concepts that are more suitable for assessing variable values, we evaluated the appropriate variables based on their ability to capture the significant aspect. For instance, to evaluate particular data elements on *Temporal Plausibility* at the data element variable level, we assessed their ability to capture temporal aspects, such as continuity of data collection and whether or not they were defined by a numeric date, where appropriate.

**Discussion**

This work is not an exhaustive list of modifications to a DQA framework for research datasets. Although there are many ways that research data and EHR data coincide in terms of data quality, the applications for verification and validation can be quite different. While we understand that the data element inventory was created to obtain a “model” for accepted variable values that work within our modified DQA framework, we acknowledge the potential limitations for its application as a comparison to other research datasets that may have a broader or narrower focus even within the scope of heart failure biomarker research studies. Our goal was to adapt a harmonized DQA framework to a clinical domain, such as heart failure, and inevitably to compile a working DQA framework that can be reusable in clinical research.

Although harmonized frameworks are typically evaluated at the data value level, we attempted to apply it at a broader element level for this study. All data elements met the criteria for *Value Conformance* by being thoroughly represented by appropriate units of measurement in their metadata descriptions. However, mapping data elements to a standard terminology, such as the Systematized Nomenclature of Medicine (SNOMED)\textsuperscript{25,26} would further ensure that data elements adhered to the best practices for research data management. There remains a subjective aspect to our approach. 88.4% of the Diagnostic Test data elements met the criteria for *Completeness*. Including dates for certain diagnostic procedures that may vary over time might have led to a more complete list. For example, including the date of “Radionuclide Ventriculography Findings” typically helps investigators track a patient’s disease progression over time if more tests are conducted at different time points.

The data elements in the inventory are considered a complete collection of appropriate variables to be included in a research dataset for heart failure biomarker research as it includes elements that encompass demographics, physical examination, tests, patient medical history, and medications. Collating the data elements for the inventory involved consulting guidelines set in literature for high research utility clinical variables in heart failure studies, as well as evaluating commonly occurring data elements that are present in open-access clinical research datasets. Some studies included other lifestyle factors, such as sources of social support, eating habits, or scales for depression and anxiety. We considered these data elements to be not as generalizable in heart failure studies according to literature guidelines previously set for EHR data elements. These guidelines included only common elements that had high clinical research utility, and as a result, the aforementioned social indices were not included in our data element inventory. *Uniqueness Plausibility* ensures that elements are not duplicated and values are not dually represented within a dataset. 85.2% of Medical History data elements adhered to the concept of *Uniqueness Plausibility*. Certain data elements, such as Family History of Coronary Arteriosclerosis, Family History of Cardiomyopathy, and Family History of Sudden Cardiac Death shared some overlapping definitions and descriptive criteria.

Because our data element inventory is presented in a single table, and there are no relationships between the data elements themselves (for example, as one would observe in an Entity-Relationship model), it did not qualify to be assessed for *Relational Conformance*. There are no additional tables in the data element inventory, although some research datasets may have separate tables for categories of measurements. Additionally, we are unable to evaluate *Computational Conformance* as we are only evaluating at the data element variable level, which provides no output variable values for us to calculate on. *Temporal Plausibility* was incalculable for Demographics and Medications.
categories as the elements were not expected to vary over time. Should there have been prescription dates for Medications, perhaps Temporal Plausibility might have been more applicable. In our inventory, however, Medications were only listed as drug classes.

Limitations
The first limitation is that there are certain data quality concepts that cannot fully apply to research data. Relational Conformance, for example, deals with the ability to navigate between different tables, which may not be necessary if the table in question is a research study data dictionary. These are structured differently and are typically not stored in the same manner as EHR systems tables.

Second, to properly assess Value Conformance, it is often necessary that data elements in research datasets are mapped to standardized biomedical terminologies. An underlying data quality issue in research data is the inability to be readily integrated or linked to relevant datasets due to a lack of standardization. This can be addressed by mapping data elements to terminologies or models to further ensure that the data elements are represented appropriately. In addition, string variables present a familiar challenge for researchers working with ontologies as mapping these terms can be difficult when compiling EHR data and when transforming data from one schema to another. We aimed to include values within a string variable that contain more structured response options. For a variable, such as “Lung (pulmonary) examination”, the values could take the form of free-text responses, which would present a challenge for data mapping. Its values, however, include structured responses, like “Clear or normal, Rales (height of rales when patient sitting upright should be noted), Decreased breath sounds or dullness, Rhonchi, or Wheezing,” which can more easily facilitate mapping to standard terminologies.

Third, because data quality assessment is a subjective task, it is necessary to have an external gold standard for comparison. We used literature guidelines that focused on EHR heart failure data elements with high research utility to create our data element inventory. The objective of this study, after all, is to evaluate data quality within open-access research datasets that focus on biomarker discovery, which are often linked to EHR data. Fourth, our methodology for assessing data quality was limited only to the data element variable level: were the data elements complete and in line with a set of external standards or common knowledge? Focusing only on evaluating data quality at the data element variable level, much like assessing a data dictionary, restricts our ability to look at the variable values and understand its accepted distributions, ranges, and completeness or missingness. Further, we were unable to conduct validation checks that may enable us to assess for computational conformance.

Future Work
Assessing data quality at the higher-level data element phase was a necessary first step in determining the ability of clinical research data to adhere to a harmonized DQA framework. To more comprehensively evaluate the utility of our harmonized framework, future projects will adapt the framework at the data value level as well. We anticipate that including data values will produce more criteria with which to better assess quality, most likely producing less adherence than what our current results exhibit. In addition, we anticipate that the addition of data elements, such as social support and insurance and billing procedures may provide us with a richer set of criteria to broaden our DQA dimensions. Incorporating more granular data values may involve either a more comprehensive data element inventory, or data transformation of several datasets to a common model so that they have a similar baseline for comparison. The latter exercise of data transformation into a common data model can also help alleviate the mapping challenges of string variable responses that can be particularly difficult to standardize. Evaluating the framework against differing granularities can more appropriately showcase its ability to be repurposed continuously in research, independent of clinical domain. Our eventual goal is to create an assessment tool to better quantify data quality using this framework. The first step is to create a stepwise script that can run on a statistical program, such as R or SAS. This would enable researchers to go through their appropriate data elements within their data dictionary and check for missingness and completeness, appropriate ranges, and distributions. They can also check Computational Conformance in this way. The second goal is to create a stepwise tool that can assess the overall

1087
Acknowledgement: This work was funded in part by NIH R01 GM105688 and R01 GM103859.

References

24. dbGaP. Framingham Cohort.


28. Yale University. The YODA Project.
A Computational Framework for a Digital Surveillance and Response Tool: Application to Avian Influenza
Eva K Lee, PhD1,2,3,*, Yifan Liu, BS1,2,3, Ferdinand H Pietz, MPH4
1NSF-Whitaker Center for Operations Research in Medicine and Healthcare; 2NSF I/UCRC Center for Health Organization Transformation; 3School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA; 4Strategic National Stockpile, Centers for Disease Control and Prevention, Atlanta, GA.

Abstract
Avian influenza viruses have caused infections and deaths in wild birds, commercial poultry, and humans. It poses an increasing threat of a pandemic. To understand the transmission dynamics of avian influenza viruses and assess the effectiveness of different containment strategies, we develop a flexible modeling framework based on multi-layer compartmental models for digital disease surveillance and response in combating pandemics. The model can accommodate other disease outbreaks under diverse settings. We demonstrate its usage on avian influenza and derive the basic reproduction number and spread characteristics. We contrast the effectiveness of different containment strategies and their combination effect in protecting both the human and the bird populations. Our system, a digital surveillance and response system (RealOpt-ASSURE), can record, monitor, and predict avian influenza outbreaks. It combines with intervention strategies to return policies and on-the-ground operations/actions that are needed for best population protection. RealOpt-ASSURE can accept heterogeneous types of surveillance data. It can help decision makers to evaluate the risk of a pandemic and choose proper containment strategies to rapidly mitigate the outbreak.

*Corresponding author: eva.lee@gatech.edu

This work is partially supported by a grant from the Centers for Disease Control and Prevention and the National Science Foundation (IIP-1516074). The findings and conclusions in this paper are those of the authors and do not necessarily represent the official position of the funding agencies.

Introduction
Avian influenza is caused by infection with avian influenza Type A viruses. These viruses were first isolated from common terns in South Africa1, and were later extensively studied. They occur naturally among wild birds and can infect domestic poultry and other animal species, including cats and swine. The highly pathogenic avian influenza (HPAI), and the very virulent types of avian influenza Type A viruses (majorly subtypes H5 and H7), are fatal for birds nearly 100% of the time2. They do not normally infect humans, but sporadic human infections with avian flu viruses have been reported, with approximately 50% mortality rate3. There is no clinical evidence that avian influenza can be directly transmitted between humans4, but such possibility exists due to mutant avian influenza viruses, and should be closely monitored to reduce the infections and mortalities among humans5.

From 1959 to early 2000, primary outbreaks of HPAI in poultry have been reported 17 times2. The more recent outbreaks started in 2003, when multiple avian influenza cases in wild birds, domestic poultry, and humans were reported worldwide. The H5N1 influenza outbreak in Southeast Asia in 2003 infected more than 130 people in Vietnam, Thailand, and Cambodia and killed more than half of them6. The same subtype appeared in wild birds and domestic poultry in China afterwards, causing human infections as well7–9. It was estimated that by 2005, this outbreak had caused more than $10 billion in losses to the Southeast Asian poultry industry9. Large outbreaks in commercial poultry farms in industrialized countries in Europe and United States have also been reported10, 11, indicating that avian influenza is not only a threat to human health, it also jeopardizes the poultry industry and has significant impact on the economy. Predicting the trend of an avian influenza outbreak and implementing an optimal intervention strategy to achieve early containment is crucial. Pandemic planning and large-scale surveillance are essential in promoting effective responses12, with real-time analysis of surveillance data being critically important to detect and predict an outbreak in a timely manner.

Numerous mathematical models have been developed to describe the spread of an epidemic to facilitate decision making using surveillance data. During the past two decades, various models were proposed for general influenza and avian influenza. Probability-based quantitative analysis was used to estimate the risk of an emerging pandemic strain of avian influenza13. Agent-based simulation of individual and community activities was developed to prevent and control influenza pandemics14. Stochastic models were developed to predict the worldwide spread of pandemic influenza based on global travel information15, 16. Compartmental models remain a popular method for pandemic studies. Ordinary differential equation-based methods were used to interpret the mutation process of avian influenza17, model the dynamics of human infections18, and investigate nonlocal epidemics19.
In this paper, we introduce a modeling framework for infectious diseases based on the compartmental modeling concept. Instead of specifying a set of compartments and a certain transmission and development structure, this modeling framework is highly generic and flexible, and can accommodate diverse types of infectious diseases. Analyzing the model using avian influenza, we perform sensitivity analysis to understand the risk of infection, and evaluate efficiency of different containment strategies in terms of both human and bird infections. We demonstrate the design of a general purpose “digital disease surveillance and response” system, RealOpt-ASSURE, for tracking the avian influenza outbreaks and predicting future disease trends; and discuss how the system can assist in decision making to control and prevent outbreaks.

Methods and Designs

Our general purpose modeling framework is based on compartmental models that combine both the disease characteristics and the response operations. Consider a homogeneous group in a compartmental model, and let \( \Phi \) denote the collection of all stages. We use the following two ways to categorize stages:

- **Passive/Active stages**: Passive stages are those under which an entity will not change its status spontaneously; while active stages are those under which an entity will change its status spontaneously. Let \( \Phi_P \) denote the collection of passive stages, and \( \Phi_A \) denote the collection of active stages. \( \Phi_P \cup \Phi_A = \Phi \).
- **Vulnerable/Contagious stages**: Vulnerable stages are those under which an entity can still be infected; and contagious stages are those under which an entity can infect other entities who are vulnerable. Let \( \Phi_V \) denote the collection of vulnerable stages, and \( \Phi_C \) denote the collection of contagious stages. We have \( \Phi_V \cup \Phi_C \subseteq \Phi \).

Define the vector \( \mathbf{y}_\phi = (\phi(t))_{\phi \in \Phi} \) as the number of entities in each compartment at time \( t \). Let \( \mu_\phi = (\mu_\phi)_{\phi \in \Phi} \) be the vector of transition rates of all stages in the model. Note that only the active stages have well-defined transition rates, thus for completeness, we define \( \mu_\phi = 0 \) for all \( \phi \in \Phi_P \). Similarly, let \( \mathbf{b}_\phi = (\beta_\phi)_{\phi \in \Phi} \) be the vector of baseline infection rate of all vulnerable stages in the model adjusted by the total population. For non-vulnerable stages, define \( \beta_\phi = 0 \) for all \( \phi \in \Phi_V \). For simplicity, we define a function \( s \) on \( \Phi : s(\phi) \) returns the set of all successor stages of \( \phi \). Let matrix \( \mathbf{C} = (c_{ab})_{a \in \Phi, b \in \Phi} \) denote the disease contagious matrix, where \( c_{ab} = -1 \) if \( a \in \Phi_V \) and \( b \in \Phi_C \), and \( c_{kb} = 1 \) for all \( k \in s(a) \) if \( a \in \Phi_V \) and \( b \in \Phi_C \). Finally, let matrix \( \mathbf{D} = (d_{ab})_{a \in \Phi, b \in \Phi} \) be the disease transition matrix, where \( d_{ab} \) is the transition probability from stage \( b \) to stage \( a \). Then the ordinary differential equations for the compartmental model with one homogeneous group can be written compactly as follows:

\[
\mathbf{y}' = (\mathbf{D} - \mathbf{I})\text{diag}(\mathbf{\mu})\mathbf{y} + \text{diag}((\mathbf{D} + \mathbf{I})\text{diag}(\mathbf{b})\mathbf{y})\mathbf{C}\mathbf{y}
\]  

(1)

where \( \mathbf{I} \) is the identity matrix. The first part of the equation models the change in compartments due to the development of disease, and the second part models the effect of new infections. If we take the natural changes of the group population independent of the disease outbreak into consideration, we may add a term \( \lambda(\mathbf{y}) \) to equation (1), where \( \lambda \) is a function of current population and depicts the natural changes in population.

The modeling framework for avian influenza consists of two parts: the compartments for human, subscripted by \( H \); and compartments for birds, subscripted by \( B \). The infections among birds are associated with contacting infectious birds, while the transmission from the susceptible to the infectious in human is associated with contacting infectious birds, as well as their saliva, mucus, and feces. The spread of avian influenza \( A \) viruses from one ill human to another is rare. However, because of the high possibility of viruses to mutate and gain such an ability, the human-to-human transmission is also included in the model. Incorporating the above features and applying formula (1) on the human and bird group individually will result in our generic modeling framework for avian influenza:

\[
\mathbf{y}_H' = (\mathbf{D}_H - \mathbf{I})\text{diag}(\mathbf{\mu}_H)\mathbf{y}_H + \text{diag}((\mathbf{D}_H + \mathbf{I})\text{diag}(\mathbf{b}_H)\mathbf{y}_H)\mathbf{C}_H\mathbf{y}_H + \lambda_H(\mathbf{y}_H)
\]  

(2-1)

\[
\mathbf{y}_B' = (\mathbf{D}_B - \mathbf{I})\text{diag}(\mathbf{\mu}_B)\mathbf{y}_B + \text{diag}((\mathbf{D}_B + \mathbf{I})\text{diag}(\mathbf{b}_B)\mathbf{y}_B)\mathbf{C}_B\mathbf{y}_B + \lambda_B(\mathbf{y}_B)
\]  

(2-2)

where \( \mathbf{b}_H \) and \( \mathbf{C}_H \) are the baseline infection rate vector and contagious matrix between the group human and bird.

In this study, we apply an SEIRD (Susceptible, Exposed, Infectious, Recovered, Deceased) model to describe the disease spread among the human population. The human population is assumed to follow a logistic growth with carrying capacity \( K_H \) and growth rate \( r \) during the outbreak. The virus transmission among birds is modeled with an SID (Susceptible, Infectious, Deceased) model. The bird population is assumed to have a constant death rate \( v \) and a variable birth rate \( \Lambda(t) \) such that the total population is constant \( K_B \). Figure 1 shows the transition diagram. Assume the probability of a human recovering from an infection is \( p \), and the basic contact rate between humans is \( \beta_H \). Under this assumption, the matrices in formula (2) for the human compartments are:
The basic contact rate between birds is $\beta_B$. The matrices for the bird compartments are

$$\mu_B = \begin{pmatrix} 0 \\ \mu_{HI} \\ 0 \\ 0 \end{pmatrix}, \beta_B = \begin{pmatrix} \beta_B \\ 0 \\ 0 \\ 0 \end{pmatrix}, D_B = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & p & 0 \end{pmatrix}, C_B = \begin{pmatrix} 0 & 0 & -1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$ 

Further, we assume the basic contact rate between human and bird is $\beta_{HB}$. The matrices for the interaction term between human and bird are

$$\beta_{HB} = \begin{pmatrix} \beta_{HB} \\ 0 \\ 0 \\ 0 \end{pmatrix}, C_{HB} = \begin{pmatrix} 0 & -1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$ 

Figure 1. Transition diagram of the compartmental model. The dashed lines refer to the contacts that cause infections.

To derive the basic reproduction number for this system, we apply the method of next generation matrix. By arranging the compartments corresponding to infections as $S_H, E_H, I_H$ and evaluating the components of next generation matrix at equilibrium, we obtain

$$F = \begin{pmatrix} 0 & \beta_H K_H & \beta_{HB} K_H \\ 0 & 0 & 0 \\ 0 & 0 & \beta_B K_B \end{pmatrix}, V = \begin{pmatrix} \mu_{HE} + r & 0 & 0 \\ -\mu_{HE} & \mu_{HI} + r & 0 \\ 0 & 0 & \mu_{BI} + \nu \end{pmatrix}.$$ 

Therefore, the basic reproduction number for the system is the spectral radius of the next generation matrix $FV^{-1}$:

$$R_0 = \rho(FV^{-1}) = \max \left\{ \frac{\beta_H K_B \mu_{HE}}{(\mu_{HE} + r)(\mu_{HI} + r)}, \frac{\beta_B K_B}{\mu_{BI} + \nu} \right\}.$$ 

The first component in the equation is the basic reproduction number within the human compartments, and the second component is the basic reproduction number within the bird compartments. Since human-to-human infection of avian influenza is inefficient and rare, we can conclude that $\beta_H K_B \ll \beta_B K_B$, thus the basic reproduction number of the system is determined by the infections among the birds. To quickly contain an avian influenza outbreak, it is desirable to reduce the basic reproduction number such that $R_0 < 1$ so that an infected bird produces less than one new infection during its infectious period. This can be achieved by either reducing the adjusted basic contact rate $\beta_B$ by quarantining the poultry farms with confirmed cases, or reducing the bird population $K_B$ by slaughtering infectious birds, both wild and live poultry.

To examine the performance of the model and understand the effectiveness of containment strategies, we perform sensitivity and scenario analysis using the 2010 avian influenza data in Egypt. We vary the contact rate between birds ($\beta_B$), and the duration of infection in birds ($\mu_B$). We report the estimated number of infections in both birds and humans, as well as time to reach containment. The simulation is performed until an asymptotic behavior is achieved, i.e., when no new cases are reported.

**Results**

**Egypt.** Egypt is considered an important poultry producer in the Arab world with 1.5 million permanent workers and 1 million temporary workers in poultry production. The combination of commercial poultry, backyard birds, and live-bird markets exposes Egypt to the danger of avian influenza outbreak. Evidence shows that stable lineages of H5N1 viruses have been established among chickens and humans in Egypt, and the control efforts and international cooperation is in urgent need. We use the data of reported infections in live poultry of an HPAI outbreak in Egypt during January to June 2010 for our analysis. The total number of reported cases in live poultry was 317,400 by containment.
Table 1. Parameters used in modeling the avian influenza outbreak in Egypt

<table>
<thead>
<tr>
<th>Name</th>
<th>Notation</th>
<th>Value</th>
<th>Literature Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial live poultry population</td>
<td>$S_B$</td>
<td>1,800,000,000</td>
<td>25, 27</td>
</tr>
<tr>
<td>Initial infected live poultry</td>
<td>$I_B$</td>
<td>3,000</td>
<td>Fitted</td>
</tr>
<tr>
<td>Average live poultry lifespan</td>
<td>$1/\nu$</td>
<td>60 days</td>
<td></td>
</tr>
<tr>
<td>Average live poultry lifespan after infection</td>
<td>$1/\mu_{H}$</td>
<td>10 days</td>
<td></td>
</tr>
<tr>
<td>Baseline infection rate among live poultry</td>
<td>$\beta_B K_B$</td>
<td>58/day</td>
<td>Fitted</td>
</tr>
<tr>
<td>Baseline infection rate between live poultry and humans</td>
<td>$\beta_B K_H$</td>
<td>$2 \times 10^{-4}$/day</td>
<td></td>
</tr>
<tr>
<td>Baseline infection rate between humans</td>
<td>$\beta_H K_H$</td>
<td>$1.5 \times 10^{-7}$/day</td>
<td></td>
</tr>
<tr>
<td>Initial human population (Egypt)</td>
<td>$S_H$</td>
<td>82,000,000,000</td>
<td>World Bank</td>
</tr>
<tr>
<td>Natural population growth rate (Egypt)</td>
<td>$r$</td>
<td>1.6%</td>
<td>World Bank</td>
</tr>
<tr>
<td>Basic reproduction number</td>
<td>$R_0$</td>
<td>1.362</td>
<td>Calculated from $R_0$</td>
</tr>
</tbody>
</table>

Figure 2 contrasts the reported cases in live poultry against the simulation result by varying the contact rate ($\beta_B K_B$) and the duration of infection ($\mu_{H}$) in live poultry. The results suggest that Table 1 provides good estimates for the reported cases (green solid curve from simulation against red dotted curve of reported cases). When the basic contact rate increases from 56 per day to 58 per day (dotted against the dashed green curves), the cumulative infections in bird compartments increase from 0.22 million to 0.46 million by June 2010 (110% increase). Varying the duration of infection in birds seems to have a larger impact on the trend of the outbreak (the three solid lines). When increasing the bird infection time from 9 days to 11 days, not only the total infections at containment sees a 4-fold increase (0.16 million to 0.6 million), the time to containment is delayed by two months as well (April 2010 to June 2010). Therefore, reducing the infection duration of birds plays a crucial role in controlling the trend of the avian influenza outbreak. This can be achieved by providing medical countermeasures to affected poultry farms, or quickly removing/isolating infectious live poultry upon identification of an outbreak.

Figure 2. Varying contact rate and duration of infection in birds and its effect on infection and containment.

Next, we focus our analysis on the infections among humans using the parameters from Table 1 and investigate the effectiveness of three containment strategies: 1) reducing contact rate to live poultry ($\beta_{HB}$), 2) slaughtering infectious live poultry ($I_B$), and 3) quarantining infectious humans ($I_H$). Each strategy is assumed to be in effect starting on February 20, 2010. Without any intervention, the total infection reaches 540 and will be contained by June 2010. Figure 3(a) shows that reducing contact rate by 20% will reduce the total infections to 453 (16% reduction), although it does not significantly improve the time of containment. However, the containment will be achieved by May 2010 with 278 infections when the contact rate is reduced by 60%. This shows that reducing contact rate is highly effective in minimizing the total infections and containment time. This strategy is also easy to implement: residents should be advised to avoid contacting birds or their saliva, mucus and feces; and poultry farm workers should be advised to use proper protective equipment while working close to the poultry.

Figure 3(b) suggests that slaughtering infectious live poultry is equally effective. It has similar efficient frontiers as reducing contact rate in controlling human infections. However, it also controls the outbreak within the live poultry. Unlike reducing contact rate, which works totally independent of the bird compartments, reducing infectious poultry also reduces the danger of susceptible birds being exposed to the virus. This leads to earlier containment within the live poultry. However, there are practical difficulties with this approach: the management of the bird carcasses and the potential economic impact. Incineration of the dead birds has been a common practice.
Figure 3(c) shows that with the current baseline infection among humans at $1.5 \times 10^{-7}$, the effect of human-to-human infection is negligible: all 5 curves are almost the same. But this could become a threat due to the high mutation probability of avian influenza viruses. Mutant species which can be easily transmitted among humans may emerge; hence epidemiologists and healthcare leaders should closely monitor such a possibility.

Figure 3(d) shows the combination effect of reducing contact rate and reducing infectious poultry. When both strategies are implemented at 20% level, the total number of infections at containment is approximately 400. Since implementing a single strategy at a high effective level may be inherently difficult, combination strategies are appealing and can generate satisfactory containment results.

A significant concern to public health officials is to understand when the intra-human infection of avian influenza will pose a major threat to public health. By varying the baseline infection rate between humans ($\beta H p_H$) while keeping other parameters fixed as given in Table 1, Figure 4 shows that the trend of avian influenza outbreak in human remains insignificant, unless the baseline infection rate is increased from $1.5 \times 10^{-7}$ to 0.15. Although such a possibility is small, it will result in more than 5,400 infections in humans at containment, 7 times more when compared to the original scenario. Since the human mortality rate is very high, it is imperative that epidemiologists and healthcare officials are vigilant in monitoring the mutations of avian influenza viruses, and take proper and timely actions should such high infectivity mutations occur.

**United States.** In this scenario, besides the factors discussed in the previous section, we also explore the impact of transportation and migration of infectious birds on disease trend of an avian influenza outbreak. For brevity, we only consider infections among birds. To model the change of bird population in each state due to commercial transportation or migration, we introduce a transportation matrix $\Omega_{N \times N} = (\omega_{ab})_{1 \leq a,b \leq N}$, where $\omega_{ab}$ represents the percentage of birds transported from state $a$ to state $b$ in a unit time, and $N$ is the total number of states considered in this setup. Each state is assigned a unique instance of equation (2-1) with its own initial bird populations, and the transportation matrix $\Omega$ is used to connect all the states. More rigorously, let $Y_{m \times N} = [y_1, y_2, \ldots, y_N]$ be the collection of stage vectors of all
states, where $m$ is the number of total compartments for birds in formula (2-1). Then the ordinary differential equation with the transportation and migration of birds can be written as

$$Y' = (D_B - I)\text{diag}(\mu_B)Y + \sum_{k=1}^{N} \text{diag}((D_B + I)\text{diag}(\beta_B)Y_k)C_BY_k\epsilon_k + \lambda_B(Y) + Y(\Omega - \text{diag}(\Omega \cdot 1))$$

where $\epsilon_k$ is the $k$th basis vector. In this scenario, we assume that the initial infection starts in Ohio with 30,000 infectious birds. If no restrictions are placed on the transportation of live poultry and no interventions are implemented, the spread of the virus is shown in Figure 5. By day 100 of the outbreak, the virus has already spread to neighboring states and most of the northeast region. By day 300, the virus has spread to the midwest and southern regions, and the trend of the outbreak is still increasing, as shown in the light green solid curve in Figure 6. The cumulative number of infections in birds in the entire United States will reach 1.8 million by day 300, and some states with high poultry production will be severely affected, causing potential dangers to the health of poultry farm workers as well as residents.

Figure 4. Sensitivity analysis of baseline infection rate between humans

We will examine two intervention strategies for this scenario: 1) reducing the probability of transporting infectious birds to other states by enforcing strict regulations and examinations on the transportation of live poultry and/or poultry products; 2) reducing the population of infectious birds by slaughtering infectious poultry and wild birds. Both strategies are assumed to be implemented one month after the initial outbreak.

Figure 5. Outbreak trend in US at different number of days after the initial outbreak; the colors refer to the number of infectious birds in each state (upper panel: real-time infections; lower panel: cumulative infections).

Figure 6(a) shows the effect of reducing transportation of infectious live poultry and controlling the migration of infectious wild birds. Without any intervention, the outbreak will start to show signs of containment by day 700, and the
total number of infectious birds at containment will be near 2.6 million. If a 20% reduction of infectious transportation is achieved, the containment will start on day 200, and the total infections in bird population at containment will be 1.1 million, a 58% reduction. If this strategy is implemented at 80% level, the outbreak will start to be contained by day 120 (almost one and a half years earlier compared to no intervention), and the total infection in bird population at containment will be 0.6 million, a 77% reduction. This result demonstrates that by imposing strict regulations and examinations on live poultry transportation, the containment will be reached earlier, and the total number of infected birds will be reduced significantly. Another advantage of this strategy is that it can confine the outbreak within a small region so that it is easier to implement other types of interventions.

**Figure 6.** Contrast of cumulative number of infected birds in US under two containment strategies.

Figure 6(b) shows the effect of slaughtering infectious birds one month after the initial outbreak. Although slaughtering infectious birds alone is not as effective as reducing transportation and does not change the disease trend significantly, it does reduce the number of infections at containment. For example, when 40% of infectious birds are slaughtered after one month of initial outbreak, the cumulative number of infections at containment will be 1.67 million, a 36% reduction compared to no intervention. The number of infections at containment will be 0.74 million, a 72% reduction, when 80% of infectious birds are slaughtered. However, this strategy has both environmental and economic concerns, and the disposal of carcasses of slaughtered birds remains a difficult problem to solve. Combination of containment strategies will be effective and promising, as shown in Figure 7. Implementing both strategies at 20% level will reduce the number of infections by containment to 0.95 million, a 63% reduction. When both strategies are implemented at 30% level, the cumulative number of infected birds will be 0.75 million, a 71% reduction.

**Figure 7.** Illustration of the effect of combination strategies in US.

**Designing A Digital Disease Surveillance Response System (RealOpt-ASSURE).** We develop a digital disease surveillance response system for infectious disease outbreaks within the architecture of RealOpt-Regional. In addition to the global operational and response capabilities inherited from RealOpt-Regional, RealOpt-ASSURE has three major functionalities: 1) recording and displaying the current reported cases, 2) predicting future infections based
on the transportation matrix \( \mathbf{A} \), and 3) determining the optimal number of partitions (point-of-dispensing sites, PODs) to dispense medical countermeasures to the affected poultry farms. To achieve real-time decision capability, the computationally intensive general-purpose disease model within RealOpt-ASSURE is approximated using a simplified inclined decay with an exponential adjustment (IDEA) model\(^{13}\).

Based on the geographical distribution of current infections, the migration patterns of wild birds, the commercial transportation of live poultry, and the location of large poultry farms which are most likely to be the victims of avian influenza outbreaks, we apply RealOpt-ASSURE to determine the operational requirement for rapid and effective containment. Figure 8(a) shows the heat map of infected sites during the 2015 avian influenza outbreak in Iowa. Figure 8(b) highlights the optimal partitions for dispensing medical countermeasures to protect the poultry farms in one of the affected locations. After establishing the dispensing partitions, decision makers can design the processes and optimize the throughput with minimum healthcare resources using RealOpt-POD\(^{34,35}\). Point-of-dispensing for human population protection can be established as walk-through or drive-through facilities or through mobile vehicles. For the live poultry, dispensing should be performed simultaneously across all partitions to ensure the maximum and most effective coverage.

With the support of the RealOpt-ASSURE digital surveillance tool and the RealOpt enterprise systems, emergency planners can prepare in advance and respond rapidly (and improvise if necessary) to local avian influenza outbreaks to achieve early disease mitigation and containment for both human and bird populations.

**Figure 8.** Screenshots of the digital surveillance tool, RealOpt-ASSURE, designed on top of RealOpt-Regional\(^\circ\). Figure 8a shows the 2015 H5N1 infected sites in Iowa. Figure 8b shows the partitions for dispensing medical countermeasures to protect live poultry. The icons represent the walk-through or drive-through point-of-dispensing sites for dispensing medical countermeasures to the human population if officials decide to initiate prophylaxis.

**Discussion**

This work describes the design of a “digital disease surveillance and response” system, RealOpt-ASSURE, in preparing for and responding to pandemic outbreak. RealOpt-ASSURE is a general-purpose modeling framework that couples compartmental models (with separate groups of compartments for human and carrier populations) with human behavior and operations processes. We demonstrate its usage in analyzing avian influenza. The RealOpt-ASSURE modeling framework is capable of tracking the development (and interplay) of epidemics in both human and carrier groups. It also incorporates the natural population growth and effect of migration and transportation. We analyze avian influenza outbreaks and investigate parameters that have significant impact on epidemic severity. We study the effect of average contact rate between birds and average time that birds stay infectious.

Analyses are performed using the 2010 H5N1 outbreak data in Egypt. Our analyses show that the trend of avian influenza outbreak is sensitive to both contact rate and infectious duration. Accurate parameter estimation of the biology of the avian influenza viruses is crucial to the model performance. We explore three containment strategies to mitigate and minimize infections among humans: 1) reducing human contact rate to birds, 2) slaughtering infectious birds/live poultry, and 3) quarantining infectious humans. The first two strategies show similar effectiveness in containing the outbreak and reducing the number of infections in humans; while quarantining infectious humans has marginal effect. This confirms that reducing human-bird contact and controlling infection among birds are key factors in containing the outbreak. The human-to-human infections effect of avian influenza is low; however, it is possible that mutations may lead to more significant intra-human infections.
Analysis is also carried out using the poultry farm and production data in the United States. Two containment strategies are contrasted: 1) reducing transportation or migration of infectious birds, and 2) slaughtering the infectious bird population. Both strategies are effective, with the former confining the outbreak within a localized region for more manageable containment operations. These findings indicate that regardless of the objectives, the outbreak in the bird population should be tackled first, and the containment strategy should be determined based on environmental and demographic information, as well as transportation/movement and bird migration patterns.

The RealOpt-ASSURE system enables decision makers to track the outbreak status, and predict possible future disease trends to assess its severity. It provides a decision support framework for policy makers to estimate the potential risk of an outbreak and contrast the cost-effectiveness of different containment strategies. Combination strategies are practical and cost-effective in achieving a disease-free equilibrium. Healthcare officials should select a strategy portfolio based on the local environment, demographics, and the composition of bird population. To further expedite the containment, the public should be educated and informed of the knowledge regarding avian influenza, as human behaviors have a profound impact on the virus transmission.

This modeling framework is highly generalizable and flexible. The modeling framework does not depend on the mechanism of the disease transmission or the biological development. It can accommodate various setups for modeling other pandemic outbreaks. For example, the model can accommodate the spatial distribution and movement of humans for global trend estimation. Other interactions as a result of human behavior, social media, and/or public policies can also be incorporated. It can also handle heterogeneous groups. For example, formula (2) consists of two sets of equations: one for the human population, and one for the bird population. Another group can be incorporated readily by adding a different set of equations with properly defined contact and disease contagious matrices.

This modeling framework can also be used for general vector-borne diseases. In Lee et al., we demonstrate its first usage for Zika virus analysis. The Zika analysis includes both human and alternate hosts.

A closer look at formula (2) shows that it only models the infection from birds to humans, but not the reverse. This reflects the reality of avian influenza. In general vector-borne diseases, infection from humans to vectors is possible. In this case, we can add the corresponding term to the carrier equations by properly defining the reverse disease contagious matrix $C_{BH}$ and contact rate vector $\beta_{BH}$. Then the equation for carrier population (2-1) becomes

$$y_H' = (D_B - I)\text{diag}(\mu_B)y_B + \text{diag}((D_B + I)\text{diag}(\beta_B)y_B)C_By_B + \text{diag}((D_B + I)\text{diag}(\beta_{BH})y_B)C_{BH}y_H + \lambda_B(y_B)$$

Combined with the equations for human population, we obtain a modeling framework for general vector-borne diseases, which can be used to model ZIKV, dengue, etc. Our previous model for ZIKV can be perfectly represented in this framework. As more biological facts of different pandemics are discovered, new stages and updated parameters can be fed into RealOpt-ASSURE continuously instead of building the entire model from scratch. This feature is particularly appealing in real-time operations and decision making, as rapid implementation of appropriate containment strategies can drastically change the overall disease trend and the outcome of an outbreak.

Informatics and social media play an important role in tracking disease spread and supporting epidemiology modeling and containing disease outbreaks. They have already demonstrated their capability in helping prevent influenza outbreaks. The most challenging part is to recognize how classical mathematical models fit and perform in this digital era. RealOpt-ASSURE allows incorporation of different levels of data within the modeling environment. It is also highly sustainable with the support of modern information technologies, which provide valuable inputs to our model. With continuous data input, the model can adjust the parameter estimations to better predict the potential outcomes of pandemics, reflect changes in the environment, human behavior, and public policies as the outbreak develops. It can provide meaningful insights to the containment strategies in real-time to better facilitate the decision-making process. It can also be adapted readily to accommodate newer disease propagation structures, additional stages introduced by intervention or hospitalization, and changes in the pattern of disease transmission, making it sustainable both theoretically and practically.

There are limitations to our study. Although the modeling framework we proposed is flexible and capable of modeling complicated scenarios, we used a relatively simple setup in the examples with homogeneous bird and human populations without any treatment or vaccination effects. Different bird species should have different infectivity rates for avian influenza, and the contact rate to birds differs between poultry farm workers and general humans. Antiviral drugs and vaccines against avian influenza also exist, and could be added to the model. These components require more parameters to estimate and may possibly lead to overfitting. However, with the availability of such data, these effects can be included in our modeling framework with additional compartments in both carrier and human populations.

Reference:

Exploring the Design and Role of Mobile Apps for Healthcare Providers to Find Teratogenic Information

Lily Lei¹, Vishwas Shetty¹, Karan Gupta¹, Janine E. Polifka, PhD¹, Glen Markham, MBA², Sarah Albee², Carol Collins, MD¹, Gary Hsieh, PhD¹
¹University of Washington, Seattle, WA; ²RightAnswer.com, Midland, MI

Abstract

Healthcare providers (HCPs) caring for pregnant patients often need information on drug risks to the embryo or fetus, but such complex information takes time to find and is difficult to convey on an app. In this work, we first surveyed 167 HCPs to understand their current teratogen information-seeking practices to help inform our general design goals. Using the insights gained, we then designed a prototype of a mobile app and tested it with 22 HCPs. We learned that HCP’s information needs in this context can be grouped into 3 types: to understand, to decide, and to explain. Different sets of information and features may be needed to support these different needs. Further, while some HCPs had concerns about appearing unprofessional and unknowledgeable when using the app in front of patients, many did not. They noted that incorporating mobile information apps into practice improves information access, can help signal care and technology-savviness, in addition to providing an opportunity to engage and educate patients. Implications for design and additional features for reference apps for HCPs are discussed.

Introduction

Uncertainty about drug safety for pregnant women is a significant problem. Birth defects are common and affect approximately 120,000 children in the United States each year.¹ There is a high incidence of comorbidities such as diabetes, depression and autoimmune diseases in women of child bearing age. Thus, it is not surprising that nine out of 10 women are prescribed one or more drugs during pregnancy.² The use of prescription medications during the first trimester of pregnancy has increased more than 60% over the last three decades².

Healthcare providers (HCPs) who care for pregnant women are concerned about the quality and accuracy of the information that is available to their patients. If the information is inappropriately frightening, women may become unnecessarily anxious about the health of their fetuses and avoid taking a medication that actually improves the likelihood of a successful pregnancy. In extreme cases, a woman may unnecessarily terminate a normal, wanted pregnancy out of fear that an exposure may have harmed the fetus³. In other circumstances, a pregnant woman may have a teratogenic exposure that could have been avoided if she or her HCP had understood the hazard more fully.² Thus, those who care for pregnant patients need to stay abreast of current teratology information treat their patients’ illnesses effectively with minimal risk to the fetus, and address their patients’ concerns appropriately.

Keeping up with the constant flow of medical research can be a daunting task for any HCP because of time constraints. Mobile devices have become an indispensable tool for HCPs in the management of disease, health data sharing and patient communication. Their portability is perfectly suited to the vagaries of the clinical setting, where HCPs can access information resources at the point-of-care (or from anywhere) any time of day. Mobile devices can provide HCPs with immediate access to evidence-based information when access to personal computers is not always feasible, such as in exam rooms or visits in clinics. But would these apps be merely mobile ports of existing teratogenic risk databases? How might we leverage mobile devices to help meet the information needs of HCPs within the time constraints of a clinic visit? Further, while much prior research has examined how the use of Health IT affects patient-provider interaction⁴⁻⁸, relatively little is known about how providers perceive and adapt mobile apps for information seeking in front of patients. To answer these questions, we (1) conducted an online survey of HCPs evaluating their teratology information needs and their use of mobile apps for seeking this type of information, (2) developed a prototype mobile application based on what we learned from the survey, and (3) tested our prototype by asking local HCPs to think aloud while completing four clinically-relevant tasks.

Background: Pregnancy Drug Risk Information Seeking

HCPs use a variety of resources to stay current on drug information, such as books, colleagues, the primary literature, online databases and internet search engines. Large electronic databases such as Micromedex®, UpToDate®, Epocrates®, and Lexicomp® are used by many HCPs to obtain information about the risks associated with the use of drugs during pregnancy. These databases provide an evidence-based summary of the available literature on the teratogenic risks associated with the use of medications during pregnancy. But this
information is a subset of the broader clinical information and therefore does not always provide a comprehensive review and evaluation of the teratology studies that have been published. Moreover, the expertise of the individuals who have summarized the teratogen information for these databases is not always transparent to the consumer. These large databases are distributed on a paid subscription basis to major hospital and university libraries; so HCPs who do not have affiliations with these institutions are limited in their ability to access this information.

A few databases exist that are designed specifically for HCPs with pregnant patients (e.g. Briggs Drugs in Pregnancy and Lactation®, Reprotox®, Shepard’s Catalog of Teratogenic Agents®, and TERIS®). These databases provide an expert review and comprehensive, evidence-based summary of the fetal risks associated with the use of specific medications during pregnancy and lactation. However, most HCPs are not aware that these databases are available through some of the larger publishers, such as Lexicomp®, Micromedex®, RightAnswer.com®, and ToxPlanet® or that they even exist.

We utilize data from the TERIS® (Teratogen Information System) database for the primary content in our app. TERIS® is a computerized, peer-reviewed database designed to assist physicians or other HCPs in assessing the fetal risks associated with exposures to drugs or chemicals in pregnant women. The database consists of a series of agent summaries, each of which is based on a thorough review of published clinical and experimental literature.

**Background: The Use of Mobile Apps in the Clinical Setting**

A 2012 survey by Manhattan Research suggests a high adoption rate of smartphones for US physicians: 81% of those surveyed own a smartphone, and 87% of them use a smartphone or tablet device in their workplace. One of the key uses of mobile devices and apps is for reference and information gathering.

However, despite their increased usage and potential, a 2012 review paper on smartphone usage among physicians and students in medicine concludes that “very few high quality-studies exist to help us understand how to best use this technology [for physicians and students].” Much of existing studies about smartphone usage in the health context focus on smartphone use for telemedicine or from the patients’ perspective for self-tracking, self-care and communication with providers (e.g.,). There are much fewer studies on HCPs’ use of smartphones for decision-making in the clinical setting, and specifically there are “very limited data” on the use of reference apps by HCPs.

Research on more traditional forms of technology does suggest that the use of smartphone apps for HCPs can improve the quality of care. Studies on the use of PDAs for data management have found that PDAs facilitated the access of patient information during ward rounds. A study on nursing student’s use of PDAs also found that “the use of the PDAs in clinical practice enhanced communication skills and contributed to the quality of care they delivered.” It both enhanced their ability and their confidence to provide clear and quality information, quickly. Recent research on the use of other Health IT, such as EMR has also generally shown positive outcomes from use, especially supportive in terms of information related tasks. However, one important drawback is that their use had a negative impact on patient centeredness; reducing eye-contact and losing rapport with patients. Insights to address these drawbacks on desktop usage include improving HCP’s technology skills and changing their behavioral style. Others have also noted that collaborative viewing, or showing the screen to patients, can create a “common information space” to facilitate interaction between patient and doctors.

**Methods: Survey**

We began our work by developing and deploying a survey targeted to HCPs. The goal of this survey was multi-fold. First, we sought to understand HCPs’ teratogen information needs and behaviors when it comes to prescribing medication for pregnant women or counseling those who have been inadvertently exposed. In addition, we sought to explore the use and role of mobile apps for teratology information seeking by HCPs.

For the open-ended questions, three of the authors who have expertise in human-centered design and clinical teratology, analyzed the responses separately to generate category codes. They then discussed each of the responses and followed an iterative process of applying open coding and axial coding to discover relationships among emerging concepts until a consensus is reached.

**1. Recruitment and Participants**

We emailed our survey link to members enrolled in the Washington, Wyoming, Alaska, Montana and Idaho medical education network, the Washington Academy of Physician Assistants, Washington State Obstetrical Association, the Midwives’ Association of Washington State, the Washington State Perinatal Regional Network, Washington State Pharmacy Association, University of Washington Department of Psychiatry and Behavioral Sciences,
MotherToBaby of Arizona, MotherToBaby Utah, and Department of Health and Emergency Medicine Services in Colorado. Initially we offered $10 gift card for completing the survey. We raised it to $20 to increase participation.

A total of 225 surveys were initiated, out of which 167 answered all the questions (74%). 115 HCPs were female (69%), and the mean age was 39 years old, with a range from 23 to 70 years of age. The mean number of years in practice was 11.5 years, with a range from 0 to 43. The HCPs’ specialties included retail and clinical pharmacists (64%), physicians (13%), midwives (7%), physician assistants (6%), nurse practitioners (4%), and others (6%). The majority (63%) of pharmacists surveyed were specialty pharmacists who practice in medical settings and focus on therapy for patients with complex disease states.

**Results: Survey Findings**

When asked about their most recent incident where they sought teratogen information, the five most common inquiries were related to infectious (19%), psychoactive (14%), gastrointestinal (8.3%), pain (7.2%), and cold (5.3%) medication. To seek the information, the vast majority of the HCPs used online databases (86%), such as Epocrates®, Micromedex®, or UpToDate®. Others used textbooks (11%) and less frequently manufacturer’s website and package inserts (2%), and other clinicians (1%).

Almost half of the HCPs surveyed expressed dissatisfaction with their recent information searches (45%). The biggest problem noted was that the resources did not provide enough information. For example, “I wasn't very happy with the online databases that I have access to. They don't give much beyond the pregnancy category.” Relatedly, others noted being dissatisfied with having to access information that was spread across different sources. “There is no single, trustworthy, up to date resource for them to consult when it comes to drug use in pregnancy”. Usability and speed were also other issues noted by respondents. “Didn't like the layout in Briggs/Lexicomp - class effect data were emphasized over specific agent. Also - references had conflicting data.” And “info limited and not presented in easy format to understand. I think pregnancy categories should not be eliminated but refined.”

1. **Teratogenic Information Needs.** As part of the survey, HCPs were presented with a clinical scenario about a pregnant patient who is being treated with escitalopram and is concerned about its risks on her baby. HCPs were given the TERIS database output for escitalopram and were asked about its content. We found that participants also desire to have additional information on alternative drugs, breastfeeding, dosage, maternal disease and actionable recommendations. For our prototype, we incorporated this additional information into the mobile app.

2. **Usage of Mobile Apps for Drug Information.** In the survey, we also asked about the use of mobile devices to search for teratogen information. About 67% (112) of our respondents have used mobile apps to find teratogen information, with the most popular app being UpToDate® (53%), followed by Epocrates® (38%).

When asked about the major limitations of existing mobile apps, 106 responded. The issues reported were similar to what they had reported with their most recent teratogen information seeking experience. The main issues (52% of responses) were related to the amount and quality of information offered in existing apps. As one respondent wrote:

> Epocrates only gives an overall summary - no specific risks; Reprotox® can sometimes be hard to digest quickly if there's a lot of research; UpToDate® doesn't always list quality of evidence and is often the recommendation of the author based on studies, which, in my experience, don't always support the conclusions drawn.

Many simply desired a centralized database of information across pregnancy and lactation resources that contain concise summaries in addition to more in depth information and links to primary resources.

The second set of limitations was related to usability (22%). Many of these respondents noted challenges with finding and navigating information via a small screen, and a desire to have more intuitively organized content. Others pointed out that logging into the app and having it sync and update often takes too much time. The third set of limitations had to do with access and cost (18%). Some interesting points to note are: Mobile apps are not always supported by existing practices, so they are often out-of-pocket costs for HCPs; not all apps are available on all phones; a lack of access to the internet when the information is needed.

3. **Information Seeking in Front of Patients.** While we were initially concerned that HCPs would not be comfortable with using mobile apps in front of patients, we found that about 60% of participants were not concerned. In fact, more than a third of the HCPs (37%) reported that their most recent teratogen search was conducted in front of patients. When asked about specific concerns with using the app in front of patients, 65 HCPs responded. For them, the primary concerns are not appearing to be professional (40%) and being perceived as
incompetent (34%). A small number (9%) also mentioned that doing so may seem like they are doing something else, such as checking personal email and some (8%) are concerned that they are not sending the right message to the patients – “Patient might feel like it's something they can do as well.”

4. Survey Discussion

In general, our survey provided confirmation that an app dedicated to teratogen information would be valuable to providers. Both having a consolidated information source and disseminating this information in the form of a mobile app are useful. While there are some concerns about appearing to be unprofessional and unknowledgeable, many seem to be already incorporating the use of mobile apps into their clinical practices.

At the same time, these results offer insights on the types of information that should be included in an app, as well as indicating the need for effective information architecture and user interaction to support the often descriptive information about teratogens. Combining our survey findings with the new FDA guidelines that sought to minimize the oversimplification of teratogenic risks, we decided to focus on the following goals in design: (1) consolidate pregnancy and lactation drug information across multiple sources to create a single repository; (2) provide a concise summary and overview of teratogenic risk information without oversimplifying risks; (3) reference sources of information; (4) have an easy to use interface with navigation and search support to move between overview and detailed information; (5) provide author or editorial information to indicate the credibility and reliability of the information; and (6) provide patient education handouts.

Mobile App Prototype

After a wide exploration of ideas through sketching, followed by wireframes in increasing fidelity, we developed an interactive prototype design using Justinmind26, a prototyping platform.

The app’s homepage shows a list of drugs with an icon denoting the teratogenic risk (Figure 1a). Users can “search by drug” through an alphabetized list of all agents, or “search by condition” through a list of maternal conditions that will lead users to an alphabetized list of agents used to treat the selected condition. The drugs can also be filtered by their risk classification; “none,” “unlikely,” “minimal,” “small,” “moderate,” “high,” “undetermined,” and “variable.”

For each drug, we used the TERIS® database for the primary teratology content in our prototype. This includes brief information About the Drug, Risk Information and Comments, Research Findings, and References. To help HCPs fully explore the risks and the benefits of treating a pregnant patient, we added a section called “Clinical Considerations.” This section provided the following information (if available): Impact of disease on pregnancy; Dose adjustments in pregnancy; Drug-associated adverse maternal reactions unique to pregnancy; Fetal/neonatal adverse reactions; and Drug effects during labor and delivery. Further, using feedback from the survey, we incorporated information on breastfeeding, dosage, and a fact sheet for the patient. The individual sections use an expandable accordion design pattern to enable users to navigate to desired information quickly (see Figure 1b). To
help users assess the credibility and currency of information, we indicated when the drug information was updated right next to the “About the Drug” heading for each drug (the first line).

Another desired feature based on our surveys was for information about alternative medications that could be used to treat a particular medical condition. Thus, in our design, users can toggle between the summary or the alternative drugs screen, which features a full list of alternative agents to the current drug selected (Figure 1c).

One of the key challenges in our design is to effectively communicate the different teratogenic risk classifications used in TERIS®. The risk classifications for teratogenic effects in the children of women exposed to the agent during pregnancy range from “none,” “unlikely,” “minimal,” “small,” “moderate,” to “high,” and “undetermined” for agents that have an unidentified risk. For some drugs, the risk was considered to be somewhere between two ratings, e.g., drug may be assessed to have a “moderate to high” teratogenic risk if used during pregnancy.

To visually communicate these risks, we developed a color-coded system for the teratogenic risk classifications. Our final design utilizes a gradation of red circles. Red was chosen since it is commonly associated with danger. The higher the risk the more saturated the red is and the lower the risk the lighter the color. For the remaining risk classifications; “none” is represented with a white circle to symbolize no risk, “undetermined” is represented with a gray circle to symbolize unknown risk. (Figure 1e).

However, the risks for some drugs required more complex representations. For example, some of the drugs have different risk assessments depending on the nature and timing of the exposure during pregnancy, such as stage of pregnancy, dosage levels, or route of exposure. Other drugs had risks that fell between two gradations of risk (e.g., between “moderate to high”). For these, we developed a single additional symbol for “variable.” Our design for “variable” is represented with three overlapping gray circles to symbolize that some drugs may have more than one risk. We thought that this single classification would be sufficient to serve its main goal, which is to notify the HCPs that the risks for this drug may be complex and they need to read further to determine if the drug’s risk applies to their particular patient’s exposure (Figure 1d).

Methods: User Study

We conducted a user study to evaluate our app design. The study consisted of a 30-minute usability testing of the prototype, where we observed participants performing 4 tasks using our prototype. We then followed up the usability testing with a 30-minute interview. The interview enabled us to more broadly talk about the app and its potential for integration into their practice workflow.

For the usability testing, we asked the participants to complete four tasks using our prototype. These tasks were common clinical scenarios that we developed based on the survey results. The scenarios provided detailed information, and asked participants to (1) review five antibiotics and select one that might have less risk to the fetus, (2) review evidence-based risk information for escitalopram for a pregnant patient, (3) find a list of drugs that have a high teratogenic risk and should be avoided during pregnancy, and (4) evaluate the risk and benefits of using ondansetron or an alternative medication to treat hyperemesis gravidarum. Each of the tasks took about 5 minutes to complete. During the tasks, the participants were asked to vocalize their thoughts, feelings, and opinions while interacting with the prototype (think-aloud). After each task, they were asked a few questions about the task. Upon completion of the four tasks, participants were asked to fill out a questionnaire which contained the System Usability Scale (SUS), which is a ten-item attitude Likert scale often used to assess usability.27 The questionnaire also included questions about participants’ demographics.

Throughout the study, we audio recorded the participants think-aloud sessions, as well as screen-captured their interactions with the mobile prototype. For analyses, we first had the study sessions transcribed, then we used the similar procedure in analyzing the open-ended survey responses.

1. Participants

HCPs from our survey who were local and expressed interest in participating in the app testing were contacted for the study. In addition, we contacted HCPs through the Washington State Department of Health and by word-of-mouth through some of the authors’ colleagues.

Twenty-two HCPs who often worked with pregnant patients were recruited to participate in our usability study. Their ages ranged from 27-63 years. The number of years they have been in practice ranged from 1-35. Their specialties included: Family Medicine physicians (7), Naturopathic physicians/nurse practitioners/midwives (6), Medical geneticists/genetic counselors (3); Obstetricians (3); Pharmacists (2); Psychiatrist (1).
**Results: User Study Findings**

Overall, participants completed the tasks with ease. When asked “how easy or difficult was it to find the information you were looking for” after each task, all of the tasks had a median rating above 4 (1 is very difficult and 5 very easy). The most common reason participants gave for assigning a low score to a task was “unfamiliarity with the app”. Once participants became familiar with the app they found the subsequent tasks much easier to complete. The overall SUS score of 80.8 is considered in the A level (the top 10% of SUS scores). It is important to note that participants reported that the four tasks used in the study represented commonly encountered clinical scenarios.

In terms of content provided, participants appreciated the consolidated information. “I like this [app] because the information is consolidated, and there's less noise. You're not looking at other stuff. It's focused on pregnancy.” (P5). They specifically complimented us on the inclusion of lactation information, patient Q&A handout, and providing the alternative drug list. Many of them asked when the app will be available for download.

1. **Three Types of Risk Information Needs**

In terms of risk classification designs, participants thought the design from lighter to darker was fairly intuitive, and were able to describe what the different colors meant, after some usage:

*Having looked at it a few times, the dark red I know means high risk, the three little bubbles I know is variable. It makes intuitive sense to me lighter to darker. I think it would just take some [time] to [get used to] it.* (P3)

However, some participants noted that the different levels of red were not memorable and sometimes hard to distinguish. Some also thought the TERIS® risk classifications were unclear since they were unable to discern the subtleties in gradations of magnitude of teratogenic risk, e.g., between “minimal” or “small.” Many even specifically asked us to include the FDA Pregnancy Categories that have been discontinued by the FDA, not realizing the change in guidelines. One potential interpretation is that HCPs need more time to adjust to the TERIS® risk classifications and a clear legend or onboarding should be added to the app to help. However, we believe there is a deeper underlying issue. There is a mismatch in what we sought to design (an informational tool), and what the HCPs generally need (an easy to use decision aid). In fact, the HCPs indicated three types of risk information risks.

**To Understand.** One of the key goals of teratogen information seeking is to learn and understand the fetal risks associated with the use of a drug during pregnancy. Indeed, in our scenarios, some participants spent a lot of time reading through the detailed teratology information as they were completing the tasks. When asked about the amount of information presented and whether they were overwhelmed, one said: “If I’m looking it up, I want that much information” (P2).

This level of detailed understanding is often needed when HCPs are exploring drugs they have not used before or if the preliminary results from searching contradict their preconceptions or when they are counseling a pregnant patient about her chances of having an adverse pregnancy outcome.

*If I am trying to counter what I normally do, like what’s my custom and practice or If I am having a question about why I wouldn’t do the things I would normally do, then I want more information than I already have. That’s primarily how I would use something like this, for something that is not already in my head.* (P4)

**To Decide.** But to understand is not the primary use for the app. Most HCPs primarily perceived the app as a decision aid. This can be seen in how the HCPs approached our scenarios. When asked to prescribe one drug out of a list of five that poses less risk to the fetus, many participants simply started with a drug they know from experience that has a minimal risk, and stopped immediately when the app confirmed that drug’s risk is minimal. They tried to base their decisions simply on the risk ratings and risk comments when possible. As one participant said:

*As a practitioner or provider, not necessarily the pharmacist, but a doc will look at this and only want to know, is the risk information, general risk data, good. And honestly that’s the bottom line. Then they’ll read more if they had time. But if they’re the exam room with the patient, they just want to know what’s the risk, what’s the data... You know, is there data, and if there’s not very good data or something like that, then they’ll go ahead and read more... Right there it tells me, minimal risk, good data, and I’m sold. They would be, too.* (P1)

This decision-aid view of risk visualization also helps explain why some of the participants thought the eight different types of risks were excessive, and that varying the gradient of the red was not an optimal design. With the
8-level risk ratings as offered by TERIS®, HCPs may perhaps gain more nuanced understanding for the differences in teratogenic risk across drugs. However, for the purposes of deciding what to prescribe, HCPs tended to think in terms of three distinct tiers of risk – safe, not safe and maybe safe. In which case, having three distinctive types of colors might be more effective (e.g., green for safe, red for not safe, and yellow for maybe safe). One HCP said: “It’s almost 3 categories for me. I think of medications as ones that are safe and can be prescribed, ones that are not safe and ones that depend on the situation” (P15). Another: “In my mind, I lump minimal, unlikely, and none all together…I would clinically use them with the same frequency” (P3).

That is also how they used the outdated FDA Pregnancy Categories, which had five levels, A, B, C, D, & X. Some HCPs felt that any drug assigned an A or B category was safe to prescribe in pregnancy and others considered drugs in categories A, B, and C to be safe. For example, one HCP said “once you saw A and B, you were good to go and didn’t have to fuss much. (P6)” Whereas another said “we know in our minds X means teratogenic, and then pregnancy category D…might be some harm and so have to look at risk benefits, and then ABC we’ll all lump together and think it’s all fine. In our heads, we think it’s a quick way to separate out the drugs. (P3)”

To Explain. We also found that an important third usage of the teratogen information is to help HCPs explain their recommendations to the patients. Some of the HCPs would use the information in the app directly to corroborate what they are saying to the patients. “[Using it] demonstrates to the patients how I approach a scenario as a clinician and make lifestyle changes to benefit their health in general. It also helps to validate what I’m saying and not that I’m just saying stuff” (P14).

At the same time, HCPs also talked about the need to translate the information to terms that patients can understand. The detailed information provided can help the HCPs understand the teratogenic risks associated with the drugs themselves, but patients have a difficult time weighing the risks with the benefits when the risks are uncertain or the data are difficult to apply to their particular circumstances. For example, when HCPs saw “odds ratio” presented in the app, they wondered: “what’s a number that I can translate to a patient so they can understand” (P3). They also talked about the benefit of having clear descriptions of risk ratings as it “gives a little bit of language when you are explaining to a patient” (P4).

2. Benefits and Challenges of Mobile Information Apps in Clinical Settings

Another question we sought to answer in this work is whether a mobile app can provide additional value to providers over existing databases that are accessible through desktops. Through our study, we found a number of potential benefits to having teratogen information on an app.

Access. Prior research has found that the use of personal digital assistants (PDAs) can improve information access for HCPs. Similarly, when asked about potential uses, HCPs conveyed that having the information on the app enabled more flexibility in use. One example is using it during in-home visits. “I attend out of hospital births and do in-home visits. Which is another reason that [having] many of these tools not on a laptop is helpful” (P4). HCPs also talked about using the app as they moved between rooms at the hospital since it can easily fit in their pockets.

Another set of benefits related to access is the speed and efficiency of use. Many HCPs indicated preferences for using the app for information seeking even when other machines were within access. They argued that the app will be faster to use without having to move in and out of other applications already running on the other machines. It also helps keep the different machines and their tasks separate, e.g., keeping the desktop dedicated to EMR.

It all comes down to speed and efficiency. If a question comes up while in a room with a patient, I have to be able to quickly access [the information]. I am just such an app person. I am just much more inclined to use an app than a website. (P3)

I have my EMR, schedule, and my email going; a lot of time I’ll open a separate function on my phone so I don’t have my EMR shut down. Also, because I have my phone, apps are pretty easy to use on it. (P4)

Technology Savviness. Another potential benefit of using a mobile app for information seeking is that it can demonstrate technology-savviness. As one HCP said: “I used to worry that they are going to think I’m not smart enough. Now I think patients will think you are more tech savvy if you’re able to show them where your data sourcing is and [is] current” (P3). And as another pointed out, “Using an app looks better than having to get the textbook out” (P8). However, there are a couple of important caveats to note. One is that this perception of technology-savviness may depend on the patient population. As P3 also mentions, it may also be because in the area where she practices the patients are generally technology savvy and the patients have themselves adopted the use of
mobile devices and apps. Second, the perception of being technology savvy is likely to depend on the HCPs’ technology use skills. As another mentioned:

*I think my only objection to looking for stuff in front of the patient is not knowing whether I am going to have a hard time finding something. If I feel like it is going to be a challenge to find something, I am not going to do it in the room. I don’t want them to see me struggling looking for the information.* (P10)

**Opportunity to Teach.** Another potential benefit of using the mobile app for teratogen information is that it provides HCPs an additional opportunity to engage the patients and to teach them. As suggested in prior work, engaging patients in the information search process can be valuable and is being practiced.

*The process can also enable HCPs to provide insights about credibility of information and help improve literacy: “I like to show patients where to get reputable information. I like to show them ‘where I’m looking up the information.’ I show them how to navigate the site and find relevant information.”* (P13)

There are pros and cons, however, with the use of mobile devices for cooperative information search. The benefit is that it is easy to just turn the phone around or give it to the patients so they can look at it (“Yeah I just pick it up and show it to them,” P4). However, compared to the computer, the small screen size makes it harder for both to look at the information together.

**Demonstration of Care.** Based on our survey results, it seems that some of the HCPs are concerned about using mobile apps in front of patients because they may be perceived as not being competent. However, some HCPs think that doing so may actually help demonstrate that they care. That they would take the additional step and time to look up something:

*When you stop and when you are looking up something, the patients are very interested that their doctor cares enough. I don’t think it shows a weakness that you don’t know the answer to the newest drug. They look at it positively saying “the doctor is looking up something specific to me. He cares about me.”* (P6)

*I’m usually transparent. I would like to be sure I’m giving the right dose, so I am going to look it up. I haven’t gotten complaints where they thought “their provider is an idiot he doesn’t know anything.” I think it’s saying I care. I want to do it right; I want to provide the right dose. No one’s ever complained about that. Safety first.* (P7)

**Discussion**

In this work, we present a design of a mobile teratogen-related information app for HCPs. We sought to address several key challenges with disseminating the information. We designed our app to consolidate different information types that are currently sought for by HCPs but spread out across different resources, as identified through our initial survey. Further, we examined the feasibility of displaying this information on mobile devices, where screen real estate is limited. We organized the in-depth information into expandable sections, and provided meta-level information about when the information on the app was last updated, in addition to editorial information. Through our study, we found that HCPs were able to complete a number of common tasks using our app and did so with ease. They also found the information presented to be valuable. However, we also uncovered challenges and opportunities in designing an app to convey health risk information, and the use of the mobile apps in the clinical setting.

**I. The Three Different Needs for Risk Information and Implications for Design**

With our design, we focused on supporting HCPs’ need to understand a drug’s potential for teratogenicity, particularly in situations where a pregnant patient has already been exposed to a medication and wants to know the likelihood that her baby will be born with a birth defect. We consolidated as much information as possible into the app, and chose a systematic approach for assessing risk that is fine-grained; one that offers 8 gradations of risk in addition to providing information about the data quality. The risk assessment is followed by a succinct narrative that describes and interprets the scientific evidence. Our app did well to support this need to understand.

However, what we found was that HCPs primarily need the app to help them make decisions about prescribing a medication during pregnancy. Given their constrained time, they often just want enough information to decide: safe, unsafe, or maybe safe. In other words, how they want to use the app and the risk ratings, is exactly what the FDA hoped to move away from when they discontinued the Pregnancy Categories. Thus, one of the challenges in using mobile technology to assess potential teratogenicity of a medication is how to compress a large body of complex (and often conflicting) data for quick decision aid without sacrificing critical aspects of the risk-benefit discussion. In fact, we believe this is a critical design challenge that generalizes to risk communication in the health domain.
One possible solution is to prepare interpretative statements for those agents that have substantial and/or conflicting evidence. This executive summary would include the risk ratings, dominant research findings, the context (dose, stage of pregnancy, etc.) for these findings (if known), recommendations for use during pregnancy, and preferred alternative drugs, if any. Topic-specific links would enable HCPs to quickly select the content most closely-related to their patient’s situation. HCPs also discussed factors other than risk that they would like to include in the app to support decision making, such as the costs of drugs and whether the drug is available over the counter.

Finally, we cannot overlook the importance of designing the app to help HCPs better explain risk to patients. In our design, we have included handouts that HCPs can print out to patients and they all appreciated that. Additional features such as being able to directly email portions of the information on the app to patients, or printing it out are also desired. But we also found that designing the information presentation to help HCPs translate the information to patients is valuable and needs to be considered in apps like these.

2. Successful Point-of-Care Integration of Teratogen Information into Clinical Workflows in Healthcare

Another goal of this work is explore the use of mobile apps for teratogen information in the clinical context. Our findings suggest several ways that the mobile app can be utilized and benefit patient-doctor interaction. Its use can provide HCPs access to information when they are away from other technology access. But even when they are in proximity of other technologies, using the app for information search might be quicker and minimize interference with other tasks. Aside from access, the app can also be used to demonstrate to patients that they care and that they are aware and proficient with new technologies. Further, its use in front of patients also provides HCPs an opportunity to engage and teach patients. As information becomes more accessible to patients, using the apps in front of patients enables HCPs to teach effective information seeking practices.

However, as we have found in our survey, a large number of HCPs have been slow to adapt the use of mobile apps in their practices and expressed concerns about appearing to be distracted or unknowledgeable. We believe part of the reason for the difference in perceptions about mobile app usage in the clinical practice is that HCPs work with different patient populations. Some work with patients where the mobile apps are highly adopted. Patients may expect their providers to be technology proficient and adopt a multitude of mobile-based tools. In addition, patients may also be accustomed to the use of mobile apps for general information seeking in their daily experiences and are less likely to perceive it negatively. Another reason for underutilization of mobile apps by HCPs may be the lack of wireless capability in some clinical settings.

But like general Health IT usage, we believe there are general strategies that can be employed to make the integration of mobile information seeking by providers in a clinical setting to be more successful. If HCPs can clearly communicate what they are doing and why they are doing it, there is no reason why the app usage needs to be perceived as negative (or any more negative than if the HCPs were to use any other resources). Further, effective engagement of patients in the process will decrease misperceptions from patients that apps are being used for personal reasons in addition to providing additional opportunities for HCPs to educate the patients.

Conclusion

In this work, we present an exploration of a mobile app prototype to support teratogen information seeking. Aside from the prototype itself, this work also contributes to the ongoing question of how to effectively present pregnancy drug risk information. Our results suggest that part of the challenge stems from the complexity and uniqueness of teratogen risk information and the different information needs of the providers – need to understand, to decide, and to explain. We suggest different features to support these three types of information needs. In addition, our findings also advance our understanding of the potential role of mobile health information apps in HCPs’ practices. Access to this type of information through mobile devices enables flexible uses (e.g., on the go), is quick and limits dependency on desktops that usually already have EMR and other applications running. Further, using apps in front of the patients may lead to potential benefits such as signaling physician care and tech-savviness, in addition to having additional opportunities to engage and educate patients. These insights can help facilitate designing for mobile apps for HCPs during provider-patient interactions.

Acknowledgement

This study was supported through SBIR Grant 1R43DD001127-01 awarded by the Centers for Disease Control & Prevention in an effort to improve medication safety for pregnant women. The authors are grateful to the HCPs who generously gave their time to participate in our study. We thank Margaret Adam, Amy Bauer, Angela Chien, Debra Doyle, Ira Kantrowitz-Gordon, Theresa O’Young, Mary Puttmann-Kostecka, Bat-Sheva Stein and WWAMI for...
their help in recruiting HCPs for the study. We would also like to thank Jina Huh, Yunan Chen, Richard Boyce, and Wanda Pratt for their feedback on this work.

References

10. UpToDate: UpToDate®, Waltham, MA. World Wide Web site: www.uptodate.com/
X Marks the Spot: Mapping Similarity Between Clinical Trial Cohorts and US Counties

Matthew C. Lenert, BA¹, Dara E. Mize, MD², Colin G. Walsh, MD MA (Primary Advisor)²

¹Vanderbilt University, Nashville, TN; ²Vanderbilt University Medical Center, Nashville, TN

Abstract

When patients and doctors collaborate to make healthcare decisions, they rely on clinical trial results to guide discussions. Trials are designed to recruit diverse participants. The question remains – how well do trial results apply to me or to people who live in our area? This study compared one complete clinical trial dataset (SPRINT) and one published study (ACCORD) to the Community Health Status Indicators dataset to assess the similarity of the trial populations to US county populations. Counties up to 495 miles to the closest SPRINT trial site and up to 712 miles to the closest ACCORD trial site had populations that were significantly more similar to the study cohort than counties farther away. The investigators detail a generalizable method for both assessing recruitment gaps in large multicenter trials and creating maps for clinicians to provide intuition on trial applicability in their area.

Background

The growing movement towards patient-centered care is an effort to improve the quality of health care delivery, thereby leading to a greater focus on shared decision making and evidence-based medicine¹. Under the shared decision-making model of patient care, it is the role of the clinician to present the best applicable clinical evidence. The role of the patient is to share their preferences and values. The physician then seeks to apply their clinical judgment to determine which therapies, diagnostic tests, and/or preventative services best fit their patient²,³. Randomized clinical trial (RCT) data often serve as the foundation of clinical evidence on which these decisions are made⁴,⁵.

The amount of data available for clinicians to digest continues to increase and integrating the latest study at the bedside or in clinic poses a challenge for practicing providers⁶,⁷. It is difficult to know – in practice and in real-time – how well a trial’s results apply to one’s patients. This task is further complicated with pervasive difficulties in determining the validity of study findings⁸. For example, the ratio of screened patients excluded from randomization has direct implications for the significance of trial results⁹. Studies may report surrogate outcomes that do not correlate with patient-valued outcomes (e.g. reporting changes in bone density for patients with postmenopausal osteoporosis is not the same as measuring rates of long-bone fractures)¹⁰. Treatment effects can be nonlinear across disease states, making the average effect deceptive¹¹. Providers often look to systemic reviews and meta-analyses as tools to consider these nuances.

Meta-analyses aggregate evidence by summarizing results to estimate an overall trend across many sample populations, but they have some stark limitations¹². It is difficult to produce a high quality meta-analysis, because one has to account for studies of varying external and internal validity in their designs. This heterogeneous mixture makes it difficult to focus on the scientific scale, and many analyses instead focus on the clinically removed statistical scale¹³. These limitations are the same limitations faced by providers, so it is not surprising that they are strong confounders.

Instead of pooling data from many small trials, there is a renewed push for building infrastructure to support large pragmatic-randomized-multicenter trials comparing an experimental treatment to the standard of care¹⁴. These trials recruit patients across the country and are carefully crafted to control for confounders. Despite accounting for population differences through the use of many sites, selection bias persists¹⁵. This bias affects the generalizability of trial results to the population at large, because the number of patients that participate in clinical trials may be a fraction of those seeking routine care¹⁶,¹⁷. Even in trials of tens of thousands of patients, trial samples may demonstrate differences from the population at large. Differences can manifest themselves in comparative functional status, age, and the number of comorbidities¹⁸. Trial results should hold for patients who are similar to trial participants, but there may be complex screening criteria or differences in subgroups that complicate comparisons¹⁹. Even at the highest level, measuring applicability of a study to the population at large remains a grand challenge.

A method to quantify similarity would help address the applicability gap. Researchers have used multiple definitions for patient similarity. The IBM TJ Watson group developed systems that combine supervised and
unsupervised approaches\textsuperscript{20,21}. Their system applies decentralized physician feedback to improve upon the base measure of similarity—in this case the Euclidean distance of a vector of clinical features for a pair of patients. The feature vector was made up of clinical variables such as International Classification of Disease (ICD) codes and objective clinical readings (e.g. blood pressure). Lee et al. took a similar approach, where they selected specific clinical features and put them into vector representation\textsuperscript{22}. This step enabled them to apply the cosine distance metric to quantify pairwise similarity, Zhang defined similarity as the Jaccard similarity over the set of ICD codes in the problem list between two patients\textsuperscript{23}. Panahiazar et al. used k-means clustering to quantify similarity as the Mahalanobis distance between a patient’s feature vector and the centroids of clustered training patients\textsuperscript{24}. There has been some exploration of making trial characteristic comparisons directly. In a small study, Cahan and Cimino devised a model that visualized where a patient would fall within the distribution of clinical and demographic features of an RCT cohort\textsuperscript{25}.

Most of these efforts focused on comparing a single patient to another patient or cluster of patients. A pairwise comparison is not ideally fitted for the problem at hand, since clinical trials tend to summarize characteristics of the cohort as well as the treatment effect. A clinical trial cohort has a distribution of clinical characteristics, while the individual patient has discrete characteristics. For example, the prevalence of smoking in the trial cohort might be 15%, but an individual currently smokes or does not. This suggests that it may be more tractable to compare the characteristics between two groups. Abstracting to the group level may lead to insights into how similar trial patients are to the local population. The investigators hypothesized that clinical trials record sufficient demographic, comorbidity, and adverse event data to readily produce a means of comparison with U.S. county-level health statistics. A quantifiable measure could then be mapped to give physicians an expedient intuition into how an RCT’s population compares to the people they see day to day. This measure may not sway a provider’s decision for a specific patient, but it may inform the (prior) probability they would recommend a specific therapy in the first place.

\textbf{Methods}

\textbf{Data Sources}

This study analyzed the complete dataset from the SPRINT Group’s blood pressure trial as well as the published data from the ACCORD Group’s type 2 diabetes trial\textsuperscript{26,27}. This study received IRB approval and completed a data use agreement to gain access to the SPRINT trial data. The SPRINT trial was a large pragmatic RCT that involved 9,361 patients across 96 clinical sites in the United States and 6 sites in Puerto Rico. This trial compared intensive blood pressure control to standard of care for patients with hypertension. The trial controlled for comorbidities such as chronic kidney disease, and several medication classes such as statins. The trial followed a clinically relevant primary outcome of myocardial infarction (MI), non-MI acute coronary syndrome, stroke, heart failure, or death attributable to cardiovascular disease. The SPRINT Group followed patients for a median of 3.26 years. The SPRINT trial planned demographic subgroups of: African-Americans, non-African Americans, ages less than 75, ages greater than or equal to 75, chronic kidney disease (CKD) present at the start of the trial, and CKD failure, or death attributable to cardiovascular disease. The SPRINT Group followed patients for a median of 3.26 years. The SPRINT trial planned demographic subgroups of: African-Americans, non-African Americans, ages less than 75, ages greater than or equal to 75, chronic kidney disease (CKD) present at the start of the trial, and CKD absent. The SPRINT trial is largely regarded as a model multi-site RCT for both its rigorous design and its methodical implementation.

The ACCORD group’s trial preceded the SPRINT trial by eight years and studied the effect of intensive control of glycated hemoglobin levels in patients with type 2 diabetes. The trial was a multi-site pragmatic trial that randomized 10,251 patients across 62 clinical sites in the United States and 15 sites in Canada. The ACCORD trial limited comorbidities and controlled for relevant demographic factors for cardiovascular and diabetic complications. The researchers also controlled for several medication classes such as beta-blockers and glucophages. The ACCORD trial measured the primary outcome of nonfatal myocardial infarction, nonfatal stroke, or death from cardiovascular causes. This trial planned for the following demographic subgroup analyses: age less than 65, age greater than or equal to 65, Caucasian, non-Caucasian, males, and females. The ACCORD group had an average follow-up period of 3.5 years. The ACCORD trial is another model exemplar of a pragmatic multi-site RCT.

The county health statistics came from the Community Health Status Indicators (CHSI) Dataset published by the Centers for Disease Control and Prevention (CDC). The CHSI dataset contains county-level social, environmental, and prevalence health data\textsuperscript{28}. This dataset is composed of a variety of CDC and Census Bureau studies from 2002 through 2014. It covers a wide array of disorders, diseases, and social behaviors. Its wide coverage of demographic and disease information at the county-level made it an ideal choice for this application. The investigators looked to map demographic and outcome data from the SPRINT and ACCORD trials to demographic and prevalence statistics from the counties based on the hypothesis that patients from a similar population should display similar features.

\textit{Mapping Between Data Sources}
Mapping variables from one data set to another requires input from clinical experts to ensure that nuances from the trial variables are accounted for and that the comparisons make clinical sense. The investigators established that a CHSI variable was comparable to a trial variable if and only if both physician authors agreed the comparison was reasonable. Demographic features such as obesity, smoking prevalence, and African-American race, mapped easily between the CHSI data and the SPRINT trial. The investigators assumed correlation between incidence rates of adverse events in the SPRINT control group and age-adjusted death rates in the CHSI county data. Outcomes of control patients were averaged to compare the prevalence of cardiac event and death, as well as stroke and death from the SPRINT control group to Coronary Heart Disease (CHD) death rates and stroke death rates in the CHSI data.

For the ACCORD trial the authors mapped the demographic features of African-American race, Hispanic race, and smoking status. The authors again used CHD death rates from CHSI and cardiac related deaths from the published data. The authors did not have access to the full ACCORD trial data; which prevented the inclusion of more features, such as the prevalence of obesity.

After selecting the variables for comparison the investigators needed to perform additional processing of the CHSI data set to ensure as fair a comparison was made as possible. Direct comparison between CHSI and trial variables is not advised, as trial participants are generally recruited conditional on the presence of some disease, disorder, or clinical state. For example, the SPRINT trial only recruited patients who had hypertension and the ACCORD trial only recruited patients who had type 2 diabetes. Patients with type 2 diabetes and patients with hypertension are subgroups of the general population, which may have markedly different characteristics.

**Missing Data**

The investigators did not find missing values in the CHSI data to be predictably missing (missing not at random). Therefore, the authors assumed that values in the data were missing at random. Predictably missing variables suggest bias in how the data were collected or how participants responded. Mean imputation was used to fill in values for variables with less than 3% (60 of 3,141 counties) of values missing. The authors used this threshold because the cost of computation time and complexity outweighed inferential benefits. The investigators used the mean of the state for mean imputation of missing county values.

Multiple imputation was used to fill in missing county values for variables with more than 3% of the feature missing. In multiple imputation, one uses a model to predict what the missing value would have been based on data one does have for that case. The investigators built one model per variable that had missing values. A bootstrapped elastic net (mixture of L1 and L2 penalty) penalized regression was used to select the features for each imputation model. Each model used all fully recorded demographic variables, features that were completed with mean imputation, quadratic, cubic, and log transforms of all features, and all interaction terms of first order features. The hyper-parameters (penalization for small regression coefficients and the ratio of L1 regularization to L2 regularization) of the elastic net regression were re-tuned for each bootstrapped sample over 200 bootstraps. The investigators optimized the hyper-parameters by minimizing the log loss of the model on the “out of bag” data. Once the most significant features were selected through the bootstrapped elastic net, the investigators needed to determine the effect or weights of those features before being able to predict the missing values. This required fitting the models over another 200 bootstraps where hyper-parameters were again retuned on each bootstrap. The resulting effects (coefficients) are the average of the coefficients for the 200 bootstrapped models. The variance of the imputed coefficients is the calculated variance from those 200 bootstrapped models.

**Conditioning on Disorders**

All CHSI variables required probabilistic conditioning on hypertension for the SPRINT trial and type 2 diabetes for the ACCORD trial. This step is crucial because all SPRINT trial participants have hypertension and all ACCORD trial participants have type 2 diabetes. The CHSI dataset is made up of multiple county averages over time along with the variance of that measure. The investigators used these moments to define a normal distribution for the county average of the various mapped variables (e.g. stroke deaths). The investigators needed to assume that each mapped variable was normally and identically distributed within a county for the years 2003-2014. This assumption means statistics over different years for the same county are independent and identically distributed. The investigators estimated the correlation parameter using a national average correlation of each variable and the conditioned variable (e.g. the correlation between obesity and hypertension).
Measuring Similarity

Similar to previous work, the investigators used feature vector representation to enable comparisons between a county and the trial population. Similarity was defined as the cosine similarity between a county feature vector and the average of the trial controls. For the SPRINT study, the feature vector was made up of the average rate of CHD death, stroke death, obesity, active smoking status, and African-American race. The investigators found the cosine similarity of each county vector to the SPRINT control group average. The same procedure was applied for the ACCORD trial control group. The cosine similarity measures were normalized and scaled to fall between 0-100, where larger values represented greater similarities.

Geographic Information System (GIS) Methods

CHSI data are keyed by state and county Federal Information Processing Standard (FIPS) codes. For example, Davidson County (county FIPS 037) in Tennessee (state FIPS 47) can then be mapped to other public data sources such as census data. The FIPS codes can be linked to GeoJSON data from the U.S. Census Bureau. The U.S. Census Bureau mappings allowed the conversion of FIPS codes to county border coordinates on GIS maps of the U.S. The investigators used SciPy and Folium (an open sourced package) for this process. The package allowed the investigators to illustrate the similarity of SPRINT/ACCORD participants to the people with hypertension or type 2 diabetes living in each county.

Statistical Analysis

The investigators sought to quantify if counties near trial sites were in fact more similar than those farther away. The investigators used latitude and longitude distances from the centroid of each county to the nearest trial site. The definition of “near” a trial site was increased one mile at a time starting at a radius of five miles. Figure 1 provides an illustration of the process of defining “near” and “far” sites. At each increment, a Wilcoxon Rank Sum test was used to examine if there was a difference in medians between the “near” sites and the “far” sites. The investigators repeated this procedure until no statistically significant difference, factoring in a Bonferroni-Holm correction, existed between counties “near” a trial site and those “far” away. The thought behind this procedure was that trial sites would mostly recruit patients within their geographic sphere of influence. However, as one expands the definition of “near”, differences should wash out as the average of the “near” group of counties becomes more like that of the “far” group of counties.

![Figure 1. Example of Near/Far County Definition From Vanderbilt University Medical Center (VUMC) Trial Site](image)

The investigators also analyzed if similarity had any correlation with differences in the rate of primary outcomes for the intervention group, essentially whether similarity was clinically meaningful. To begin this analysis, the investigators grouped the SPRINT participants by trial site. Next, the investigators calculated the similarity of the site-specific control group to the entire trial control group average, using the same feature vector of CHD death, stroke death, obesity, active smoking status, and African-American race. The investigators defined clinical significance as the absolute value of the site-specific difference in the primary outcome rate for the intervention group with the trial average primary outcome rate for the intervention group. Basically, does site similarity to the trial average (as measured by the control group) explain divergence in effectiveness of the intervention in the intervention group? Robust regression was applied to the measure of clinical significance with the following explanatory variables: similarity, number of participants at the site, and an interaction term. The investigators used the number of participants and the interaction term to address issues with discreteness. For example, a site with 3 participants can only observe rates of 0%, 33%, 67%, and 100%; such quintiles are far from a continuous measure.
Robust regression makes a Huber-White correction to the model variance, which often widens the confidence intervals for the model coefficients.

Results

The characteristics of the CHSI dataset are summarized in Table 1. All of the variables except for African-American race and Hispanic race appear to be symmetrically distributed. Symmetry is important for the rapid convergence of the sampling distribution (county measures over different years) to the normal distribution. The CHD variable was the only variable with missing values subjected to mean imputation (less than 3% missing). All other variables with missing values went through the multiple imputation pipeline. Table 1 and Table 2 are summarized to the national level for readability, but all the analysis was done at the county level.

Table 1. CHSI Data Characteristics

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>Median</th>
<th>Number Missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHSI Hypertension</td>
<td>32436 ± 7104</td>
<td>32200</td>
<td>551/3141 (18%)</td>
</tr>
<tr>
<td>CHSI % African-American</td>
<td>9.0%</td>
<td>2.1%</td>
<td>0/3141 (0%)</td>
</tr>
<tr>
<td>CHSI % Hispanic</td>
<td>7.0%</td>
<td>2.3%</td>
<td>0/3141 (0%)</td>
</tr>
<tr>
<td>CHSI Diabetes Type 2</td>
<td>10094 ± 3224</td>
<td>10000</td>
<td>682/3141 (22%)</td>
</tr>
<tr>
<td>CHSI Obese per 100,000</td>
<td>30779 ± 6146</td>
<td>30660</td>
<td>236/3141 (8%)</td>
</tr>
<tr>
<td>CHSI Smoker per 100,000</td>
<td>21281 ± 6314</td>
<td>20800</td>
<td>430/3141 (14%)</td>
</tr>
<tr>
<td>CHSI Stroke Death per 100,000</td>
<td>54±19</td>
<td>53</td>
<td>474/3141 (15%)</td>
</tr>
<tr>
<td>CHSI CHD Death per 100,000</td>
<td>249 ± 80</td>
<td>245</td>
<td>48/3141 (1.9%)</td>
</tr>
</tbody>
</table>

After imputation, the CHSI variables were scaled to probabilities and transformed into conditional probabilities, e.g. the probability of smoking given hypertension (SPRINT) or type 2 diabetes (ACCORD). Table 2 reports the conditioned national averages at the 95% confidence level. Again, the analyses of this study took place at the county level, not the national level, making Table 2 a summarized view of the data.

Table 2. CHSI Conditioned Variable National Averages

<table>
<thead>
<tr>
<th>Variable</th>
<th>95% Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHSI % African-Americans with Hypertension</td>
<td>[8.48%, 9.50%]</td>
</tr>
<tr>
<td>CHSI % Obese with Hypertension</td>
<td>[34.95%, 36.02%]</td>
</tr>
<tr>
<td>CHSI % Smoker with Hypertension</td>
<td>[22.94%, 23.67%]</td>
</tr>
<tr>
<td>CHSI % Stroke Death with Hypertension</td>
<td>[0.10%, 0.11%]</td>
</tr>
<tr>
<td>CHSI % CHD Death with Hypertension</td>
<td>[0.48%, 0.53%]</td>
</tr>
<tr>
<td>CHSI % African-American with Type 2 Diabetes</td>
<td>[8.48%, 9.50%]</td>
</tr>
<tr>
<td>CHSI % Hispanic with Type 2 Diabetes</td>
<td>[6.58%, 7.45%]</td>
</tr>
<tr>
<td>CHSI % Smoker with Type 2 Diabetes</td>
<td>[23.89%, 25.13%]</td>
</tr>
<tr>
<td>CHSI % CHD Death with Type 2 Diabetes</td>
<td>[0.46%, 0.52%]</td>
</tr>
</tbody>
</table>

The investigators compared each county data point to the average mapped characteristics of the trial control groups. The investigators used the trial control groups, because some of the features used were part of the primary outcome measured (CHD death and stroke death). These variables would likely be confounded by the intervention. Table 3 visualizes what the trial populations looked like for both the SPRINT trial and the ACCORD trial.

Table 3. Trial Control Group Data Characteristics

<table>
<thead>
<tr>
<th>Trial</th>
<th>African American</th>
<th>Current Smoker</th>
<th>Cardiac Event Death</th>
<th>Hispanic</th>
<th>Obesity</th>
<th>Stroke and Death</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPRINT</td>
<td>1493/4673 (31%)</td>
<td>590/4673 (13%)</td>
<td>67/4673 (1.4%)</td>
<td>NA</td>
<td>1952/4673 (42%)</td>
<td>17/4673 (0.3%)</td>
</tr>
<tr>
<td>ACCORD</td>
<td>968/5123 (19%)</td>
<td>702/5123 (14%)</td>
<td>124/5213 (2.4%)</td>
<td>379/5123 (7.4%)</td>
<td>NA</td>
<td>NA</td>
</tr>
</tbody>
</table>
Figure 2 visualizes the similarity of all US counties to the SPRINT trial. The pins represent SPRINT Trial sites and were placed by latitude and longitude independent of the shading. The darker shading in the figure corresponds to greater similarity between that county and the SPRINT control group. One can see that the South Eastern United States has the greatest concentration of both trial sites and darkly shaded counties.

Figure 2. Applicability Map of SPRINT Trial to US Counties

The distribution of similarity scores is not symmetric across the similarity range. Similarity tends to cluster in two places, around a similarity of 60 and around a similarity of 95. The histogram does not describe an obvious distribution. Similarity also appears to have more significance in relative terms than in absolute terms.

Figure 3. Distribution of SPRINT Similarity Across U.S. Counties
Figure 4 shows the applicability scores of the ACCORD trial to each US county. One can see again that the South East and Eastern seaboard appear fairly well represented by the trial population. The western and central counties appear tend to be less similar.

The distribution of similarity for the ACCORD trial, shown in Figure 5, also appears to be bimodal and asymmetric. In this histogram, similarity is concentrated around 75 and around 98. Similarity also appears to be more concentrated and less differentiated than in the SPRINT trial.
The investigators tested if the similarity of counties near trial sites was greater than counties farther away from any one trial site. For the SPRINT trial the investigators observed a statistical difference in similarity between near counties and far counties up until the inflection point of 495 miles from the nearest trial site. Past this distance, similarities no longer differed between the two groups. In the ACCORD data, differences in control group similarity between counties near trial sites and counties far from trial sites were significant up until 712 miles. Note that these distances are measured from a trial site to the geometric center of a county.

When assessing if similarity had any correlation with differences in intervention outcomes the investigators wanted to get a sense about the distribution of differences across sites. Figure 6 presents a box plot of differences in the rate of primary outcomes (in the intervention group) by SPRINT trial site. Note that in the SPRINT trial, the primary outcome was a negative (MI, non-MI acute coronary syndrome, stroke, heart failure, or death attributable to cardiovascular disease). The distribution is centered at zero, but deviances are skewed toward higher rates of the primary outcome than the average.

![Figure 6. SPRINT Intervention Group Difference in Primary Outcome Rates Grouped by Site](image)

The next part of the analysis took the absolute value of all the site intervention group differences and regressed them on site control group similarity, total number of patients recruited at the site, and an interaction term. Table 4 shows the confidence intervals of all the regressors for this model. Both similarity and the number of site participants were statistically significant predictors. The interaction term was not significant, but was left in the model because it had a significant contribution to the model $R^2$. This simplistic robust regression was able to explain 36% of the variance of site-specific intervention group differences in the rate of primary outcomes. The similarity coefficient is negative meaning that greater similarity leads to smaller differences in intervention group outcomes. More site participants also corresponded to smaller intervention group differences in outcomes.

Table 4. Robust Regression Coefficients of SPRINT Trial Sites

<table>
<thead>
<tr>
<th>Regressor</th>
<th>95% Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control Group Similarity</td>
<td>[-0.55, -0.03]*</td>
</tr>
<tr>
<td>Number of Participants at Site</td>
<td>[-1.27, -0.01]*</td>
</tr>
<tr>
<td>Interaction Term</td>
<td>[0.00, 0.01]</td>
</tr>
<tr>
<td>Intercept</td>
<td>[8.00, 38.27] *</td>
</tr>
</tbody>
</table>
Discussion

The primary findings of this work demonstrate that the geographic similarity of trial cohorts to population data can be estimated using publicly available health indicators. Proximity to study sites correlated with increased similarity from trial controls to county data. This result is intuitive and face valid, yet had not been demonstrated using this kind of approach, to the investigators’ knowledge. In the SPRINT trial most of the country, geographically speaking, is lightly shaded, meaning the trial cohort does not represent them as well. A possible rationale for this result is that the SPRINT trial population generally has more African Americans, fewer smokers, and more people with high Body Mass Indices (BMI) than what the authors found the national hypertensive population to be. Both trials performed various subgroup analyses on different demographic and clinical criteria. While neither publication mentioned oversampling explicitly, both research groups had predefined subgroups in mind during study design. The predefined subgroup analyses may have influenced the demographics of recruiting with regard to racial features, nevertheless the overall results the researchers reported reflect the mixture of who they recruited.

This study demonstrated weak correlation between cohort similarity and outcome similarity. This result offers further face validity to the method, because one might expect results to hold in a population that looks similar to the population used for testing. The regression results also demonstrate the dangers of over generalizing results; intervention outcomes can differ significantly from the trial reported average. Further study is needed to better establish the link between population similarity metrics and differences in intervention outcomes.

The investigators conjecture that comparing variables that are more contextually similar will increase the power of this method. For example, many studies report the mean BMI and its standard deviation; this cannot be mapped to the CHSI data set easily. Instead, it would be of greater value for this method to know the percentage of trial participants considered obese. The investigators were able to use the prevalence of obesity in the case of SPRINT trial because they had received the data as part of a New England Journal of Medicine (NEJM) Challenge. Generally, the full trial data set is kept privately as the hard-earned spoils of the trial authors’ large and taxing effort to design, fund, and execute the trial. The decision to share SPRINT trial data in rapid fashion by the NEJM was not without controversy, but reflects a cultural trend toward open data.

The strengths of this work include the application of publicly available data to generate a generalizable method for assessing similarity of trial controls to county population health data. The CHSI data cover many of the most impactful and prevalent diseases and demographic behaviors in the population. The investigators’ method is scalable to other trials, because very little trial specific data are needed. The data needed by this method can be aggregated, lessening the risk of health privacy issues. Mapping fields between CHSI and a trial can largely be done with control population summary statistics and tables of adverse outcomes. These two sources of information are often published as part of the trial results. Open source software was used to generate all results to encourage generalizability. While this method does require some clinical expertise for variable mapping and some work is required for each trial, the results can be served across the country, hinting that this approach has economy of scale returns. Another strength of the approach is its use of the normal distribution to condition general health statistics onto sub populations. The properties of the normal distribution allow one to condition on an infinite number of conditions using similar methods.

A major limitation of this approach is the lack of joint probabilities in CHSI data with respect to outcomes such as stroke death and risk factors such as hypertension. Having joint and marginal probabilities would allow for flexible conditioning and would not require the assumptions made here. This approach may benefit from better assumptions on distribution (e.g. Gamma or Poisson). These bivariate distributions are challenging to find conditional expectations for analytically. Altering the distributional assumptions represents future work. This method may also benefit from using a likelihood-based approach to compare the likelihood function of one population to another. Another assumption was the use of equal weights for all population features. It is likely that some population characteristics affect study applicability more than others. Future work should include replicative efforts with other trial data and assessment of whether the applicability differences actually manifest.

Conclusions

Large public data sets with a broad coverage of demographics, lifestyle variables, and clinical states empower researchers to think at a national level. The rapid assessment of an interactive map comparing trial cohorts to county populations, such as the one presented, may help inform shared decision-making between providers and patients. Specifically, this work would help providers set a prior probability of study applicability to their local population. This may also help providers quickly prioritize which trial results to integrate into their practice. Furthermore, trial administrators could use this method to assess geographic recruitment gaps in rapid fashion.
References
Environmental Reservoirs of Nosocomial Infection: Imputation Methods for Linking Clinical and Environmental Microbiological Data to Understand Infection Transmission

Julia Lensing\textsuperscript{1}, MS, K. Vilankar\textsuperscript{2}, MS, Hyojung Kang\textsuperscript{2}, PhD, Donald E. Brown\textsuperscript{2}, PhD, Amy Mathers\textsuperscript{3}, MD, Laura E. Barnes\textsuperscript{2}, PhD

\textsuperscript{1} Department of Systems Engineering, United States Military Academy, West Point, NY, USA
\textsuperscript{2} Department of Systems and Information Engineering, University of Virginia, Charlottesville, VA, USA
\textsuperscript{3} Division of Infectious Diseases and International Health, Department of Medicine, University of Virginia, Charlottesville, VA, USA

Abstract

The transmission of hospital-acquired Carbapenem-resistant Enterobacteriaceae (CRE) is a serious and growing concern in hospitals worldwide. Previous research of CRE found that traditional patient-to-patient transmission of the bacteria does not fully account for all cases of transmission. Recent efforts to further understand modes of transmission found identical genomes of CRE in patient sinks as was found in cultures collected from patients, indicating that environmental reservoirs could be playing a larger role in transmission than was first realized. This study evaluated imputation methods for linking multiscale clinical and environmental microbiological data. We then utilized the imputed data set to model the risk of CRE presence in sinks between culture dates. We demonstrated that imputation based on expert knowledge of the unique factors of the physical hospital layout and patterns of occurrence throughout hospital sinks provided the best representation of sink positivity and also identified several significant risk factors for explaining environmental contamination. This work helps to more clearly define the mechanism and risk of transmission from a wastewater source to hospitalized patients in a world with increasingly antibiotic-resistant bacteria which can thrive in wastewater environments and cause infections in vulnerable patients.

Introduction

The spread of hospital-acquired antibiotic-resistant bacteria is a serious and growing problem. The Center for Disease Control and Prevention (CDC) estimates that more than two million people are sickened each year with antibiotic-resistant infections resulting in at least 23,000 deaths [1]. Of the many bacteria that are growing in drug resistance, there is particular concern with gram-negative pathogens as they are becoming increasingly resistant to nearly all drugs, including Carbapenems. Carbapenems are widely considered to be the strongest class of antibiotic in use today and are typically administered as a last resort in treating gram-negative infections [1]. Because of their resistance to the strongest antibiotics available, infections caused by Carbapenem-resistant Enterobacteriaceae (CRE), most commonly found in a patient’s gastrointestinal tract, are difficult to treat and can result in death [2].

Traditionally, understanding and tracking of hospital acquired infections has focused on direct contact transmission where infected and colonized patients act as reservoirs of transmission between uncolonized patients and hospital workers [3]. A large healthcare facility in the United States has been tracking a low-level spread of a nosocomial CRE pathogen between August 2007 and September 2015 with approximately 450 patient colonizations. Since the discovery of CRE transmission, active surveillance of all patients has been instituted and strict contact precautions have been in place for patients who revealed a positive culture for the bacteria. Additionally, large-scale cleaning, disinfecting, sterilization, replacement, and limited-use interventions have been conducted, but these methods have produced only temporary successes in reducing the spread of the bacteria. Recent efforts to further understand transmission of the waterborne bacteria led hospital epidemiologists to begin sampling wet surfaces including patient sinks, toilets, and hoppers for the CRE pathogen. As a result, it has become increasingly recognized that non-patient reservoirs within the hospital setting may play a larger role in the transmission of drug resistant pathogens than was first realized [4] [5].

Beginning in September of 2013, nearly 3500 samples of sink drains, sink p-traps, toilets, hoppers, and other wet surfaces have been collected periodically throughout the hospital looking for the presence of CRE. Sampling found identical genomes in samples of patients as was found in the sinks, indicating that sinks, or other environmental factors, could be acting as reservoirs for transmission [6]. However, manpower and financial constraints limited the frequency and volume of samples taken over time leaving a large gap in knowledge of the lifecycle of the bacteria in sinks and
resulting in a sparse data set with samples at irregularly spaced intervals. In order to understand the sink positivity between sample dates and to utilize the environmental microbiological sample results for further modeling, the data must be at equally spaced intervals, and thus missing data must be cleaned by removal or imputation.

Missing data are often unavoidable in epidemiological and clinical research. They can bias study results distorting statistical parameter estimates and decreasing the statistical power of a study. Numerous methodologies have been proposed to handle missing data [7]. We hypothesized that using standard methods of imputation would not provide the best representation of the environmental samples in this study because the study data include all three types of missing data, and the percentage of missing observations is significantly larger than the number of non-missing data when translated to daily measurements for modeling.

To determine which imputation method provides better representation of the presence of CRE in patient sinks over the study period, we examined two non-standard imputation methods as well as two regularly practiced methods. After determining the imputation method that resulted in the most accurate representation of the environmental sample results, we used the imputed data set for further modeling to understand the significant risk factors in predicting the presence of CRE in patient sinks.

The objectives of this study are twofold: 1) to evaluate imputation strategies for continuous representation of irregular sampled environmental microbiological data and 2) to use the selected imputation approach for modeling of the environmental reservoir positivity as a function of other significant risk factors including colonized patients, environmental interventions, and characteristics of patient room sinks.

1 Related Work

1.1 Patient Infection Risk Factors

The transmission of CRE infection occurs when a non-colonized person comes into direct contact with an infected or colonized patient, through intermediate carriers such as healthcare workers, or through contact with contaminated environmental reservoirs such as sinks and toilets, among others. Previous studies [8, 9] from the US and Israel have shown that the primary risk factors for patient acquisition include ICU stays, long-term hospitalization, transplantation and antibiotics. Another outbreak of CRE at the Tisch Hospital at New York Medical Center included 24 infected patients in intensive care units (ICUs) over the course of a year. Similar to other outbreaks, risk factors for infection during this outbreak included prolonged hospital stay, a stay in the ICU, and ventilator usage [10].

1.2 Environmental Risk Factors

Risk factors from the studies conducted in the US and Israel did not completely explain infection transmission. Previous studies [4, 11, 12] provide evidence that environmental reservoirs are a source of infection and transmission. Additionally, a study from Spain [4] described an outbreak due to multidrug-resistant *Klebsiella oxytoca* in an ICU where damp environmental reservoirs were linked to bacterial transmission. In that study samples collected from sinks, drainpipes, and traps showed that only one storage sink, which had its drainpipes connected to two other sinks, was found to be positive. The connecting drainpipes were also found to be positive. Furthermore, this study showed that the outbreak was completely eradicated after replacing the horizontal drainage system that connected the two impacted sinks. In conclusion, this study stated that wet environmental reservoirs should be considered when strictly applied traditional control measures are not efficacious.

A study from France [12] found that sinks were frequently contaminated in ICUs as a result of their use in disposing of patient bodily fluids and were a potential source of extended-spectrum beta lactamase-producing *Enterobacteriaceae* (ESBLE), thus increasing risk in the environment of patients as a consequence of the splash-back effect. Recent research from Wolf et al. [11] demonstrated that sinks acted as a source of infection by verifying that the ESBLEs recovered from patients were identical to those that had been previously recovered from sinks. The outbreak described in a Colombian [13] study found that the likely cause of the infections was the improper design of sinks; in particular, the joints of the sinks to the walls were not sealed, leading to facilitation of colonization.
1.3 Modeling Infection Transmission

Methods including logistic regression have been used in previous studies to identify risk factors associated with nosocomial transmission. A study [14] conducted at Roosevelt hospital in New York City used logistic regression models to evaluate efficacy of infection control measures in preventing the transmission of multidrug-resistant tuberculosis. The study found that distance from infected patient room is a significant predictor of nosocomial transmission.

The logistic regression approach has also been used in some studies [15, 16] to identify patient risk factors for CRE transmission in hospital settings. Research by Papadimitriou-Olivgeris et al. [15] focused on patient characteristics, diagnosis, and procedures to determine that prior ICU stay, duration of previous hospitalization, diagnosis of chronic obstructive pulmonary disease, carbapenem administration, and beta-lactamase administration were significant risk factors. Similarly, Tuon et al. [16] considered procedures such as mechanical ventilation and found that urinary catheter devices and central venous catheter devices were significant risk factors, along with advanced age and antibiotic exposure to ciprofloxin.

Work by [17] focused on modeling the nosocomial transmission of carbapenem-resistant bacteria using logistic regression and random forest models to determine important risk factors for CRE transmission. Both models showed that distance to the infected room was one of the significant predictors. One of the models showed that the proximity to sinks is important in predicting patient infection. This study also found that the cumulative presence of positive patients in the same room as a sink, distance from the bed to sink, and sink design are significant predictors of sink positivity. However, the model was constructed using limited data and a restricted time range, which leaves many questions about the role of sinks in the spread of infections.

Studies [4, 11–13] have demonstrated that sinks play a role in infection transmission but did not highlight additional environmental risk factors responsible for sink contamination. This study improves on the understanding of sink contamination by highlighting important variables. This study is similar to [17] in terms of the modeling approach used, but it differs in the level of spatiotemporal variables from environmental data. In addition to examining the presence of positive patients in the same room over time, this study also considers the status of neighboring rooms and sinks as potential risk factors.

1.4 Imputation in Healthcare Studies

Missing data is a frequent issue in epidemiological and health sciences research. While most modeling techniques would simply call for the removal of incomplete cases, doing so can cause bias and loss of information [18]. Imputation methodologies can be used to replace missing data with substituted values in order to create a complete data set. Standard methods for imputation such as k-nearest neighbor, last observation carried forward, and mean value can lead to bias and are not effective in handling categorical variables [19]. These methods are also typically used to address only one of the three types of missing data - missing completely at random, missing at random, and missing not a random.

Missing data in health care settings is a result of many complex sources of information and their incompatibility between each other and also due to the personal nature of care given that varies among patients [20]. A recent study of outcomes of trauma cases found that using a combination of existing imputation methods, specific to each individual data set in the analysis significantly improved the accuracy of predicted mortality after trauma over the more common method of handing missing data by simply removing all unknown cases [20]. Additionally, a study on classification of respiratory patterns involving imputation of missing data found that self-organizing maps (SOM) machine learning techniques were more effective in predicting patterns than other more commonly used statistical methods such as mean/mode imputation and multiple linear regression [21].
Table 1: Description of Variables for Imputation of Sink Positivity

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
<th>Input Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unit</td>
<td>unit of hospital</td>
<td>ICU1, ICU2, ICU3</td>
</tr>
<tr>
<td>PTBR</td>
<td>patient (PT) or bathroom sink (BR)</td>
<td>PT, BR</td>
</tr>
<tr>
<td>Lag0</td>
<td>presence of positive patient for sink/day at time t=0</td>
<td>“0”, “1”</td>
</tr>
<tr>
<td>Lag1</td>
<td>presence of positive patient at time t-1</td>
<td>“0”, “1”</td>
</tr>
<tr>
<td>Lag2</td>
<td>presence of positive patient at time t-2</td>
<td>“0”, “1”</td>
</tr>
<tr>
<td>Lag3</td>
<td>presence of positive patient at time t-3</td>
<td>“0”, “1”</td>
</tr>
<tr>
<td>Intervention0</td>
<td>intervention in place at time t=0</td>
<td>“0”, “1”</td>
</tr>
<tr>
<td>Intervention1</td>
<td>intervention in place at time t-1</td>
<td>“0”, “1”</td>
</tr>
<tr>
<td>Intervention2</td>
<td>intervention in place at time t-2</td>
<td>“0”, “1”</td>
</tr>
<tr>
<td>Intervention3</td>
<td>intervention in place at time t-3</td>
<td>“0”, “1”</td>
</tr>
<tr>
<td>CumSum7</td>
<td>cumulative number of positive patients in 7 days prior</td>
<td>[0, 7]</td>
</tr>
<tr>
<td>CumSum14</td>
<td>cumulative number of positive patients in 14 days prior</td>
<td>[0, 14]</td>
</tr>
<tr>
<td>SinkDesign</td>
<td>sink design based on unique values from sink variable description</td>
<td>25 designs “A-Y”</td>
</tr>
<tr>
<td>SinkBedDis</td>
<td>measurement of distance from sink to bed</td>
<td>[4, 14] feet</td>
</tr>
</tbody>
</table>

2 Methods
2.1 Data

This Institutional Review Board (IRB) approved study used clinical and environmental data from a major U.S. hospital. Clinical data included culture dates of approximately 130 patients who tested positive for CRE and their room movement information from September 2013 to September 2015. Environmental data included 967 positive and negative swab and liquid samples from 166 sink drains and sink p-traps from floor 3-8 in the hospital over the same 740 day period. Twenty six attributes describing the physical characteristics of the sink were manually collected. The attributes include surrounding countertop area, faucet, exposed piping, and the presence of other wet areas such as showers, hoppers, and toilets. Lastly, dates and units of any clinical interventions that were performed during the period in question were collected.

For this analysis, our environmental microbiological sample dataset was transformed into to a sink-day format that includes a record for each sink and each day during the study period. We defined positive (“1”) population as the set of sink-days for which a known positive culture was found, negative (“0”) population as the set of sink-days for which a known negative culture was found, and unknown (“U”) population as the set of sink-days for which there was missing sample data or no environment sample taken. Of our data set with approximately 81,000 cases, only about 1,000 (1.2%) had known culture values.

Table 1 describes the variables used in this study. Variables Lag0, Lag1, Lag2, and Lag3 indicate whether on the given day, a positive patient stayed for any length of time in the given room. The lag number corresponds to the time with Lag0 indicating the current day, Lag1 indicating one day prior, and Lag2 indicating two days prior. If a positive patient moved through the room, a binary response of “1” was indicated at the appropriate lag value. Similarly, CumSum7 and CumSum14 were created to demonstrate the cumulative number of days in the past 7 or 14 days respectively that a positive patient was in the given room with the given sink. Variables Intervention0, Intervention1, Intervention2, and Intervention3 were used to indicate whether an intervention was performed on the current day or one of the prior three days. The timing of different interventions was considered since infection control leadership hypothesized that these interventions influenced environmental positivity. SinkDesign represents 25 unique sink characteristics throughout the 166 sampled sinks. SinkBedDis represents the distance of the patient bed to the sink.
2.2 Imputation of Environmental Sample Data

We evaluated four imputation methods, two non-standard methods based on expert guidance (Linear and Midpoint methods) and two traditional imputation methods that are frequently used in health care studies (Logistic Regression and Multivariate Imputation by Chained Equations). In the linear imputation method, a positive sample is carried forward to the next negative sample while linearly imputing the probability of the sink being positive between negative and positive observations. The midpoint method carries the first observation forward to the midpoint between two observations and carries the next observation backward to the midpoint. The logistic regression method trains a set of data on the given samples and predictor variables and predicts on the missing samples. The mutivariate imputation by chained equations (MICE) method uses the statistical package MICE in R to perform multiple imputation over 10 iterations of known variables to predict a binary response [22] [23].

For the purpose of imputation, we consider each sink as a univariate time series. Each imputation method was implemented on each sink over the period of 740 days resulting in a value for each sink-day that will be used in modeling. Predicted values are replaced by any known values. Figure 1 shows visual depiction of the sample data and each of the four methods on one sink over the first 350 days.

![Figure 1: Example plot of Imputation Methods for Patient Sink](image)

2.3 Modeling and Feature Selection

We employed two modeling techniques to identify risk factors that affect sink positivity. First, Bayesian model averaging (BMA) [24] was used to explore all possible subsets of variables from Table 2 and identify the most significant predictors. Based on each subset model’s posterior probability, which can be estimated using Bayesian information criterion (BIC), the models that were at least 20 times less likely than the best model were removed. Among the rest of the models, any model which was less probable than its nested subset model was also excluded. The remaining models were averaged to estimate variable importance as well as mean and standard deviation of the coefficients. Secondly, a random forest model using an ensemble technique was developed to examine variable importance and to predict sink positivity. The random forest model was trained with 10-fold cross validation, using 500 trees with a maximum node size of 5.
Table 2: Variable Selection for Sink Risk Modeling

<table>
<thead>
<tr>
<th>Variable</th>
<th>Values</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>P</td>
<td>+/-</td>
<td>Patient was positive on current day</td>
</tr>
<tr>
<td>P_{lag7}</td>
<td>+/-</td>
<td>Patient positive in last 7 days in same room</td>
</tr>
<tr>
<td>P_{lag14}</td>
<td>+/-</td>
<td>Patient positive in last 14 days in same room</td>
</tr>
<tr>
<td>P_{lag21}</td>
<td>+/-</td>
<td>Patient positive in last 21 days in same room</td>
</tr>
<tr>
<td>P_{lag28}</td>
<td>+/-</td>
<td>Patient positive in last 28 days in same room</td>
</tr>
<tr>
<td>status(P_1)</td>
<td>+/-</td>
<td>Patient in adjacent rooms that share common plumbing with next room was positive</td>
</tr>
<tr>
<td>status(P_2)</td>
<td>+/-</td>
<td>Patient in adjacent rooms that do not share common plumbing with next room was positive</td>
</tr>
<tr>
<td>status(S_1)</td>
<td>+/-</td>
<td>Patient in adjacent rooms that share common plumbing were positive</td>
</tr>
<tr>
<td>status(S_2)</td>
<td>+/-</td>
<td>Patient in adjacent rooms that do not share common plumbing were positive</td>
</tr>
<tr>
<td>I_3</td>
<td>0 or 1</td>
<td>Intervention occurred in the same room in last 3 days</td>
</tr>
<tr>
<td>I_5</td>
<td>0 or 1</td>
<td>Intervention occurred in the same room in last 5 days</td>
</tr>
<tr>
<td>I_7</td>
<td>0 or 1</td>
<td>Intervention occurred in the same room in last 7 days</td>
</tr>
<tr>
<td>S_{lag30}</td>
<td>0 or 1</td>
<td>Same room sink was positive in last 30 days</td>
</tr>
</tbody>
</table>

3 Results

3.1 Imputation Methods

We evaluated the four imputation methods using the data that consists of 967 known positive samples. We began by evaluating a simple generalized linear model (logistic regression) of the given samples as a response to the extracted and selected features in Table 1 where $g_t$ annotates the probability of a positive sink given the selected set of predictor variables.

\[
g_t = \beta_0 + \beta_1(\text{Unit}) + \beta_2(\text{CumSum14}) \\
+ \beta_3(\text{SinkDesign}) + \beta_4(\text{Intervention2})
\] (1)

Next, we sequentially add in an imputed value from each method at t-3 (3 days prior to sink-day), t-7 (7 days prior to sink-day), and t-15 (15 days prior to sink-day) to determine which imputation method provided the most increase in the prediction accuracy of the sink based on 10-fold cross validation.

\[
g_t = \beta_0 + \beta_1(\text{Unit}) + \beta_2(\text{CumSum14}) \\
+ \beta_3(\text{SinkDesign}) + \beta_4(\text{Intervention2}) \\
+ \beta_5(\text{Imputed Value})
\] (2)

Our baseline data set consisted of the original 967 sink-days while the imputed data sets provides us the ability to use over 80,000 completed sink-days. Table 3 shows the performance of the four imputation methods. It indicates that
Table 3: Imputation Method Accuracy Results

<table>
<thead>
<tr>
<th>Method</th>
<th>t-3</th>
<th>t-7</th>
<th>t-15</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear</td>
<td>82.4%</td>
<td>79.9%</td>
<td>78.1%</td>
</tr>
<tr>
<td>Midpoint</td>
<td>98.2%</td>
<td>86.1%</td>
<td>80.5%</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>69.7%</td>
<td>72.0%</td>
<td>70.4%</td>
</tr>
<tr>
<td>Mice</td>
<td>71.2%</td>
<td>71.0%</td>
<td>71.3%</td>
</tr>
<tr>
<td>Baseline (no imputation)</td>
<td>70.4%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

regardless of time interval, the Midpoint imputation method provides the highest accuracy of prediction of positive sink compared to the baseline (without any imputed values).

3.2 Logistic Regression Model

Based on the results of BMA, we selected the variables whose importance was higher than 10 percent for the final logistic regression model. Table 4 shows the odds ratios and confidence intervals of the variables included in the final model. The odds ratio denotes the increase in probability of a sink becoming positive given the presence of any variable. The confidence interval of odds ratio was calculated using a bootstrap method.

The results showed that the odds of a sink becoming positive increases by almost 13-fold (95% CI: 12.39-14.18) when the status of the sink was positive in the last 30 days. Intervention has a negative effect on sink positivity. This is consistent with the general understanding that interventions would reduce the odds of sink contamination for a short duration. The odds ratio reduces by 50% if an intervention is performed within the last 7 days. Instances of positive sinks in adjacent room\(_1\) which share common plumbing increase the chances of a sink becoming positive by 80% (95% CI:1.68-1.93). Instances of positive sinks in adjacent room\(_2\) that do not share common plumbing increase the chances of a sink becoming positive by only 13% (95% C.I: 1.03-1.24).

The presence of a positive patient in the same room increases the odds of a sink becoming positive by 1.92-fold (95% CI: 1.46-2.53), whereas the presence of any positive patient within the past 14 days increases the odds by 1.78-fold (95% CI: 1.52-2.08). Furthermore, the presence of a positive patient in an adjacent room\(_1\) which shares common plumbing increases the odds of sink positivity by 1.13-fold (95% CI: 0.92-1.40), while the presence of a patient in adjacent room that do not share common plumbing increases the odds by 1.70-fold (95% CI: 1.37-2.09). The overall significance of the model was evaluated using the likelihood ratio test, and the result indicates the model is significant at \(\alpha=0.05\) level.

Table 4: Odds Ratio with Confidence Intervals - Logistic Regression

<table>
<thead>
<tr>
<th>Variable</th>
<th>Odds Ratio</th>
<th>95% C.I</th>
</tr>
</thead>
<tbody>
<tr>
<td>(P) Same Room Patient Status</td>
<td>1.92</td>
<td>1.46 - 2.53</td>
</tr>
<tr>
<td>(P_{lag14}) Positive patient last 14 days</td>
<td>1.78</td>
<td>1.52 - 2.08</td>
</tr>
<tr>
<td>(P_1) Adjacent Room(_1) Patient status</td>
<td>1.13</td>
<td>0.92 - 1.40</td>
</tr>
<tr>
<td>(P_2) Adjacent Room(_2) Patient Status</td>
<td>1.70</td>
<td>1.37 - 2.09</td>
</tr>
<tr>
<td>(S_1) Adjacent Room(_1) Sink Status</td>
<td>1.80</td>
<td>1.68 - 1.93</td>
</tr>
<tr>
<td>(S_2) Adjacent Room(_2) Sink Status</td>
<td>1.13</td>
<td>1.03 - 1.24</td>
</tr>
<tr>
<td>(S_{lag30}) Sink status in last 30 days</td>
<td>13.25</td>
<td>12.39 - 14.18</td>
</tr>
<tr>
<td>(I_7) Interventions carried out</td>
<td>0.50</td>
<td>0.45 - 0.56</td>
</tr>
</tbody>
</table>

3.2.1 Random Forest Models

The results for Random Forest are shown in Figure 2, which summarizes the importance of variables by mean decrease accuracy (MDA). MDA can be used as a feature selection method because it shows the impact of each variable on the
model accuracy. The results show that 7 of the 9 variables included in the logistic regression were also included in the random forest model, while the accuracy slightly increased (81%) compared to Logistic regression (80%). Both models show identical true positive rate, which is important in our case since the cost associated with misclassification of a positive sink as negative is higher than falsely classifying negative sink as positive. Additionally, we see a higher specificity based on random forest model compared to logistic regression. Table 5 summarizes the results from both models.

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
<th>AUC</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic Regression</td>
<td>80%</td>
<td>78.8%</td>
<td>84.3%</td>
<td>73.2%</td>
</tr>
<tr>
<td>Random Forest</td>
<td>81%</td>
<td>80.4%</td>
<td>84.3%</td>
<td>76.6%</td>
</tr>
</tbody>
</table>

### 4 Conclusions

The modeling results indicate that variables such as presence of positive sink in adjacent rooms that share common plumbing, status of the patient in the same room, status of sink in the past 30 days, presence of a positive patient in last 14 days, presence of a positive patient in the adjacent room and interventions performed in the past 7 days are significant risk factors in explaining sink contamination. The findings on the presence of a positive patient in last 14 days, status of sink in the last 30 days and interventions implemented in the last 7 days are consistent with the understanding shared by infection control practitioners we consulted with.

Previous research [25] has shown that biofilms found in sinks were linked to outbreaks. Some research [26,27] has also shown that these biofilms are resistant to traditional disinfectant methods. Our results indicating the significance of adjacent rooms that share common plumbing can be explained by the probable presence of biofilms in sink drain walls, traps or drainpipes connecting the two room sinks. Despite the timely implementation of intervention strategies, it could be likely that the presence of biofilms contributed to the adjacent sinks becoming positive.

Another spatial factor that we found to be significant in the model is the presence of a positive patient in the adjacent room. The model results show that the odds of sink positivity increase by 1.13-1.70 fold when the adjacent rooms have a positive patient. We would expect this to be true given that positive patients in the adjacent room would use the sink (adjacent room sink). Thus, it is likely that they would contaminate the sink of the room that they are staying in.

Developing an imputation method for bacterial presence between known samples provided valuable insight into the growth and movement of the bacteria throughout sinks in the hospital. Additionally, the ability to complete a data set of environmental testing allows for the capability for it to be included in other models, such as patient risk modeling,
that will provide a more comprehensive predictive capability. In modeling the presence of CRE in sinks, this research aimed to understand what factors contribute to sink positivity. Knowledge gain in both capacities can lead to better understanding the role of environmental reservoirs in the spread of the bacteria and could be used to support changes to hospital policy and procedures that ultimately aid in the containment or eradication of the bacteria from the hospital.

Acknowledgment

This work was supported by the Coulter Translational Research Partnership and the University of Virginia Health Center.

References


Bootstrap-based Feature Selection to Balance Model Discrimination and Predictor Significance: A Study of Stroke Prediction in Atrial Fibrillation

Xiang Li, PhD¹, Zhaonan Sun, PhD², Xin Du, MD³, Haifeng Liu, PhD¹, Gang Hu¹, Guotong Xie¹

¹IBM Research - China, Beijing, China
²IBM T.J. Watson Research Center, New York, USA
³Department of Cardiology, Beijing Anzhen Hospital, Beijing, China

Abstract

Atrial fibrillation (AF) is a common cardiac arrhythmias, which increases the risk and severity of ischemic stroke. For predicting ischemic stroke in AF patients, a risk prediction model that can achieve both good model discrimination (e.g., AUC) and statistical significance of predictors is required in real clinical practices. In this paper, we propose a new bootstrap-based wrapper (Boots-wrapper) method of feature selection, and apply this method on Chinese Atrial Fibrillation Registry data to develop 1-year stroke prediction models in AF. The proposed method can heuristically search a subset of features to maximize the discrimination of the prediction model and minimize the penalty for the non-significant features. To achieve robust feature selection, we perform bootstrap sampling to get a more reliable estimate of the variation and significance statistics. The experimental results show that Boots-wrapper can balance model discrimination and statistical significance of features for developing AF stroke prediction models.

Introduction

Atrial fibrillation (AF) is a common cardiac arrhythmias, affecting over 2.7 million Americans¹ and approximately 10 million adults in China². AF significantly increases the risk and severity of ischemic stroke³. Though warfarin and radiofrequency ablation (RFA) are both effective therapies in preventing ischemic stroke for AF patients⁴, these therapies are normally recommended to high-risk AF patients⁵ because warfarin may have side effect of bleeding and RFA is expensive. Therefore, it is critical to accurately identify the truly high-risk AF patients that have high probability to develop ischemic stroke within relative short-term periods (e.g., 1 year). For this purpose, in our previous AF studies⁶⁷⁸, we identified potential predictors for ischemic stroke and built prediction models using different feature engineering and machine learning methods, based on Chinese Atrial Fibrillation Registry (CAFR) data⁹, which can achieve significantly better model discrimination in terms of the area under the ROC curve (AUC) than the widely-used risk models such as CHADS², CHA²DS²-VASC¹⁰ and Framingham Score¹¹.

However, there is a factor we did not take into account in our previous studies – the statistical significance of predictors, which is as important as discrimination for a risk model to be adopted in many clinical scenarios. Actually, because the purpose of stroke prediction is to help clinicians identify high-risk AF patients and then decide to whom warfarin and/or RFA should be used, only those who have not been treated with warfarin or RFA at baseline are of interest for developing stroke prediction model. This criteria greatly decreases the number of patients that can be used in modeling. Furthermore, compared to long-term risk prediction (e.g., 5-year stroke risk prediction by Framingham score¹¹), relative short-term prediction (e.g., 1-year) is more practical for clinical decision-making. However, a 1-year stroke prediction model had to be built on a highly imbalanced dataset, because there was not enough cases of ischemic stroke onset within 1 year after baseline. For this dataset with limited patient numbers and rare positive cases, the statistical significance test of predictors in the multivariate risk model is even more important. Therefore, the objective of this study is to identify potential predictors for developing 1-year prediction models for AF patients, by balancing model discrimination and statistical significance of predictors.

Feature selection methods in machine learning¹² can be used to automatically test and select predictors from a large number of candidate features. There are three main supervised feature selection strategies: filter, wrapper and embedded optimization. Filter methods calculate a score to represent the relevancy of a feature against the outcome and filters the features based on the score. Though the p-value from chi-squared or other tests can be used as the relevancy score to filter the features that have significant univariate correlation with the outcome, it can neither ensure
the statistical significance of features in a multivariate model, nor the discrimination of the model. Wrapper methods utilizes a specific learning model (e.g., logistic regression) to select the subset of features that provides the best performance for a specific metric using a heuristic search strategy. The current wrapper subset selection method can optimize the model discrimination (e.g., AUC) of a subset of features by cross-validation, but cannot ensure the statistical significance of selected features. The traditional stepwise selection method for generalized linear model (GLM, including logistic regression and Cox regression) is also an example of wrapper method, which uses the Score statistics to evaluate the goodness-of-fit and constrains the statistical significance of selected features in the model using a hard constraint (e.g., p-value of an entered feature must be less than 0.05). However, the hard constraint of p-value may result in under-fitting for a small imbalanced dataset. Finally, embedded optimization methods incorporate feature selection directly into the learning process of a model. For example, the Lasso method introduces L1-norm regularization to GLM, which can achieve feature selection by shrinking the coefficients of low relevant features to zero during model training. However, it is difficult to constrain statistical significance of features in the objective function of model optimization.

In this paper, we propose a new bootstrap-based wrapper (Boots-wrapper) method of feature selection, and use this method to build stroke prediction models for AF patients. The proposed method can select a subset of features to balance the discrimination (e.g., AUC) of the GLM model built on the selected features and the statistical significance of features, by maximizing the model discrimination and minimizing a soft penalty for the non-significant selected features. To achieve robust feature selection, we perform bootstrap sampling instead of cross-validation to get a more reliable estimate of the variation in feature coefficients, and also use the p-value from bootstrap test to get more reliable significance statistics. We applied Boots-wrapper on a training dataset derived from CAFR data to develop 1-year prediction models of ischemic stroke for AF patients, and compared its performance on a separate testing dataset with that of existing feature selection methods. The experimental results show that Boots-wrapper can balance model discrimination and statistical significance of features for developing AF stroke prediction models.

Methods

Cohort

The objective of this study is to build 1-year ischemic stroke prediction models for AF based on the Chinese Atrial Fibrillation Registry (CAFR) data'. The CAFR study started from the year of 2011 to 2015, and has enrolled more than 17,000 AF patients from 32 hospitals in Beijing, China. The study collected the patients’ demographics and clinical information at baseline, and followed up the patients every 6 months. At every follow-up visit, the clinical events such as ischemic stroke were collected. In this study the patients of interest are those who had not been treated with warfarin or RFA at baseline. From the CAFR data, we identified 3738 AF patients who meet this criteria, where 143 patients (3.83%) are cases who had ischemic stroke within 1 year after baseline, and 3593 patients are control instances who completed 1-year follow-ups and did not have ischemic stroke within 1 year. To make the result models more interpretable, the clinician selected 51 features as candidate features, including demographics, medical histories, vital signs, laboratory test results and treatments.

In this study, we used the data of patients registered in 2011 and 2012 as the training set (2872 patients, where the proportion of positive cases is 4.04%) to develop the risk prediction models, and the data of patients registered after 2013 as the testing set (866 patients, where the proportion of positive cases is 3.12%) to evaluate the performance of the models.

Method Framework

The goal of our method is to select a feature set that can maximize the discrimination of the GLM model (e.g., logistic regression) built on the selected feature set, and minimize the penalty for the non-significant selected features:

$$L = \text{discrimination (features)} - \text{penalty (non-significant features)},$$

where discrimination can be measured using the metrics such as accuracy, sensitivity, specificity, or AUC. However, it is very difficult to optimize this objective using an embedded feature selection method that is built in the process of model learning. Therefore, we propose a wrapper method of feature selection to heuristically optimize the objective. Figure 1 shows the pipeline of our proposed approach Boots-wrapper, which follows the general framework for wrapper methods of feature selection. Given the training dataset and the testing dataset, the approach will perform the following main steps:
• **Feature Set Search**: searching a subset of features from all possible feature subsets. Some existing heuristic search strategies, such as hill-climbing, best-first, and genetic algorithms\(^{11}\) can be used to search the feature set.

• **Bootstrap-based Feature Set Evaluation**: evaluating the performance of the GLM models that are trained on the above subset of features. In this step, our feature set evaluation method, which uses the bootstrap-based algorithm to evaluate model discrimination and statistical significance of features, is different from the previous wrapper method and can achieve more reliable estimation. We will give more details about the evaluation method in the following sub-section.

• **Feature Set Search** and **Feature Set Evaluation** are repeated until the performance is optimized and the final predictor set is found.

• **Final Model Training**: developing the final GLM model using the selected set of predictors.

• **Final Model Evaluation**: evaluating the performance of the final GLM model on a separate testing dataset.

---

**Figure 1.** Pipeline of Boots-wrapper approach

**Bootstrap-based Feature Set Evaluation**

Given a set of features, the feature set evaluation method include the following steps:

(1) **Bootstrap sampling**: As shown in Figure 2, given a training dataset with \( N \) instances, the dataset is randomly sampled with replacement \( N \) times to obtain a set of data samples, which is referred to as bootstrap-sampled dataset. Accordingly, the remainder validation set includes all the data samples that are in the whole training dataset but not in the bootstrap-sampled dataset.

(2) **Model training**: As shown in Figure 3, the bootstrap sampled dataset is used to train a prediction model (GLM) from the Bootstrap-sampled dataset

(3) **Model validation**: The discrimination of the prediction model (e.g., AUC) is evaluated on the remainder validation dataset.

Step (1)–(3) are repeated \( M \) times.
Figure 2. Examples of Bootstrap Sampling

<table>
<thead>
<tr>
<th>ID</th>
<th>Age</th>
<th>Sex</th>
<th>Stroke</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>76</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>59</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>45</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Bootstrap-sampled dataset

Repeat M times

Validation dataset

<table>
<thead>
<tr>
<th>ID</th>
<th>Age</th>
<th>Sex</th>
<th>Stroke</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>76</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 3. Examples of model training and validation based on bootstrap-sampled data

<table>
<thead>
<tr>
<th>ID</th>
<th>Age</th>
<th>Sex</th>
<th>Stroke</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>59</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>45</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Prediction models

\[ y = \text{logistic} \left( \beta_0 + \beta_1 x_1 + \beta_2 x_2 \right) \]

\[ y: \text{Stroke}, \quad x_1: \text{Age}, \quad x_2: \text{Sex} \]

\[ \beta_1^{(1)} = 0.65 \]

\[ \beta_2^{(1)} = 0.02 \]

\[ \beta_1^{(2)} = 0.62 \]

\[ \beta_2^{(2)} = -0.05 \]

\[ \beta_1^{(3)} = 0.68 \]

\[ \beta_2^{(3)} = 0.10 \]

Validation dataset

<table>
<thead>
<tr>
<th>ID</th>
<th>Age</th>
<th>Sex</th>
<th>Stroke</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>76</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Model discrimination (e.g., AUC)

\[ D^{(1)} = 0.70 \]

\[ D^{(2)} = 0.68 \]

\[ D^{(3)} = 0.66 \]

(4) **Significance statistics of features**: As shown in Figure 4, the coefficients of features from all repetitions are used to compute the p-values of features. For the coefficient \( \beta_i \) of each feature \( i \), we first compute its mean \( \bar{\beta}_i \) and standard error \( \sigma_i \) of all repetitions. Then we compute the p-values by performing z-test: \( p_i = Prob(z \geq \bar{\beta}_i / \sigma_i) \)

(5) **Discrimination Statistics**: The average discrimination statistics (e.g., AUC) of all repetitions \( \bar{D} \) is calculated.

(6) **Performance Evaluation**: The discrimination statistics and the p-values of features are combined to estimate the performance of the feature set:

\[ L = \bar{D} - \alpha \cdot \sum_{i=1}^{k} (I(p_i \geq \theta) \cdot p_i) \]

where \( k \) is the number of risk factors. If \( p_i \geq \theta \) then \( I(p_i \geq \theta) = 1 \), else \( I(p_i \geq \theta) = 0 \). \( \alpha \in [0,1] \) is the parameter to balance the importance of prediction performance and statistical significance. \( \theta \) is the pre-defined threshold for p-value (e.g., 0.1, 0.05, 0.01).
Results

We evaluated the performance of Boots-wrapper in building 1-year ischemic stroke prediction models for AF patients from CAFR dataset. As our dataset is highly imbalanced and has very rare positive instances (3.12% in the testing set), AUC and the area under the precision recall curve (AUPR) were used to evaluate the prediction performance of models, because AUC and AUPR can provide a more informative assessment than accuracy, sensitivity and specificity for imbalanced data. Note that the baseline of AUPR for our testing dataset is the average precision of randomly predicting the risk, which equals the 1-year stroke occurrence rate (i.e., 0.031 for the testing set).

In the experiments, we used AUC as the discrimination metric of the Boots-wrapper method, and set the parameters $M = 100$ and $\theta = 0.05$. Different values of parameter $\alpha$ were tried in our experiments. We use logistic regression, which is a GLM with a logit link function and a binomial distribution, to build prediction models based on the selected features. The proposed Boots-wrapper method was compared with the state-of-the-art feature selection methods, including:

- **Filter.** The p-value from two standard statistical tests, which are the Chi-square test for categorical features and the ANOVA F test for numeric features, were used as the relevancy score to select the features whose p-value is less than a predefined threshold.
- **Lasso.** We used Lasso\textsuperscript{14} to introduce L1-norm regularization to logistic regression for embedded feature selection.
- **Stepwise.** The stepwise selection method for logistic regression was applied to maximize the Score statistics using the whole training dataset and rigidly constrain p-values of selected features to be less than a predefined threshold.
- **Cross-validation (CV) based wrapper.** We applied the cross-validation based wrapper subset selection method\textsuperscript{14} to search the subset of features that can achieve optimized performance (AUC) by cross validation.
Table 1. AUC and AUPR of the multivariate logistic regression models built on different patient sets and feature sets

<table>
<thead>
<tr>
<th>Feature selection method</th>
<th>Parameter</th>
<th>No. of features</th>
<th>Training set</th>
<th>Testing set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>AUC</td>
<td>AUPR</td>
</tr>
<tr>
<td>No feature (i.e., random prediction)</td>
<td></td>
<td>0</td>
<td>0.500</td>
<td>0.040</td>
</tr>
<tr>
<td>CHADS$_2$ score</td>
<td></td>
<td>5</td>
<td>0.662</td>
<td>0.067</td>
</tr>
<tr>
<td>CHA$_3$DS$_2$-VASc score</td>
<td></td>
<td>7</td>
<td>0.665</td>
<td>0.067</td>
</tr>
<tr>
<td>All features</td>
<td></td>
<td>51</td>
<td>0.747</td>
<td>0.129</td>
</tr>
<tr>
<td>Filter</td>
<td>$p &lt; 0.05$</td>
<td>20</td>
<td>0.715</td>
<td>0.106</td>
</tr>
<tr>
<td>Filter</td>
<td>$p &lt; 0.001$</td>
<td>6</td>
<td>0.688</td>
<td>0.093</td>
</tr>
<tr>
<td>Lasso</td>
<td>$C = 0.1$</td>
<td>19</td>
<td>0.726</td>
<td>0.094</td>
</tr>
<tr>
<td>Lasso</td>
<td>$C = 0.02$</td>
<td>9</td>
<td>0.663</td>
<td>0.070</td>
</tr>
<tr>
<td>Stepwise</td>
<td>$p &lt; 0.05$</td>
<td>5</td>
<td>0.701</td>
<td>0.091</td>
</tr>
<tr>
<td>Stepwise</td>
<td>$p &lt; 0.2$</td>
<td>12</td>
<td>0.719</td>
<td>0.107</td>
</tr>
<tr>
<td>CV wrapper</td>
<td></td>
<td>12</td>
<td>0.727</td>
<td>0.093</td>
</tr>
<tr>
<td>Boots-wrapper</td>
<td>$a = 0.1$</td>
<td>6</td>
<td>0.705</td>
<td>0.093</td>
</tr>
<tr>
<td>Boots-wrapper</td>
<td>$a = 0.001$</td>
<td>8</td>
<td>0.713</td>
<td>0.093</td>
</tr>
</tbody>
</table>

Table 2. Multivariate logistic regression models built using different feature selection methods

<table>
<thead>
<tr>
<th>Variable</th>
<th>Model:</th>
<th>Boots-wrapper (0.001)</th>
<th>Stepwise (0.05)</th>
<th>CV wrapper</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>exp($\beta$)</td>
<td>p-value</td>
<td>exp($\beta$)</td>
</tr>
<tr>
<td>Intercept</td>
<td></td>
<td>0.001</td>
<td>&lt;0.001</td>
<td>0.003</td>
</tr>
<tr>
<td>Age /10 years</td>
<td></td>
<td>1.540</td>
<td>&lt;0.001</td>
<td>1.535</td>
</tr>
<tr>
<td>Prior thromboembolism</td>
<td></td>
<td>2.099</td>
<td>&lt;0.001</td>
<td></td>
</tr>
<tr>
<td>Intracranial hemorrhage confirmed by CT/MRI</td>
<td></td>
<td>3.383</td>
<td>0.005</td>
<td>3.361</td>
</tr>
<tr>
<td>Total bilirubin /1 mmol/l</td>
<td></td>
<td>1.032</td>
<td>0.054</td>
<td>1.035</td>
</tr>
<tr>
<td>Fasting blood glucose /1 mmol/l</td>
<td></td>
<td>1.110</td>
<td>0.058</td>
<td>1.125</td>
</tr>
<tr>
<td>Left ventricular posterior wall thickness /1 mm</td>
<td></td>
<td>1.142</td>
<td>0.073</td>
<td></td>
</tr>
<tr>
<td>Left ventricular ejection fraction /10 %</td>
<td></td>
<td>0.868</td>
<td>0.171</td>
<td></td>
</tr>
<tr>
<td>Total cholesterol /1 mmol/l</td>
<td></td>
<td>1.156</td>
<td>0.197</td>
<td></td>
</tr>
<tr>
<td>Prior ischemic stroke confirmed by CT/MRI</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Current smoking</td>
<td></td>
<td></td>
<td></td>
<td>2.242</td>
</tr>
<tr>
<td>Current alcohol drinking</td>
<td></td>
<td></td>
<td></td>
<td>1.754</td>
</tr>
<tr>
<td>Abnormal Left ventricular diastolic function</td>
<td></td>
<td></td>
<td></td>
<td>0.515</td>
</tr>
<tr>
<td>hemoglobin /10 g/L</td>
<td></td>
<td></td>
<td></td>
<td>1.338</td>
</tr>
<tr>
<td>Aspartate transaminase /10 U/L</td>
<td></td>
<td></td>
<td></td>
<td>1.084</td>
</tr>
</tbody>
</table>

1135
In the first experiment, the initial predictor set was set to empty (i.e., no predictor was forced in the models). Then each feature selection method was performed on the original candidate feature set to select predictive features. The wrapper methods (Boots-wrapper, Stepwise and CV wrapper) were performed using a forward hill-climbing search strategy. The performance of the models built on the selected features were also compared with that of the existing risk scores: CHADS², CHA²DS²-VASc, as well as the model built on the whole feature set. As shown in Table 1, Boots-wrapper achieved better AUC and AUPR than the other feature selection methods on the testing dataset, while selecting a small set of features. The prediction performance of our models also outweigh the CHADS² and CHA²DS²-VASc scores.

Table 2 gives the details of some models built on the selected features. As the Stepwise method with the entered p-value < 0.05 rigidly constrains the p-values of selected features, it only selected 5 predictors into the model. In our experiment, we also changed the threshold of entered p-value to 0.2 and selected 12 predictors, but the prediction performance of the model built on the testing set obviously decreased due to over-fitting (AUC = 0.661, AUPR = 0.055). Without constraining the statistical significance of the selected features, the CV-based wrapper method identified 12 features, where 8 of them are not statistical significant, and it also resulted in over-fitting (AUC = 0.681). The proposed Boots-wrapper method selected 8 predictors when setting α = 0.001, where 5 features are not statistical significant, and achieved the best AUC and AUPR on the testing set. When setting α = 0.1, Boots-wrapper selected 6 predictors, where only 1 features are not statistical significant, and also achieved decent testing performance. The results show that Boots-wrapper can balance the model discrimination and statistical significance of predictors, as well as reduce over-fitting.

In the second experiment, we introduced prior knowledge to the models to make the models more interpretable, by adding some known predictors (age, prior thromboembolism, congestive heart failure, hypertension and diabetes) into the initial predictor set (i.e., these predictors were forced in the models). Then each wrapper feature selection method was performed on the remainder candidate feature set to select additional predictive features. As shown in Table 3, Boots-wrapper also achieved better AUC and AUPR than Stepwise and CV based wrapper methods on the testing dataset with prior knowledge. Table 4 gives the details of some models built on the combination of knowledge-based predictors and selected predictors. We can find that some of the knowledge-based predictors are actually not statistical significant in the model built on CAFR dataset. In addition to these known predictors, Boots-wrapper also discovered other features that can improve model discrimination, where some of the selected features are statistical significant. Finally, the prediction performance and predictor number of some models in the above experiments are also compared in Figure 5.

Table 4. Multivariate logistic regression models built using different feature selection methods with prior knowledge

<table>
<thead>
<tr>
<th>Variable</th>
<th>Knowledge (exp(β) p-value)</th>
<th>Boots-wrapper (0.1) (exp(β) p-value)</th>
<th>Boots-wrapper (0.001) (exp(β) p-value)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>0.002 &lt;0.001</td>
<td>0.001 &lt;0.001</td>
<td>&lt;0.001 &lt;0.001</td>
</tr>
<tr>
<td>Age /10 years</td>
<td>1.422 0.001</td>
<td>1.444 0.001</td>
<td>1.475 &lt;0.001</td>
</tr>
<tr>
<td>Prior thromboembolism</td>
<td>2.138 &lt;0.001</td>
<td>2.162 &lt;0.001</td>
<td>2.084 &lt;0.001</td>
</tr>
<tr>
<td>Congestive heart failure</td>
<td>1.311 0.187</td>
<td>1.310 0.190</td>
<td>1.165 0.470</td>
</tr>
<tr>
<td>Hypertension</td>
<td>1.353 0.235</td>
<td>1.339 0.252</td>
<td>1.297 0.313</td>
</tr>
<tr>
<td>Diabetes</td>
<td>1.084 0.700</td>
<td>0.837 0.478</td>
<td>0.848 0.512</td>
</tr>
<tr>
<td>Fasting blood glucose /1 mmol/l</td>
<td></td>
<td>1.144 0.036</td>
<td>1.140 0.042</td>
</tr>
<tr>
<td>Bleeding</td>
<td></td>
<td>1.840 0.059</td>
<td></td>
</tr>
<tr>
<td>Intracranial hemorrhage confirmed by CT/MRI</td>
<td></td>
<td>3.159 0.008</td>
<td></td>
</tr>
<tr>
<td>Total bilirubin /l μmol/l</td>
<td></td>
<td>1.033 0.052</td>
<td></td>
</tr>
<tr>
<td>Left ventricular posterior wall thickness /l mm</td>
<td></td>
<td>1.128 0.108</td>
<td></td>
</tr>
</tbody>
</table>
Table 3. AUC and AUPR of the multivariate logistic regression models built on different patient sets and feature sets with prior knowledge

<table>
<thead>
<tr>
<th>Feature selection method</th>
<th>Parameter</th>
<th>No. of features</th>
<th>Training set</th>
<th>Testing set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>AUC</td>
<td>AUPR</td>
</tr>
<tr>
<td>Knowledge-based features</td>
<td></td>
<td>5</td>
<td>0.692</td>
<td>0.075</td>
</tr>
<tr>
<td>Stepwise + knowledge</td>
<td>$p &lt; 0.05$</td>
<td>8</td>
<td>0.701</td>
<td>0.096</td>
</tr>
<tr>
<td>Stepwise + knowledge</td>
<td>$p &lt; 0.2$</td>
<td>12</td>
<td>0.713</td>
<td>0.110</td>
</tr>
<tr>
<td>CV wrapper + knowledge</td>
<td></td>
<td>17</td>
<td>0.732</td>
<td>0.107</td>
</tr>
<tr>
<td>Boots-wrapper + knowledge</td>
<td>$\alpha = 0.1$</td>
<td>7</td>
<td>0.693</td>
<td>0.084</td>
</tr>
<tr>
<td>Boots-wrapper + knowledge</td>
<td>$\alpha = 0.001$</td>
<td>9</td>
<td>0.701</td>
<td>0.097</td>
</tr>
</tbody>
</table>

Figure 5. AUC and AUPR of different models, evaluated on the testing set
Discussion

In this study, we compared several feature selection methods in developing 1-year prediction models of ischemic stroke for AF patients. The logistic regression method did not work well on the whole feature set due to severe overfitting (training AUC = 0.747, testing AUC = 0.667), but the prediction performance can be improved by appropriate feature selection. However, the Filter and Lasso methods did not achieve obvious performance improvement. That is probably because Filter is a univariate feature selection method and the selected feature set still have high redundancy, while Lasso tends to remain high variance features, such as numeric values of laboratory tests, but some predictive features are binary features with relatively low variance.

We also compared three wrapper methods of feature selection: CV-based wrapper, Stepwise and our proposed Boots-wrapper. The CV-based wrapper method uses cross-validation to evaluate the performance of a subset of features, and does not constrain statistical significance of features. The Stepwise method can constrain the p-value of statistics test (e.g., Wald test) for each feature in a GLM model using hard p-value thresholds. Our proposed method uses a soft penalty to non-significant features to balance the model discrimination and significance of selected features, making the result prediction models more flexible and practical for small and imbalanced dataset. The identified risk predictors, odds ratios and p-values in logistic regression were verified by clinicians, concluding that the selected predictors are interpretable and reasonable to clinicians.

Bootstrap approaches has been widely used in model validation. A filter feature selection method based on bootstrapping was also proposed. In this study, we developed a novel wrapper feature selection method based on bootstrapping, which uses bootstrap sampling to validate model discrimination while evaluating the significance of features by calculating the p-values of bootstrap test. Compared with cross-validation, bootstrap can get a more reliable estimate of the variation in feature coefficients and p-values. Both k-fold cross-validation and leave-one-out cross-validation cannot obtain reliable estimate of the variations. For k-fold cross-validation (e.g. k = 5 or 10), the sample size in the training data could be very different from the whole dataset, therefore the variation could be over-estimated. On the other hand, leave-one-out cross-validation tends to under-estimate the variations since the training sample set at each fold is almost identical. Bootstrap uses the "sample with replacement" approach to overcome the aforementioned issues: 1) the training sample size is the same as the whole dataset; 2) the training sample set at each Bootstrap iteration has reasonable amount of variation. With a reliable estimate of the variations, the selected feature set is more robust. Secondly, conventional statistical tests such as Wald test usually impose distributional assumptions on the data, and are based on asymptotic theories. Distribution of the real-world data may not always fit the distributional assumptions. In addition, the behavior of a test statistic under finite sample could be different from the theoretical results based on asymptotic assumptions (sample size goes to infinity). Bootstrap is a non-parametric approach to estimate the variations and to perform hypothesis testing. No distributional assumption is imposed on the real-world data. Also, by "sampling with replacement", we ensure that the sample size in each bootstrap sample is the same as the real data. Therefore, the test statistic from bootstrap can be more reliable than the ones from conventional methods.

The major disadvantage of the proposed Boots-wrapper method is the high time complexity. As the wrapper framework is very computationally expensive, the method is not practicable for large datasets. Fortunately, it is very easy to parallelize the bootstrap sampling algorithm because each repetition of sampling is independently of each other. Therefore, the method could be run in a distributed system to speed up the computation.

Conclusion

AF is a common cardiac arrhythmias, which increases the risk and severity of ischemic stroke. For predicting ischemic stroke in AF patients, a risk prediction model that can achieve both good model discrimination (e.g., AUC) and statistical significance of predictors is required in real clinical practices. In this paper, we propose a new bootstrap-based wrapper (Boots-wrapper) method of feature selection, and use this method on CAFR data to develop 1-year stroke prediction models in AF. The proposed method can heuristically search a subset of features to maximize the discrimination of the prediction model and minimize the penalty for the non-significant features. To achieve robust feature selection, we perform bootstrap sampling to get a more reliable estimate of the variation and significance statistics. The experimental results show that Boots-wrapper can balance model discrimination and statistical significance of features for developing AF stroke prediction models.
References

A Multi-scale U-Net for Semantic Segmentation of Histological Images from Radical Prostatectomies

Jiayun Li, MS1,2, Karthik V. Sarma, MS1,2, King Chung Ho, MS1,2, Arkadiusz Gertych, PhD3,4, Beatrice S. Knudsen, MD, PhD4,5, Corey W. Arnold, PhD1,2

1Department of Bioengineering, University of California, Los Angeles, CA, USA; 2Computational Integrated Diagnostics, Departments of Radiological Sciences and Pathology and Laboratory Medicine, University of California, Los Angeles, CA, USA; 3Department of Surgery, Cedars-Sinai Medical Center, Los Angeles, CA, USA; 4Department of Pathology and Laboratory Medicine, Cedars-Sinai Medical Center, Los Angeles, CA, USA; 5Department of Biomedical Sciences, Cedars-Sinai Medical Center, Los Angeles, CA, USA

Abstract

Gleason grading of histological images is important in risk assessment and treatment planning for prostate cancer patients. Much research has been done in classifying small homogeneous cancer regions within histological images. However, semi-supervised methods published to date depend on pre-selected regions and cannot be easily extended to an image of heterogeneous tissue composition. In this paper, we propose a multi-scale U-Net model to classify images at the pixel-level using 224 histological image tiles from radical prostatectomies of 20 patients. Our model was evaluated by a patient-based 10-fold cross validation, and achieved a mean Jaccard index of 65.8% across 4 classes (stroma, Gleason 3, Gleason 4 and benign glands), and 75.5% for 3 classes (stroma, benign glands, prostate cancer), outperforming other methods.

Introduction

Prostate cancer is the most common and second most deadly cancer in men in the United States [1]. A key component of prostate cancer staging and treatment selection is the Gleason grading system, in which histopathological slides of prostate tissue are assigned grades that represent the aggressiveness of the cancer. The Gleason scale ranges from Gleason 1 (G1) - Gleason 5 (G5), with a score of G1 indicating cancer that closely resembles normal prostate glands and a score of G5 indicating the most abnormal histopathology, which is associated with the highest mortality risk. Final Gleason scores are generated by summing the first- and second-most prevalent patterns in the tissue sections. Currently, Gleason score is the best biomarker in predicting long term outcome of prostate cancer [2–4]. Yet, a recent clinical trial found no significant difference in mortality at 10 years between patients on active surveillance and immediate surgery [5], underscoring the need for more effective risk stratification tools to improve the outcomes of treatment deferral. Studies have also demonstrated the prognostic value of quantitative pathology features, such as the percent of Gleason 4, for diagnosis and treatment planning [3,6,7]. In addition, Gleason scores are assigned manually through pathologist review, a process that has been shown to have low inter-observer agreement across pathologists, especially when differentiating Gleason 3 (G3) vs Gleason 4 (G4), a distinction that may have substantial impact on further care [8–10].

A computer aided diagnosis (CAD) tool for performing Gleason scoring would provide a repeatable method for grading cancers and may be used as a pre-step for quantitative pathology feature extraction, thus providing a more precise assessment of cancer stage and treatment planning. In this paper, we propose a multi-scale U-Net CNN model for pixel-wise Gleason score prediction. We also compare the performance of several machine learning approaches to Gleason score assignment, including a pixel-wise deep convolutional neural network (CNN), a standard U-Net, the proposed multi-scale U-Net, and the previous work by Gertych, et.al.[11] on 224 histological image tiles from radical prostatectomies of 20 patients.

Previous Work

Previous work has been done in developing an automatic Gleason grading system to help improve diagnosis accuracy and achieve quantitative histological image analysis. A commonly used approach is to extract tissue features and apply classifiers on pre-selected small image tiles, each of which only contains one tissue class. Farjam, et.al. [12] developed a method to segment prostate glands with texture-based features, and then used the size and shape features of glands to classify image tiles into benign or malignant glands. Nguyen, et.al. [13] used structural features of prostate glands
to classify pre-extracted regions of interest (ROIs) into benign, G3, and G4, achieving an overall accuracy of 85.6%.

In the work by Gorelick, et.al. [14], a two stage Adaboost model was applied to classify around 991 sub-images extracted from 50 whole-mount sections of 15 patients. They achieved 85% accuracy for distinguishing high-grade (G4) cancer from low-grade cancer (G3).

However, the above algorithms require a set of pre-extracted image tiles with homogeneous tissue content, which may not be generalizable to larger and more heterogeneous images. Moreover, accurate localization of such small image tiles is a non-trivial problem [15]. Rather than attempting to classify the entire image tile, some efforts focus on segmenting and classifying glands with glandular features such as lumen shape and nuclear density [16,17], but these efforts require well-defined gland boundaries, and may not be applicable for high-grade prostate cancer with few recognizable glands.

Instead of using features from segmented glands, Gertych, et.al. [11] used intensity and texture features from joint histograms of local binary patterns and local variance to segment stroma (ST), prostate cancer (PCa), and benign glands (BN). In their two-stage classifier, a support vector machine (SVM) was trained with local intensity histogram to separate ST and epithelium (EP) areas, and then a random forest (RF) classifier was trained to segment BN and PCa. They obtained an average Jaccard index ($J$) of 59.5% for segmenting ST and EP areas. For separating BN and PCa, they achieved a $J_{\text{BN}}$ of 35.2% ± 24.9%, and a $J_{\text{PCa}}$ of 49.5% ± 18.5% in the test set. Although their model was able to do pixel-wise classification on heterogeneous image tiles, they did not address the problem of differentiating high-grade (G4) versus low-grade cancer (G3).

Additional previous work has explored the use of neural network models to learn features directly, rather than using handcrafted features. Deep convolutional neural network (CNN) models have demonstrated high performance in a variety of natural image analysis tasks [18–21]. Litjens, et.al. [22] implemented a deep convolutional neural network (CNN) to detect cancerous areas on prostate biopsy slide images at 5x magnification. They achieved around 0.90 AUC, but did not address the challenge of distinguishing high-grade versus low-grade cancer.

Pixel-wise deep convolutional networks are difficult to apply to pathology image analysis due to the high resolution of digital pathology slides [23,24], direct analysis of which would generally require more memory than is available on a graphics processing unit (GPU). Two approaches to handling this challenge are resolution downsampling and patch extraction. In down-sampling, high resolution images are scaled down to more manageable sizes, at the cost of the loss of potentially discriminative fine details. In patch extraction, images are divided into (possibly overlapping) sub-patches that are then treated as independent training samples. This approach allows for the analysis of full resolution data, but may lead to an intractable number of potential patches, requiring subsampling of the dataset. In both of these methods, an overall prediction is created for the image or the patch.

Shelhamer and Long, et.al. [24,25] proposed a fully convolutional network (FCN) that can be trained from end to end to output pixel-wise predictions for an entire input image patch (rather than a single prediction for the patch). In order to get dense predictions for each pixel, they used up-sampling operations, and replaced the final fully connected layer with an N×1×1 convolution layer, which output probabilities for N classes. Shallow layer features were fused with deep layer features to mitigate the challenge that intensive up-sampling can lead to coarse segmentation results. The model obtained a mean $J$ of 67.5% and showed 30% improvement on PASCAL VOC 2011 test datasets.

The U-Net architecture proposed by Ronneberger, et.al.[26] extended the FCN by adding a relative symmetric up-sampling path to down-sampling path, creating a U-shaped network architecture. Another important modification of U-Net was the use of an overlap-tile strategy for large image segmentation, in which a slightly larger tile is used as input and predictions are produced for the centered small tile. This method achieved an average $J$ of 77% on a cell segmentation task [26].

While much work has been done in histological image analysis of prostate cancer, few addressed the problem of differentiating high-grade versus low-grade cancer. In this paper, we developed a multi-scale U-Net to predict four tissue classes at once (ST, BN, G3, and G4). We compared the proposed method with a pixel-wise CNN, a standard U-Net and a previously developed model by Gertych et.al.[11] that uses a combined SVM and RF classifier. Our multi-scale U-Net outperformed all other models and achieved the highest mean $J$ of 65.8% across four classes.

Methods

Dataset

Radical prostatectomy specimens from 20 patients with a diagnosis of G3 or G4 prostate cancer according to the contemporary grading criteria [27,28] were retrieved from archives in the Pathology Department at Cedars-Sinai.
imbalance of the dataset, in this methodology, individual potential patches from different classes would have unequal probability of being sampled. Because of the class probability of being sampled. Training was performed using an RMSProp (LR = 0.001, optimizer using Keras [32] with Tensorflow [33] on two NVIDIA Titan X GPUs with synchronous gradient updates and a batch size per GPU of 50 patches. In order to saturate the GPUs during training, patch sampling was run in threads with separate state; one sampler thread was used per GPU. Training was performed over 25 “epochs” of 100,000 patches.

For evaluation, every possible patch was extracted from tiles in the testing set, and any patches that would have extended outside of the bounds of the original tile were discarded. Class predictions were obtained for these patches from the network, and each pixel was assigned a class based on the maximum prediction probability for that pixel.

Semantic Image Segmentation with U-Nets

Convnets produce dense predictions by extracting patches around every pixel, which can be inefficient even for images with moderate size. The FCN proposed by Shelhamer and Long, et.al. [24,25] uses up-sampling and fully convolutional layers to generate pixel-wise predictions efficiently in a single pass. The pooling operation makes CNNs relatively invariant to spatial transformations and also reduces spatial resolution of feature maps. To enable making local predictions with global context, the U-Net [26] extends an FCN with a U-shape architecture, which allows features from shallower layers to combine with those from deeper layers [24,26].

One intuitive way of performing semantic segmentation with FCN is to use the entire image as the input. However, training FCN with large images may require a huge number of samples and also can cause high GPU memory requirement. To solve these problems, large images are divided into several relatively smaller patches, and the overlap-tile strategy is used for seamless segmentation [26]. This requires the size of the patch to be carefully chosen so that the patch can be segmented with sufficient contextual information. Yet, the size of cellular structures such as glands may vary greatly as shown in (Figure 1). To better segment tissue structures with variable size, we propose a multi-scale U-Net architecture that incorporates patches (subtiles) of three different sizes: 400x400, 200x200, and 100x100 to explicitly provide contextual information at multiple scales [34]. To handle border patches that cause one of these patch sizes to extend past the boundary of a given image tile, the tile is padded with reflection of the border [26]. A detailed overview of our multi-scale U-Net architecture is shown in (Figure 2). Instead of taking the whole 1200 x1200 image tile as input, we divided images into 100x100 subtiles and extracted the three patches of varying size around each of these image subtiles. Features from different sizes of patches were then concatenated together and used as...
inputs for the multi-scale U-Net model. The commonly used fully connected layer was replaced by a 4x1x1 convolutional layer that output pixel-wise probabilities for four classes (G3, G3, ST, and BN).

Figure 1. Variations in gland size. (a) shows a tile with heterogeneous Gleason grades (G3, G4 and benign glands). Pathologist annotation mask is shown in (b). The high-grade cancer (G4) areas are shown in red, low-grade cancer (G3) areas are denoted as pink, benign glands are indicated by green, and stroma areas are represented by blue. These images demonstrate the heterogeneity of glands both between grades (e.g. glands A and C) and within the same grade (e.g. glands A and B).

Figure 2. Architecture of the multi-scale patch-based U-Net. The whole image was divided into multiple non-overlapping 100x100 subtiles. To capture contextual information, a 200x200 patch (framed in yellow) and a 400x400 patch (framed in black) were extracted around each centered 100x100 patch (framed in red). Features of different sizes were either down-sampled or up-sampled to 200x200, and concatenated into 64x200x200 feature maps that were input to a U-Net model. The final layer output a 4x100x100 probability map, each channel of which corresponded to a probability map of one class.
In this experiment, we trained two FCN models. The first was the baseline U-Net model that followed an existing work [26]. The other is the multi-scale U-Net. Both models were trained with batch gradient descent (batch size: 25) and backpropagation. A momentum of 0.9 and a learning rate of 0.05 were used. A heuristic was followed to improve the learning of deep neural network model [19], where the learning rate was decreased by 10x when validation errors stopped decreasing. Models were implemented in Torch7 [35], and the training was done on two NVIDIA Titan X GPUs. The dataset of 20 patients was divided into 10 folds resulting in two patients in each fold. This patient-based cross validation ensured independence of training and testing data.

**Evaluation Metrics**

Overall pixel accuracy, mean accuracy for each class, and Jaccard index are three commonly used evaluation metrics for multi-class semantic image segmentation. Overall pixel accuracy measures the proportion of correctly classified pixels, however, it can be biased by imbalanced datasets. Mean single-class accuracy calculates the average proportion of correctly classified pixels in each class, which can also be biased by imbalanced datasets and overestimates the true accuracy due to combining multiple negative classes into one inference class [36–38]. Jaccard index, also known as intersection-over-union, overcomes the limitations of overall pixel accuracy and mean accuracy since it considers both false positives and negatives.

Here, we report Jaccard index and overall pixel-wise accuracy for our models, which can be obtained from a pixel-wise confusion matrix $C$. $C_{ij}$ is the number of pixels labeled as $i$ and predicted as $j$. The total number of pixels with label $i$ is denoted as $T_i = \sum_{j=1}^{N} C_{ij}$, where $N$ is the number of classes. The number of pixels predicted as $j$ is represented as $P_j = \sum_i C_{ij}$ [36]. The Jaccard index for class $i$ is then defined as follows:

$$J_i = \frac{C_{ii}}{T_i + P_i - C_{ii}}$$  

(1)

The overall pixel-wise accuracy (OP) is defined as

$$OP = \frac{\sum_i C_{ii}}{\sum_i \sum_j C_{ij}}$$  

(2)

**Results and Discussion**

**Model Comparison**

For the pixel-wise deep CNN model, class predictions were produced for a testing set comprising 30,170,133 pixels in 37 tiles across 3 patients. For the standard and multi-scale U-Net models, pixel-wise confusion matrices were summed across all 10 folds. In the first evaluation, true positive, true negative, false positive, and false negative rates for each class were calculated for all pixels in the dataset. Gleason 3 and Gleason 4 predictions were summed into a single inference class (PCa) for evaluation. For comparison, results from a baseline SVM + RF model by Gertych et.al. [11] are also included. The Jaccard index and overall pixel accuracy of each model are reported in (Table 1). The analysis was also performed without combining Gleason 3 and Gleason 4 into a single class, with performance shown in (Table 2). In both cases, the same network (trained on separate classes) was used for prediction. The multi-scale U-Net architecture achieved the highest Jaccard index in both segmentation tasks: mean $J = 75.5\%$ for 3 class segmentation and mean $J = 65.8\%$ for 4 class segmentation. Both the U-Net and multi-scale U-Net models outperformed the pixel-wise CNN and the SVM-RF model by Gertych et.al. [11].

**Table 1.** Model performances on segmenting prostate cancer (PCa), benign glands (BN) and stroma (ST).

<table>
<thead>
<tr>
<th></th>
<th>$J_{PCa}$</th>
<th>$J_{BN}$</th>
<th>$J_{ST}$</th>
<th>Mean J</th>
<th>OP</th>
</tr>
</thead>
<tbody>
<tr>
<td>U-Net</td>
<td>74.3%</td>
<td>70.6%</td>
<td><strong>80.1%</strong></td>
<td>75.0%</td>
<td>86.6%</td>
</tr>
<tr>
<td>Multi-scale U-Net</td>
<td><strong>74.7%</strong></td>
<td><strong>72.6%</strong></td>
<td>79.3%</td>
<td><strong>75.5%</strong></td>
<td><strong>86.7%</strong></td>
</tr>
<tr>
<td>Pixel-wise CNN</td>
<td>66.0%</td>
<td>59.0%</td>
<td>71.0%</td>
<td>65.0%</td>
<td>63.9%</td>
</tr>
<tr>
<td>Gertych, et.al. [11]</td>
<td>49.5%</td>
<td>35.2%</td>
<td>59.5%</td>
<td>48.1%</td>
<td>n/a</td>
</tr>
</tbody>
</table>
The approximated inference time for each model is also measured. It took about 2 hours for pixel-wise deep CNN model, around 3 seconds for U-Net model, and 9 seconds for multi-scale U-Net to generate predictions for a 1200x1200 tile on one NVIDIA Titan X GPU. Dense predictions can be much more efficiently produced by FCN.

Table 2. Model performances on segmenting Gleason 4 (G4), Gleason 3 (G3), benign glands (BN), and stroma (ST).

<table>
<thead>
<tr>
<th>Model</th>
<th>J_{G3}</th>
<th>J_{G4}</th>
<th>J_{BN}</th>
<th>J_{ST}</th>
<th>Mean J</th>
</tr>
</thead>
<tbody>
<tr>
<td>U-Net</td>
<td>45.8%</td>
<td>60.9%</td>
<td>70.6%</td>
<td>80.1%</td>
<td>64.4%</td>
</tr>
<tr>
<td>Multi-scale U-Net</td>
<td>49.8%</td>
<td>61.5%</td>
<td>72.6%</td>
<td>79.3%</td>
<td>65.8%</td>
</tr>
<tr>
<td>Pixel-wise CNN</td>
<td>23.0%</td>
<td>25.0%</td>
<td>59.0%</td>
<td>71.0%</td>
<td>45.0%</td>
</tr>
<tr>
<td>Gertych, et.al. [11]^a</td>
<td>n/a</td>
<td>n/a</td>
<td>35.2%</td>
<td>59.5%</td>
<td>47.4%</td>
</tr>
</tbody>
</table>

^a The previous model (SVM+RF) by Gertych, et.al. only addressed three class segmentation by combining G3 and G4 to PCa.

Segmentation results generated by U-Net and multi-scale U-Net for two representative image tiles are shown in (Figure 3). Our models performed well in segmenting different tissue types on image tiles with heterogeneous content, but both models struggled with some border areas due to a lack of contextual information. The small high-grade gland marked by a white arrow in the second row in (Figure 3), for example, was segmented as low-grade gland by both models.

In cases in which global information may be more important for class prediction, the multi-scale U-Net showed superior performance. As shown in (Figure 4), the single input U-Net misclassified areas with dense nuclei on a large benign gland. However, the multi-scale U-Net was able to segment this area correctly. Though both models could segment large irregular high-grade glands very well (Figure 3), they had limited power in segmenting poorly-formed high-grade areas, as shown in the first row of (Figure 5). Models could detect the approximate location of high-grade glands.
cancer, but failed to segment the exact areas. Segmentation performance of both models decreased on tiles with a mixture of small high-grade glands and small low-grade glands. The highest Jaccard indices for G3 and G4 achieved by the multi-scale U-Net were 49.8% and 61.5%, respectively. This reflects the reality that differentiating G3 and G4 is a challenging task, even for pathologists. The inter-observer agreement of clinical pathologists for distinguishing G3 from G4 is between 25% to 47% [11,39]. A larger training dataset that represents more of the natural variance of these cancer grades could allow for improving the models’ ability to discriminate between these classes.

**Conclusion**

In this paper, we addressed the challenge of segmenting different tissue types on heterogeneous histological image tiles by using deep learning techniques. The performance of three different deep learning models (pixel-wise CNN, U-Net, multi-scale U-Net) were evaluated and compared using the Jaccard index and overall pixel accuracy. All three models outperformed a reference algorithm on three-class (ST, BN, PCa) segmentation. Both the U-Net and multi-scale U-Net models achieved a higher Jaccard index than the pixel-wise model. The multi-scale model with three types of inputs (400x400, 200x200, 100x100) showed superior performance as compared with the original U-Net, likely due to its ability to explicitly make use of more global information without overly increasing memory requirements during model training.
There are some limitations in our work. Models were only trained on image tiles, rather than whole histological images. Though our method can be extended to whole image segmentation by splitting these images into non-overlapping tiles, the prediction accuracy for boundary patches could be influenced by lack of contextual information and changes in class balance. Also, our model did not perform as well in segmenting G4 cancer with less differentiated glands. Exploring other approaches, such as the use of two separated models with two scales of inputs [40], could improve performance in the future. We also plan to investigate the influence of global versus local features on predicting dense labels, and will perform further evaluations of our models with whole histological images and extend our algorithm to a computerized tool which can be used to extract reliable and reproducible quantitative features from histological images.

Acknowledgments

The authors would like to acknowledge support from the UCLA Radiology Department Exploratory Research Grant Program (16-0003). KVS acknowledges support from an AMA Foundation Seed Grant, NIH NCI F30CA210329, NIH NIGMS GM08042, and the UCLA-Caltech Medical Scientist Training Program.

References


32. Chollet F. Keras.


A hybrid Neural Network Model for Joint Prediction of Presence and Period Assertions of Medical Events in Clinical Notes

Rumeng Li\textsuperscript{1}, Abhyuday N Jagannatha\textsuperscript{1}, Hong Yu, PhD\textsuperscript{2,3}

\textsuperscript{1}University of Massachusetts, Amherst, MA, USA; \textsuperscript{2}Department of Quantitative Health Sciences, University of Massachusetts Medical School, Worcester, MA; \textsuperscript{3}Bedford VAMC and CHOIR, Bedford, MA, USA

Abstract

In this paper, we propose a novel neural network architecture for clinical text mining. We formulate this hybrid neural network model (HNN), composed of recurrent neural network and deep residual network, to jointly predict the presence and period assertion values associated with medical events in clinical texts. We evaluate the effectiveness of our model on a corpus of expert-annotated longitudinal Electronic Health Records (EHR) notes from Cancer patients. Our experiments show that HNN improves the joint assertion classification accuracy as compared to conventional baselines.

Introduction

In recent years natural language processing techniques have demonstrated increasing effectiveness in clinical text mining. Electronic health record (EHR) narratives, e.g., discharge summaries and progress notes contain a wealth of medically relevant information like diagnosis information, adverse drug events etc. Automatic extraction of such information and representation of clinical knowledge in standardized formats could be used for a variety of purposes like clinical event surveillance and decision support, pharmacovigilance and drug efficacy studies etc. Extracting information from EHR narratives presents challenges unique to this domain. Unlike open-domain data, clinical text normally contain a substantial amount of in-domain terminology and domain-specific knowledge. Therefore, accurately recognizing and understanding the medical entities in clinical text become essential for useful information extraction.

This work addresses the problem of identifying assertion in EHRs. Assertion is an important attribute to any event in information extraction. In EHRs, assertion can be understood as a physician’s belief status with regards to a particular patient’s medical problem. Specifically, as shown in Table 1, a medical problem could be current or happened in the past. The problem could be present, absent, or hypothetical and conditional. Knowing the assertion status of a clinical event (e.g., bleeding) is important for physicians to make clinical decisions (e.g., prescribing anticoagulants). Therefore assertion identification of clinical events is critical for information extraction and data mining from EHRs. Assertion identification was also one of the 2010 i2b2/VA challenge task\textsuperscript{1}, which classified assertion as present, absent, possible, conditionally present, hypothetically present and not associated with patients. In our task, we classify the presence status of a clinical event conditioned on the current time period. We have the same six categories for presence assertion as the 2010 i2b2/VA challenge task, and we have extra four classes of period assertion for the same clinical event as current, history, future, and unknown to capture the temporal information about statements. To better describe the task and our model, we present an exemplary explanation in Table 1.

<table>
<thead>
<tr>
<th>Adverse Drug Event</th>
<th>Period</th>
<th>Presence</th>
</tr>
</thead>
<tbody>
<tr>
<td>He has fever (caused by the drug)</td>
<td>Current</td>
<td>Present</td>
</tr>
<tr>
<td>He had fever (due to the drug)</td>
<td>History</td>
<td>Present</td>
</tr>
<tr>
<td>He has no fever (from the drug)</td>
<td>Current</td>
<td>Absent</td>
</tr>
<tr>
<td>His fever (caused by the drug is) resolved</td>
<td>History or current</td>
<td>Present or Absent</td>
</tr>
<tr>
<td>He has a fever, (possibly caused by the drug)</td>
<td>Current</td>
<td>Possible</td>
</tr>
<tr>
<td>He might have a fever</td>
<td>Current</td>
<td>Possible</td>
</tr>
<tr>
<td>If he is infected/(takes the drug), he will run a fever</td>
<td>Future</td>
<td>Conditional</td>
</tr>
<tr>
<td>He may develop a fever (with this drug)</td>
<td>Future</td>
<td>Hypothetical</td>
</tr>
</tbody>
</table>
Table 1 shows the period and presence assertion categories and corresponding representative texts to the clinical event “fever” (an adverse drug event). Our goal is to develop natural language processing (NLP) approaches to automatically identify both the belief status of a clinical event and its period status. It is important to identify both types of assertions as the task represents a more accurate scheme for reasoning about the physician’s belief status of the patient’s medical problem. For example, to identify the assertion in the sentence “His fever caused by the drug is resolved”, one needs to consider the dependency between period and presence. In this particular scenario, “fever” is present if period is history. In contrast, “fever” is absent if period is current. Also as the presence and period assertions are related to each other or even conditionally dependent on each other, a joint learning model can be of advantage. We incorporate such relations in our model and propose a neural network based framework to jointly predict the two types assertions for a given medical entity.

Previous efforts for the assertion identification (or classification) task include rule-based and machine-learning-based methods. Rule-based approaches required hand-crafted rules, which limited their performance. Therefore, it is no surprise that in the 2010 i2b2/VA challenge task, eight of the top 10 participating systems employed machine-learning approaches (e.g., SVM-based classifier, sometimes employing millions of features). With the recent advance in deep learning, neural network models have shown in automatically capturing semantics and syntax as compared to traditional SVM based models, which require significant feature engineering. Additionally, neural network models also have the added advantage of capturing long-distance dependency in text.

In this study, we explored deep learning models. Specifically, we used recurrent neural network with gated recurrent unit (GRU) to represent a clinical event using the left and right context of the event in its sentence. For each generated hidden unit, residual neural network was used for better representation generation. After combining the entity and context representations with additional attention weights, the framework outputs two labels for presence and period assertion. In this architecture, the two tasks leveraged a common feature-set generated by the recurrent neural networks and then used different attention weights for each assertion task. We also used two extra parameters to add mutual influence for the final prediction of the two types of assertion. Experiments on an expert-annotated EHR narratives show the effectiveness of our deep learning model. We provide more details about our model in Section 3.

The main contributions of this paper include a novel neural network architecture that not only leverages recurrent residual network for assertion classification task, but also jointly predicts both the presence and period assertions in one framework. Our method obtains good results on both types of assertion classification tasks. The rest of the paper is organized as follows: we present the related work in Section 2 and introduce our proposed model in Section 3. In Section 4 we report the experiment results and our analysis. And we conclude the paper in Section 5.

Related Work

Medical Assertion Classification

Determination of the assertion status of clinical events is an important area of clinical NLP research. Previous efforts mainly include rule-based methods and machine learning approaches. Popular rule-based methods include the NegEx algorithm and ConText algorithm. The NegEx algorithm is a simple regular expressions algorithm to determine whether a medical entity is present or absent in a patient. The ConText algorithm extends the NegEx algorithm to detect four assertion categories: absent, hypothetical, historical, and not associated with the patient. Uzuner et al. studied the rule-based Extended NegEx system and a SVM-based Statistical Assertion Classifier (StAC) and showed that a machine learning approach achieved competitive results for assertion classification. Four assertion classes as present, absent, uncertain in the patient, or not associated with the patient were used in their system. Wu et al. conducted a multi-corpus analysis of negation detection and concluded that it was easy to optimize for a single corpus but not to generalize to arbitrary clinical text.

The 2010 i2b2/VA Challenge designed a specific assertion classification task. For each “problem” concept mentioned in a clinical text, systems were built to classify the concept’s status associated with the patients as “present”, “absent”, “possible in the patient”, “conditionally present”, and “hypothetically present”, or mentioned in the patient report but “associated with someone other than the patient” based on the context that describes it. The task as a multi-class categorization problem allows the use of machine learning classification methods. SVMs were still the common theme for the task. For some SVM models, millions of features were employed from lexical, syntactic to contextual level.
In addition to those text features, concept-mapping features derived from existing annotation tools like cTAKEs, MetaMap (UMLS) etc. are also used. The best system addressed the task in two stages: in stage 1, assertion class predictions were generated for every word that was part of a “problem” concept by using three parallel different SVM classifiers. In stage 2, a secondary classifier predicted a class for the complete concept, based on the separate per-word predictions from the ensemble of stage-1 classifiers, by using SVM-multi-class with a linear kernel. Later Kim et al. revised their participating system and added specific features to improve the performance on minority classes (e.g., the conditional class) and obtained better results.

Our task is more challenging than the 2010 i2b2/VA Challenge assertion task in that we not only classify the presence assertion, we also jointly classify the period assertion as being “Current”, “History”, “Future”, or “Unknown” in a joint model architecture.

Related Neural Networks

Our deep learning model is based on neural network models to learn feature representations for clinical events and their context for classification. Our model is related to learning representations for long text (sentence/paragraph/document), an important task which draws much efforts. Recurrent neural networks (RNN), and their variants are widely used. Closely related work to our model is the recurrent neural network with gated recurrent unit (GRU) and deep residual network.

RNN can be used effectively to learn distributed representations over a variable-length sequence. At each time-step, it takes both the output of the previous step and the current token as input, convolutes the inputs, and forwards it to the next step. A gated recurrent unit (GRU), a variant of RNN, was proposed by Cho et al. to make each recurrent unit to adaptively capture dependencies of different time scales. It has similar unit as the long-short term memory unit (LSTM) with two gating units named reset and update gates modulating the flow of information inside the unit. However, without having a separate memory cells as LSTM. GRU has been proved comparable results and faster training with less parameters than LSTM. Figure 1 provides an illustration of RNN and GRU.

The deep residual network has two significant characters compared with RNN. The first one is the residual learning. In a neural network model, normally the data is passed from one layer to the adjacent layer. In the residual network, an additional layer is used to connect layers that are far away. During the back propagation, errors can be passed from a higher layer to a lower layer directly. This character is of advantage as it may capture the long-distance context determining the assertion. The second character is the depth of such models. A typical residual network has hundreds of layers, which is much deeper than most existing models. Thus the training of such a model becomes a challenge. In addition, given the number of the layers, the number of parameters also exceeds most networks. When trained on a small dataset, a deep residual network may suffer from over-fitting. The residual neural network has been proved useful in capturing information from images for classification. A number of variants have been introduced for a series of tasks, we do not go in to the details due to the space limitation.

Attention mechanism is also adopted in our system. Attention mechanisms in neural networks are inspired by the presence of attention in human visual system. Human beings’ visual system is able to focus on the most salient part of an image and adjust the focal point over time. The concept of “attention” has gained popularity in training neural networks and have been applied to various computer vision and NLP tasks and we don’t enumerate here.
Figure 2: A Residual Learning Block. Compared with a traditional multi-layer perceptron, the change is that some layers that used to be non-adjacent are connected.

In the following Section, we will show how to employ the recurrent neural network with gated recurrent unit and residual network in the targeted task.

Methods

We address the assertion classification task as a supervised learning problem and build one framework to jointly predict both presence and period assertion labels. Given a sentence with an annotated clinical event, the model predicts both the presence tag and the period tag for the event. The presence tag is chosen from six categories and the period tag has four categories.

Given a clinical event and its sentence, our HNN model use three recurrent neural networks with gated recurrent unit (GRU) and residual networks to generate representations respectively for the clinical event, sentence tokens to the left of the clinical event entity, and the right. These representations are further passed to following layers conducting interactions and predicting the tags.

For better illustration, we provide Figure 3 and Figure 4 to show how the proposed HNN model works for the medical entity “fever” in the sentence “He has fever caused by the drug”. The sentence is splitted into three parts, the entity, its left neighbor and its right neighbor. We firstly have a representation for the entity “fever”, as this is a single word, RNN GRU is not needed to generate the representation, we can just use the embedding of “fever” instead. But for an entity with longer sequence, for example like “lung cancer” or any entity longer, an RNN model is needed for generating the representation. Two RNNs are also employed for “He has” and “caused by the drug” which are the left and right neighbors of the entity. The three representations we get from the three RNNs are fed to the following part of the network as shown in Equation 1. In Equation 1, R_{Block} stands for residual learning block. L stands for linear function.

\[
\begin{align*}
    h_{i_{tmp}} &= GRU(h_{i-1}, x_i) \\
    R_{Block}(h_{i_{tmp}}) &= \\
    &\text{Dropout}(L(Dropout(L(\text{tanh}(h_{i_{tmp}})))))) \\
    \text{repeat twice} \\
    \text{ResidualNet}(h_{i-1}, x_i) &= R_{Block}(h_{i_{tmp}}) \\
    \text{repeat 7 times} \\
    h_i &= \text{ResidualNet}(h_{i-1}, x_i)
\end{align*}
\]

Equation 1 and Figure 4 explain the residual recurrent neural networks with gated recurrent unit (GRU) we used in this work.

Their representations are then fed to the linear transformation node which outputs the final score. The linear function

---

1Please note that in this example sentence, “drug” is also a medical entity, but we just address the entity “fever” as an example here.
takes as input the final left representation $out_{L_i} = h_i^{final}$, final right representation $out_{R_j} = h_j^{final}$ and the final representation for the medical entity $out_{E_k} = h_k^{final}$, and outputs the comprehensive representation for this entity $R_e$ as in Equation 2.

$$R_e(out_{L_i}, out_{R_j}, out_{E_k}) = f(out_{L_i}W_0 + out_{R_j}W_1 + out_{E_k}W_2 + b)$$ (2)

Then we get to consider features that are specific to each sub-task. We assume that for each sub-task, the model should attend to different parts of the original sentence. In the mean time since the two sub-tasks are related as we have stated, the knowledge what sub-task A is attending to will help improve the performance of sub-task B, we formulate the representations of the two sub-tasks to interact with each other. Equation 3 defines the details.

**Step 1**

$$R_e^1 = f_1(R_e) + \sum_{i=1,n} att^1_i w_i; w_i \in S$$

$$R_e^2 = f_2(R_e) + \sum_{i=1,n} att^2_i w_i; w_i \in S$$ (3)

**Step 2**

$$R_e^1 = R_e^1 + \alpha R_e^2$$

$$R_e^2 = R_e^2 + \beta R_e^1$$

We use $S$ to represent the sentence which contains the entity. In Equation 3 $n$ is the number of its containing words. $w_i$ is the embedding of the $i^{th}$ word in $S$. We use $R_e$ to distribute attentions over the $n$ words in $S$. The weights are calculated using a softmax over the sum of all $w$ in $S$. $att^2 = Softmax(Linear^3([R_e \cdot \sum_{w_i \in S} w_i])$. Step2 defines our way to interact the two sub-tasks with weights $\alpha$ and $\beta$. $\alpha$ and $\beta$ are learned according to $\alpha = WR_e^2 + b; \beta = W'R_e^1 + b'$.

To boost performance, we add some basic features here including the entity position in the sentence, entity length, entity bag-of-words features, sentence length, number of nouns in the sentence, number of verbs in the sentence, the verb tense in the sentence, and part-of-speech tags of words in the sentence. They are concatenated with $R_e^1$ and $R_e^2$.

The score for each label was fed into the softmax classifier using $R_e^1$ and $R_e^2$ to make predication of the presence and period assertion.

Equation 4 defines our loss function:

$$loss = hinge_{loss}(pred_{label1}, gold_{label1}) + hinge_{loss}(pred_{label2}, gold_{label2})$$

$$hinge_{loss} = \frac{1}{N} \sum_{n=1}^{N} \sum_{k=1}^{K} \left[ \max(0, 1 - \delta(\{l_n = k\}t_{nk})) \right]^p$$ (4)

**Experiments**

**Datasets**

We used an annotated corpus of 1089 EHR notes. The corpus comprises of a three-year longitudinal provider notes of 21 cancer patients, which include progress notes and discharge summaries – essentially all note types in the longitudinal EHRs of the cancer patients. Each note was annotated by two clinical professionals. For each annotated medical entity, including four different types respectively as Drug (Medication information of drug and its attributes), Indication (reason for prescribing medication), ADE (Adverse Drug Event as an injury resulting from the normal use
Figure 3: The architecture of the proposed model.

Figure 4: How the residual network functions in the proposed model. Note that it contains dropout which is not shown in this figure explicitly.
of a drug at a normal dose), and SSLIF (other signs which the physician can observe and symptoms which the patient reports) in the corpus, presence and period assertions are used together to assign property to it. Presence assertions include six categories which are respectively present, absent, possible, conditional, hypothetical, and not associated with patient. The period assertions include four categories as current, history, future, and unknown. Presence and period assertion annotation statistics in the corpus are provided in the Table 2 and Table 3.

**Table 2:** Distribution of Presence Assertions.

<table>
<thead>
<tr>
<th></th>
<th>Training No.</th>
<th>Training Proportion %</th>
<th>Testing No.</th>
<th>Testing Proportion %</th>
<th>Total No.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Present</td>
<td>13960</td>
<td>52.09</td>
<td>6978</td>
<td>52.07</td>
<td>20938</td>
</tr>
<tr>
<td>Absent</td>
<td>10680</td>
<td>39.85</td>
<td>5391</td>
<td>40.23</td>
<td>16071</td>
</tr>
<tr>
<td>Hypothetical</td>
<td>582</td>
<td>2.17</td>
<td>307</td>
<td>2.29</td>
<td>889</td>
</tr>
<tr>
<td>Possible</td>
<td>679</td>
<td>2.53</td>
<td>315</td>
<td>2.35</td>
<td>994</td>
</tr>
<tr>
<td>Conditional</td>
<td>68</td>
<td>0.25</td>
<td>36</td>
<td>0.27</td>
<td>104</td>
</tr>
<tr>
<td>Not Patient</td>
<td>830</td>
<td>3.10</td>
<td>373</td>
<td>2.78</td>
<td>1203</td>
</tr>
</tbody>
</table>

**Table 3:** Distribution of Period Assertions.

<table>
<thead>
<tr>
<th></th>
<th>Training No.</th>
<th>Training Proportion %</th>
<th>Testing No.</th>
<th>Testing Proportion %</th>
<th>Total No.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Current</td>
<td>20881</td>
<td>77.92</td>
<td>10485</td>
<td>78.25</td>
<td>31366</td>
</tr>
<tr>
<td>History</td>
<td>5156</td>
<td>19.24</td>
<td>2503</td>
<td>18.68</td>
<td>7659</td>
</tr>
<tr>
<td>Future</td>
<td>696</td>
<td>2.60</td>
<td>369</td>
<td>2.75</td>
<td>1065</td>
</tr>
<tr>
<td>Unknown</td>
<td>66</td>
<td>0.25</td>
<td>43</td>
<td>0.32</td>
<td>109</td>
</tr>
</tbody>
</table>

**Experiments Setup**

Our experiments used word embeddings of 100 dimensions learnt on a combined corpus of PubMed open access articles, English Wikipedia and an unlabeled corpus of around hundred thousand Electronic Health records which are not the EHR used in our dataset\(^1\). We used Chainer, a flexible framework for neural networks for the implementation of the proposed model\(^2\). We report the micro-averaged and macro-averaged accuracy results and also the performance on each class.

**Experiment Results**

For comparison, we used a standard SVM and a standard LSTM model as baseline. To be specific, the Linear SVM classifier\(^1\) with bag-of-words features is used. The LSTM is implemented using Chainer. It runs through the whole sentence, without splitting the sentence according to entities as what we do in the proposed model. Bag-of-words features is also employed. In addition, the presence assertion and period assertion classification task are conducted separately for the baseline systems.

Experiment results for presence assertion is shown in Table 4 and results for period assertion is shown in Table 5. We use HNN as abbreviation for our model. As can be seen, the proposed model by jointly training the two task in one unified framework obtain good results in both two tasks.

**Analysis**

As mentioned in the previous sections, HNN addresses the combined classification tasks of presence and period assertion in one joint framework. Assertion prediction is a complex task from the NLP perspective. The difficulty

---

\(^2\)https://chainer.org/

\(^1\)Trained using LIBSVM\(^2\)
of this task is further exacerbated by the fact that we predict assertions on four different underlying medical entities (Drug, ADE, Indication and SSLIF). The difference among these entity types induces a domain difference in assertion prediction which increases the complexity of our task.

The HNN model in general achieves competitive results on both tasks. The overall result of all methods on the period task is lower than that on the presence task. One possible factor behind this discrepancy is that an EHR document is largely written in past or present perfect tense, irrespective of the relative timeline of events in that document. Due to this ambiguous use of tense in EHRs, learning algorithms are often forced to rely on contextual clues, which are more difficult to recognize and extract. Presence assertions on the other hand have more direct linguistic clues such as negation. However, even though the period assertion task is a more difficult one, our system is able to outperform the baseline by a large margin. In fact, the performance gap between HNN and the next best baseline (around 3.4% micro accuracy and 18.4% macro accuracy) for period task is significantly higher than that in presence task. This indicates that HNN can learn to recognize a more diverse set of contextual patterns, and is less dependent on tense based indicators.

The SVM model is ineffective for presence and period assertion in minority classes. LSTM model also only shows small improvements in minority class prediction. In contrast, our HNN model leads to significant improvement on several minority classes such as “Not Patient”, “Possible”, “Future” and “Unknown”. This performance indicates that HNN can better capture the semantics of minority class and subsequently learn to generalize better from fewer examples.

All the three models get zero accuracy on the conditional and hypothetical classes (except a very small accuracy of our model on conditional class) because these two are the rarest classes in our dataset. As a result, our data-driven models do not have enough corresponding positive samples for meaningful training.

Conclusion

We have proposed a novel hybrid neural network framework for prediction of assertions in Electronic Health Records. Our model can jointly predict both the presence and period assertion values ascribed to the medical entities associated with patients in clinical texts. Our experimental results show that the HNN model leads to significant improvement in both minority and majority class for period and presence assertion.
Acknowledgement

We thank the UMassMed annotation team: Feifan Liu, Elaine Freund, Weisong Liu, Steve Belknap, Nadya Frid, Alex Granillo, Heather Keating, and Victoria Wang for creating the gold standard evaluation for this work. We also thank the anonymous reviewers for their comments and suggestions. This work was supported in part by the grant HL125089 from the National Institutes of Health. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsor.

References


17. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems 2013 (pp. 3111-3119).
SpindleSphere: A Web-based Platform for Large-scale Sleep Spindle Analysis and Visualization

Xiaojin Li, MS\textsuperscript{1,2}, Licong Cui, PhD\textsuperscript{1,3}, Shiqiang Tao, PhD\textsuperscript{1}, Ningzhou Zeng, MS\textsuperscript{1,3}, Guo-Qiang Zhang, PhD\textsuperscript{1,3}

\textsuperscript{1}Institute for Biomedical Informatics, University of Kentucky, Lexington, KY
\textsuperscript{2}Department of Electrical Engineering and Computer Science, Case Western Reserve University, Cleveland, OH
\textsuperscript{3}Department of Computer Science, University of Kentucky, Lexington, KY

Abstract

Sleep spindles are a hallmark of stage 2 non-REM sleep that contain information about heritable traits that play an important role in neurological diseases. One of the key challenges in leveraging spindles for clinical research is the lack of a data processing pipeline and web-based platform for managing and visualizing spindle-specific data at scale. We propose SpindleSphere, a scalable integrated data management and visualization platform for spindle research. SpindleSphere has several features: (1) standardized, metadata-based search and query of annotated polysomnography (PSG), the gold standard for sleep diagnosis; (2) event-specific signal exporting; (3) interface for interactive waveform visualization; (4) multi-scale spindle rendering; and (5) parallel algorithm in MapReduce for detection of spindle segments. SpindleSphere provides real-time visualization of multi-modal signals from National Sleep Research Resource (NSRR) for spindle characterization. Preliminary evaluation of SpindleSphere was performed on the NSRR (130 GB of PSG data from 300 subjects).

1 Introduction

Sleep spindles (or spindles, as shown in Figure 1) are one of the most important clinical phenomena in sleep studies\textsuperscript{1}. Sleep spindle characteristics (e.g. spindle density) are reflected in many disorders such as mental retardation\textsuperscript{2}, epilepsy\textsuperscript{3}, schizophrenia\textsuperscript{4,5}, and sleep disorders\textsuperscript{6,7}. Spindles also play an important role in memory reorganization and consolidation\textsuperscript{8–10}. As an electroencephalographic (EEG) phenomenon, the spindle is traditionally characterized as a group of rhythmic signals from between 12 and 14 Hz with a maximum of 3 seconds in duration\textsuperscript{11}. Spindles are most frequently observed during stage 2 Non-rapid eye movement (NREM) sleep (Rechtschaffen and Kales rules\textsuperscript{12}). Spindles are traditionally characterized as one of the hallmarks of stage 2 NREM sleep. Changes in spindle characteristics result from interactions of several regions of the brain, which are linked to learning, behavioral arousal, and sensory gating during wakefulness\textsuperscript{13}. Therefore, identifying variations in spindle characteristics could have diagnostic value as a biomarker for disease states.

Spindle analysis typically involves the following steps (Figure 2). First, the overnight PSG data (European Data Format (EDF) files) are obtained from the data repository and stored on a local machine. Second, using the annotation file, overnight or partial signals of useful channels or montages are extracted and then used as the input for spindle detection algorithms. After spindle detection, detected spindles are visualized and reviewed using existing EDF tools or other software. The last step allows the user to perform specific analysis using verified spindles. Using this workflow, there are several challenges including storage limitations, redundancy and low scalability, data management and visualization limitations, and data format and cross-platform limitations.

Storage limitations. During data acquisition, each EDF file is stored but only a small part is used in spindle detection. Therefore, storing whole EDF file dataset causes storage issues when detecting spindles with large datasets, and simply expanding storage will not allow us to cope with the rapidly expanding volume and increasing complexity of biomedical data.

Redundancy and low scalability. Storing datasets as files, at the abstraction layer, presents barriers that make data access, processing and analysis tasks even more difficult, redundant, and ineffective. Current spindle detection programs...
extract a specific channel or montage from an EDF file for every independent spindle detection study, which is a redundant and time-consuming process.

**Spindle visualization and reviewing limitations.** Most sleep-related datasets do not provide spindles that have been identified and verified by experts. As such, it is difficult to handle large datasets using existing EDF tools in visualizing and reviewing the automatically detected spindles, especially comparing the performance of different automatic spindle detection approaches. Moreover, most existing EDF tools are stand-alone software, which makes it a challenge to share the detected spindles since it requires both sides have the same software and signal files.

![Sleep spindle analysis (on the local machine)](image)

**Figure 2:** The traditional workflow of spindle analysis.

In this paper, we propose a scalable integrated data management and visualization platform (SpindleSphere) for spindle research to address the challenges presented here with a new efficient and effective spindle analysis workflow. SpindleSphere is a web-based application developed using Ruby on Rails framework and a special spindle format as the back end. It has a set of key characteristics described below that are not available elsewhere.

**A scalable spindle processing pipeline,** which includes 1) standardized, metadata-based search and query of annotated PSG signal data. We specially designed new data schemas for efficient storing, exporting and visualizing PSG data in our database; 2) interfaces for querying, importing, event-specific extraction, and repackaging of PSG data. Files that only contain useful signal data require less storage space and can be partially downloaded instead of using the entire large EDF file; in addition, there is no longer a need to execute the redundant and time-consuming step of extracting signals from specific channels.

**A web-based spindle analysis and visualization platform,** which includes 1) online visualization of single, multiple, or cross-sectional event-specific waveforms; 2) multi-scale spindle visualization and reviewing interface, which, for the first time, provides an intuitive and efficient solution for spindle reviewing, verification and comparison in sleep-related studies; 3) leveraging the advantages of the Web application, it is unnecessary to download and install specific EDF tools, so the results can easily be shared with others; moreover, the spindles can be reviewed on any computer with a browser on it, i.e. no platform limitations.

**Large-scale processing of PSG data with parallel computing.** We designed a key-value-based JavaScript Object Notation (JSON) format for representing spindle data, which breaks through format limitations (existing EDF is not compatible with input file format for parallel computing) and is suitable for both sequential and parallel computing. In order to verify the feasibility of SpindleSphere on a parallel programming model, we designed a parallel spindle detection algorithm based on Mölle’s work. Then we applied both sequential and parallel algorithms to over 130GB of PSG data from NSRR and evaluated the performance. The evaluation results indicate that the data format in SpindleSphere supports cloud computing models and provides high scalability for large-scale dataset processing.

## 2 Background

### 2.1 National Sleep Research Resource (NSRR)

The National Sleep Research Resource is a comprehensive, well-integrated retrospective repository of sleep and sleep-related data, consisting of bio-physiological signals linked to risk factor and outcome data for participants in nearly 20,000 sleep studies with thousands of overnight sleep recordings, collected as a part of prior and ongoing NIH-supported efforts. The NSRR data repository encompasses a wide variety of signals, data collection protocols, and processing algorithms, thus representing a significant but under-utilized resource of “big data.”

For this paper, we use a dataset that consists of 300 subjects from the Cleveland Children’s Sleep and Health Study (CCSHS). The CCSHS is one of the largest population-based pediatric cohorts with objective sleep studies. The
cohort is a stratified random sample of full-term and preterm children born between 1988 and 1993 at 3 hospitals in the Cleveland area. The cohort includes 517 children, studied at 8 to 11 years of age with in-home sleep studies, acoustic reflectometry, blood pressure, and other sleep-related measures. The study was approved by the Institutional Review Boards of all participating institutions.

2.2 Automatic spindle detection algorithm

It is difficult, time-consuming and labor-intensive to recognize and count sleep spindles in an overnight recording visually, let alone perform these identifications within a large sleep-related dataset, such as NSRR. Additionally, the visual analysis may produce subjective results since it is an operator dependent task. Therefore, developing an automatic sleep spindle detection system can reduce manual work required in sleep studies. Several approaches for spindle detection from the EEG waveform have been proposed.

In order to enhance spindle activity during the depolarizing phase, Mölle et al. proposed an approach that considered the temporal dynamics between faster rhythmic activities and slow oscillations during NREM sleep. Acir and Güzeliş proposed a two-stage classification procedure for automatic recognition of sleep spindles using artificial neural networks (ANN). Duman et al. proposed an algorithm integrating techniques in a decision tree: short time Fourier transform and multiple signal classification algorithms. Wamsley et al. implemented a wavelet-based approach with a complex Morlet wavelet. Tsana and Clifford proposed a method based on continuous wavelet transform with a Morlet kernel function, then identified spindles with a large power of the coefficients in frequencies. In this paper, we designed a parallel algorithm based on Mölle’s work to evaluate the performance of the new workflow using SpindleSphere.

2.3 Apache Hadoop and MapReduce

Apache Hadoop is an open-source framework using the MapReduce programming model, which was developed for processing large-scale datasets. MapReduce is an implementation for processing big datasets using a parallel and distributed algorithm; it consists of two functions named map and reduce. A complex and intensive task is divided into several independent tasks that are executed on distributed nodes in a cluster using the map function, and the results are summarized using the reduce function. The Hadoop framework uses Hadoop Distributed File System (HDFS) for large-scale data storage and management, and it provides a reliable and scalable solution for parallel applications.

In this paper, we used MapReduce to process multiple procedures in spindle detection in parallel to improve the performance of the approach.

3 Methods

3.1 A new spindle analysis workflow using SpindleSphere

To address the limitations in traditional workflow, we propose a new efficient and effective spindle analysis workflow with SpindleSphere shown in Figure 3, which includes 1) PSG data from NSRR that will be imported to our EDF database, with a scalable processing pipeline that associates related specific metadata and provenance information to each signal segment; 2) user requests are processed by a database consisting of metadata-annotated signal segments; 3) users can download event-specific signal segments to their local machine for further analysis; 4) user-contributed and additional spindle data source can be imported to our database; 5) online visualization of single, multiple, or cross-sectional event-specific waveforms, and multi-scale spindle visualization and reviewing.

**Figure 3:** The new workflow of sleep spindle research.
3.2 System architecture

SpindleSphere is a scalable integrated data management and visualization platform for spindle studies. It provides an intuitive web interface for clinicians and researchers to query, visualize and review detected spindles with large PSG signal datasets. Figure 4 shows the system architecture of SpindleSphere: 1) backend database structure to store parsed EDF files that are systematically linked to NSRR; 2) query engine for processing user requests, such as annotation-based signal rendering and exporting of label-specific signal segments; 3) interface design and information display since PSG data are content rich, especially after they are carefully annotated. Overall, the SpindleSphere platform used Model View Controller (MVC) architecture pattern and was implemented using Ruby on Rails technology and Data-Driven Documents (D3), which is an open source JavaScript library for data visualization. Moreover, we addressed scalability of SpindleSphere in two ways: using a highly scalable database system as the backend and support for parallel computing. We describe different components of SpindleSphere in following subsections.

3.3 Data model

For data storage, we selected MongoDB\textsuperscript{21} (version 3.2.7) as the backend database for data storing and management. MongoDB is a free and open-source cross-platform document-oriented database program. It is classified as a NoSQL database program and uses JSON-like documents with schemas, which provide flexible data modelling, considerable scalability and improved data management performance. Three different data models in MongoDB were built to store each signal fragment with its EDF header and annotations. The main idea is to break large, unstructured, sequential data files into minimal, semantically meaningful, fragments. Such fragments can be indexed, assembled, retrieved, rendered, or repackaged on-the-fly. Therefore, based on these models, PSG data can be viewed and exported in various ways for different objectives, such as visualizing and exporting overnight EEG-C3 channel signal or sleep stage 2 C3-A2 montage signal, which is much more flexible compared with original EDF files. In addition, we designed another model to store the characteristics of spindle, which provides high flexibility for visualizing different kinds of spindles, especially in spindle the reviewing function of SpindleSphere. For example, SpindleSphere can visualize the spindles detected from different algorithms at the same time by assigning different event concepts.

3.4 Interface

To support the requirements in spindle research, such as exporting and viewing signals corresponding to special events, we designed and prototyped SpindleSphere with a set of key features not available elsewhere: (1) standardized, metadata-based search and query of annotated PSG signal data; (2) interfaces for querying, importing, event-specific extraction, and repackaging of PSG data; (3) online visualization of single, multiple, or cross-sectional event-specific waveforms; (4) multi-scale spindle reviewing interface, which, for the first time, provides an intuitive and efficient solution for spindle reviewing, verification and comparison. The interface of SpindleSphere incorporates a rich set of query composition and visualization functionalities, including a) signal query and exporting interface, and b) signal visualization and spindle reviewing interface.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{system_architecture.png}
\caption{The system architecture of SpindleSphere.}
\end{figure}
Signal query and exporting interface. The SpindleSphere interface allows users to select a subject study, render the signal with interesting events, and apply montage function. It provides several widgets for user interaction, such as subject selection, montage selection, signal channel selection, spindle display mode and event display selection. These options are implemented as several simple drop down menus with the ability to select various values.

Signal visualization and spindle review interface. The SpindleSphere provides a range selector allowing the user to visually zoom into a particular section of the signal by dragging the scrollbar or the edges of the navigator window to the area of interest. In addition, users have the option of zooming/panning the time-range of interest using the mouse. For spindle reviewing, once the spindle display mode is selected, multiple colorful areas will be shown associated with related channels. Each area indicates a spindle that was detected by the specific approach, and colors represent various spindle detection approaches. The spindle area also can be zoomed in/out using the scrollbar.

3.5 Parallel spindle detection algorithm for computational scalability

In order to verify that SpindleSphere supports parallel computing (cloud computing) and provides more efficient workflow for spindle research, we designed and implemented a parallel version of automatic spindle detection algorithm proposed by Mölle et al.\textsuperscript{14}, named PASD (Parallel Algorithm for Spindle Detection). We used the output of SpindleSphere, i.e. the signal data exported by SpindleSphere, as the input file for PASD, and then compared the results between the original algorithm and the parallel version using our own spindle reviewing interface.

The workflow of the original algorithm proposed by Mölle et al.\textsuperscript{14} is shown in Figure 5(a). For each subject’s signal, it has six steps listed as follows: 1) apply band pass filter to each continuous sleep stage 2 signal. Each subject may have more than one continuous sleep stage 2 signal, therefore, a loop is implemented to process all sleep stage 2 signals; 2) aggregate the results of step 1 and calculate the standard deviation; 3) calculate the amplitude threshold using the result of step 2; 4) apply the same band pass filter to overnight signal from the subject; 5) threshold processing is employed on the filtered overnight signal and then the spindle candidates identified; 6) verify the time duration of each spindle candidate, and then output the final results.

However, two of these steps can be optimized using parallel computing. At first, it is unnecessary to apply bandpass filter to sleep stage 2 signals one by one, and all sleep stage 2 signals can be filtered concurrently due to the independence among each one. Second, step 4 does not need to be executed after the first 3 steps are finished, and it can be done with sleep stage 2 signals in step 1. Leveraging the advantage of parallel computing, multiple procedures can be processed concurrently, so we designed a parallel version PASD based on the MapReduce framework, which is illustrated in Figure 5(b). The input of PASD is the signals of all subjects, including sleep stages 2 signals and overnight signals. In the Map function, all the signals are filtered with bandpass filter, and split to each node in the cluster. In the Reduce function, the sleep stage 2 signals and overnight signal for each subject are grouped together according to the same value of key. Then, three steps mentioned above, including step 3 (T), step 5 (S) and step 6 (M), are executed. The final output consists of detected spindles of all subjects.

![Figure 5: The workflow of spindle detection algorithms. (a) The original algorithm. (b) The parallel algorithm (PASD); In the reduce phase, T is the Threshold step, S is the Segment&thresholding step, and M is the Maximum&minimum step in (a).](image-url)
4 Results

SpindleSphere was used to process PSG data of 300 subjects collected from consented subjects enrolled in the CCSHS study at NSRR. On average, an EDF file is in the size of 452 MB and 18.8% of the file (85 MB) is used in spindle detection. In the following, we first present the interface results of SpindleSphere, including query, exporting and visualization interface; we then describe the performance comparison, including storage space and execution time, of spindle detection using the traditional and new workflows.

4.1 Interface results

The features of the signal query and exporting interface are as follows: 1) Channel/montage selection: after selecting a subject, the user can select the appropriate channel or montage to visualize. A montage is the waveform representing the voltage difference between two different channels, and the deleted signals from the relevant channels were calculated on the fly in response to the user query. Users may also choose to create their own montages using the “Add montage” interface, which has the ability to manage the combination of signals for custom montages; 2) Event selection: the user can choose to view event-related signals to better navigate through the data. The events-related signals are stored in annotation_signals model, and this model enables visualization of the correct signal data segments with specialized events in response to the user query; 3) Spindle display mode: the user can choose to view the annotated spindle signals to review the detected spindles with multi-scale, such as spindles overview, spindles embedded in record and integrated spindles.

**Figure 6:** The query interface and the exporting interface of SpindleSphere.

**Signal query interface.** Figure 6(I) shows the components in query interface, including subject selection, channel type selection, channel/montage selection, montage management, annotation category selection, specific annotation selection, spindle display mode, export annotated data and select button. In subject selection, the user can select the subject’s EDF file they are interested in. In channel type selection, there are two types of signals that can be selected; “original channels” is the original channel in EDF file without any computation, and “montages” is the voltage difference between two different channels. Montage management provides the functions for adding and deleting montages. In annotation category selection, the user can select different types of events, such as wake, REM sleep and stage 2 sleep; then, the signal of the specific event can be selected in specific annotation selection. Opening the show spindle
mode, the detected spindles are labeled using color-areas in the record; otherwise, it only shows the pure signals of each selected channel or montage. Moreover, opening the only spindle mode shows only detected spindles without irrelevant data points. Clicking export annotated data opens a new page with the signal exporting interface. The query is executed by clicking the select button.

Signal exporting interface. Figure 6(II) shows the components in the exporting interface, including subject selection, channel type selection, channel/montage selection, montage management, annotation selection and download button. In contrast to the components in the query interface, subject selection and annotation selection enable the user to select multiple subjects and annotations and download the corresponding files.

The features of the signal visualization and spindle reviewing interface are summarized as follows: 1) functional scrollbar with preview of signals; 2) rendering of long records in a single page with scrollbar for ease of navigation; 3) channel/montage view property editor, such as color configuration; 4) zoom in/out function on both time-range and signal amplitude using functional scrollbar; 5) crosshair that locates each data point and shows its actual value; 6) multi-scale spindle visualization and reviewing using both square waveform and color-area, which are designed for users to review spindles in both integrity and detail.

The signal visualization and spindle reviewing interface. The components are shown in Figure 7, including annotation navigator, functional time-scale scrollbar, signal color configuration, amplitude-scale scrollbar, signal ruler, and spindle reviewing. The user can navigate to other annotation signals, which are listed in specific annotation selection in the query interface, using annotation navigator. Functional time-scale scrollbar provides multiple functions including preview of detected spindles, preview of both signals and spindles, and timespan selection, which enable the user to enlarge and show details of interesting parts of the signal; in addition, amplitude-scale scrollbar provides the function to zoom in/out of the amplitude scale. Moreover, the user can choose different colors for a specific signal using signal color configuration. Signal ruler is designed for verifying the value of each data point in an efficient way. Spindles are visualized with the colorful background in spindle reviewing. Different colors represent different kinds of spindles, such as the spindles detected by different algorithms as shown in F.1 to F.3 in Figure 7. Clicking the circle in the legend of F part can highlight the corresponding spindles. The integrated spindles without irrelevant data points in F part using only spindle mode (as shown in G part).

Figure 7: The signal visualization and spindle reviewing interface of SpindleSphere.
4.2 Spindle detection evaluation

A comparative evaluation was performed to measure the advantages of using SpindleSphere platform for processing, visualizing and reviewing spindle signal data. Two kinds of workflows were applied to the CCSSHS spindle study. For the traditional one, we first downloaded the EDF data from NSRR and stored it on our local machine, and then we detected spindles using the original method. For the new workflow with SpindleSphere, we downloaded the signal data from the exporting interface and applied both original and parallel algorithms. In addition, we imported our results into SpindleSphere, and then compared detected spindles using different methods; the result shows that the spindles detected by the original and parallel algorithm are the same.

Detection performance. In this paper, we mainly focus on the improvement of storage space and execution time, since the parallel algorithm only modified the process of the original algorithm without changing its precision and recall. The details of storage space on the local machine of two kinds of workflows are shown in Figure 8(a). The results indicate that the traditional workflow costs almost six times more than the new one does. Figure 8(b) shows the execution time of three different spindle detection procedures, including the original method with the traditional workflow, the original method with SpindleSphere, and PASD with SpindleSphere. The original method was executed on the local machine with Intel Core i7 2.93 GHz processor and 16 GB main memory, and PASD was executed on a 10-node cluster configured with Intel Xeon E5-2670 processor and 15 GB main memory per node. The results indicate that, for both the original method and PASD, SpindleSphere provides a competitive efficient workflow for spindle-related research. It is nonsensical to compare with applying PASD in the traditional workflow since it needs lots of hours to reformat EDF files to suitable input format. In contrast, the data downloaded from SpindleSphere can be directly used as the input for PASD.

Figure 8: Comparative performance evaluation of SpindleSphere versus the traditional workflow. (a) **Traditional**: the traditional workflow, **With SpindleSphere**: the new workflow with SpindleSphere. (b) **Ori. Traditional**: the original method in the traditional workflow, **Ori. SpindleSphere**: the original method in the new workflow, **PASD. SpindleSphere**: the parallel algorithm in the new workflow. (c) **a1**: algorithm 1, **a2**: algorithm 2, **a3**: algorithm 3.

Multi-method comparison. SpindleSphere offers an easy-to-use system to review and compare spindles across detection algorithms. Figure 8(c) shows the reviewing of spindles detected by three different approaches mentioned in
Simon Warby’s study\textsuperscript{24}. In this way, we can have an intuitive and convenient way to choose which approach has the best performance without knowing the labeled spindles. Figure 8(c) contains three subplots, and each of them shows the highlighted spindles with the corresponding approach. The results demonstrate that algorithm 1 \((a_1)\) and algorithm 3 \((a_3)\) detected similar spindles, and algorithm 2 \((a_2)\) detected the minimum number of spindles among three approaches.

5 Discussion

With the growing demand for multi-center collaborative studies in sleep research using large datasets, there is a need to adopt emerging integrated platforms to meet research requirements, such as less storage utilization, less redundant and highly scalable processing pipeline, intuitive and efficient reviewing system, multiple platform supporting and support for parallel computing. The results of the comparative performance evaluation demonstrate the significant advantages of using SpindleSphere in spindle research with very large PSG signal datasets. SpindleSphere is used in several spindle-related studies, including parallel spindle detection development and spindle analysis with sleep disorders. Some interesting discoveries were found using spindle reviewing interface. In addition, the data model and framework of SpindleSphere can be extended to other medical and clinical studies using large-scale bio-signal data, such as sleep stage analysis and epilepsy seizures detection.

Limitation. There are a few limitations of this paper, such as spindle editing, exporting and sharing. For future development, we will implement an interface for editing spindles, which will provide an easy-to-use spindle annotation tool for experts. The current version of SpindleSphere does not support downloading spindle data, but we are in the process of implementing spindle exporting function in the next phase of development. Moreover, statistical analysis of spindles and results sharing function will be implemented in the future.

6 Conclusion

In the sleep spindle studies, PSG data are often used as the gold standard. However, lack of a PSG data processing pipeline and web-based platform for managing and visualizing spindle-specific data on a large scale is one of the key challenges in leveraging spindles for clinical research. To address this challenge, we proposed a scalable web-based data management and spindle analysis and visualization platform, SpindleSphere. It provides a scalable spindle processing pipeline, and an intuitive web-based platform for clinicians and researchers with querying, exporting and visualizing capabilities of large-scale PSG signal data for spindle research. The results of performance evaluation of the SpindleSphere demonstrated a performance improvement in data storing, exporting, processing and visualizing with more than 130GB of data from 300 subjects from NSRR.

Acknowledgment. This work is supported in part by University of Kentucky Center for Clinical and Translational Science (Clinical and Translational Science Award UL1TR001998), National Sleep Research Resource (NHLBI R24 HL114473) and National Science Foundation under MRI Award No.1626364. We thank the Amazon Web Services (AWS) Cloud Credits for Research program that enabled the implementation of our parallel algorithm.

References


Representation of Social History Factors Across Age Groups:
A Topic Analysis of Free-Text Social Documentation

Elizabeth A. Lindemann, BS1, Elizabeth S. Chen, PhD2, Yan Wang, PhD3,
Steven J. Skube, MD1, Genevieve B. Melton, MD, PhD1,3
1Department of Surgery and 3Institute for Health Informatics University of Minnesota,
Minneapolis, MN; 2Center For Biomedical Informatics, Brown University, Providence, RI

Abstract
As individuals age, there is potential for dramatic changes in the social and behavioral determinants that affect health status and outcomes. The importance of these determinants has been increasingly recognized in clinical decision-making. We sought to characterize how social and behavioral health determinants vary in different demographic groups using a previously established schema of 28 social history types through both manual analysis and automated topic analysis of social documentation in the electronic health record across the population of an entire integrated healthcare system. Our manual analysis generated 8,335 annotations over 1,400 documents, representing 24 (86%) social history types. In contrast, automated topic analysis generated 22 (79%) social history types. A comparative evaluation demonstrated both similarities and differences in coverage between the manual and topic analyses. Our findings validate the widespread nature of social and behavioral determinants that affect health status over populations of individuals over their lifespan.

Introduction
Increasingly, the importance of social and behavioral factors on an individual’s health status are being recognized in clinical decision-making and effective population health management. As use of electronic health record (EHR) systems increases, there is a concomitant increase in EHR documentation that can be used to understand both how social history is documented and how these factors change over an individual’s lifespan. Previous studies1-9 have examined how individual social, behavioral, and environmental factors can be represented with controlled data representations as well as how these factors differ across documentation sources. For example, Rajamani et al.1, Aldekhyyel et al.2, and Lindemann et al.3 examined how occupation is documented in standards and reports, free-text comment fields in an EHR, and free-text documentation from multiple clinical note sources, respectively. Chen et al.4,5, Wang et al.6,7, Winden et al.8, and Carter et al.9 examined documentation of tobacco, alcohol, and drug use within free-text fields of the EHR. Finally, Winden et al.10 analyzed living situation, living conditions, and residence in standards. Since social history is often documented as free text and not structured data within clinical notes or in free-text documentation fields of the social history section of the EHR, this documentation and its associated topics, particularly with respect to an individual’s lifespan, has had limited characterization. Ultimately, gaining a better understanding of the social history information documented provides an opportunity to understand these factors across populations and across the lives of patients.

As a patient ages, there is some evidence that relevant social history topics change dramatically. For infants and toddlers, many of the social determinants of health (SDOH) of parents and caretakers are particularly important including their mental health status and social support11-14. Within the literature, many of the direct references to the SDOH of infants and toddlers (ages zero to 24 months)15 may refer to breastfeeding habits16,17, residence18, living conditions19, and palliative care for neonatal infants who have acute or potentially chronic care needs20-22. Residence is defined as the physical dwelling in which an individual lives or physical location, while the term living conditions refers to qualitative factors within these environments, such as secondhand smoke exposure, rodent infestations, water quality, and other relevant factors. As children enter early childhood, defined as ages two through five years by the National Institute of Child Health and Human Development (NICHD)15, topics such as living conditions and residence remain important23,24, and other pediatric-specific patient factors are of increasing importance like sleep habits25 and factors that may cause potential increase in obesity rates26. As children reach middle childhood (6-11 years)15 and enter school systems, exposures related to this new environment are increasingly discussed in literature, such as bullying27 and school performance. During this time, relational dynamics between both peers and family members and psychosocial effects are also increasingly important28, and factors associated with diet and obesity remain popular topics29.
When individuals reach adolescence (12-18 years)\textsuperscript{15}, SDOH topics broaden considerably and may more closely match social history topics for adult age groups. Documentation of alcohol use, drug use, and tobacco use now reflect individual use rather than exposure from parents or caretakers\textsuperscript{30}. Also, sexual history topics may be mentioned during adolescence without the context of abuse\textsuperscript{30}. In recent years, mentions of social media habits increase dramatically within literature as a factor of Internet presence\textsuperscript{31}. Specifically of interest is the impact of social media usage on psychological state and its association with bullying and social dynamics\textsuperscript{32}.

The importance of standardizing how SDOH topics are entered into the EHR for clinical decision-making has been recognized by the National Academy of Medicine\textsuperscript{33}. Towards this goal, understanding the current state of SDOH entry is important for determining what information is most valued by providers. In this study, we sought to identify how EHR social history topics change within clinical documentation as individuals age, and how social history topics vary by demographics, including gender, race, and ethnicity through automated topic analysis and manual analysis of social history topics in social history documentation for a large integrated healthcare system that serves both metropolitan and rural communities with primary and tertiary care.

**Methods**

At a high level, this study sought to provide a: (a) detailed analysis of social history topic variation by age and other demographics and (b) validation of the automated topic analysis through a separate manual analysis of social history documentation. An overview of our approach is provided in Figure 1 and was composed of four high level steps: (1) data collection, (2) topic analysis of pediatric and adult social history documentation, (3) manual validation of pediatric and adult social history documentation, and (4) comparison of social history topics from both analyses.

Figure 1. Overview of assessment of social history topics by age and other demographics. Populations and demographics of the patient set were defined first, while manual analysis and topic analysis were performed simultaneously.

**Data Collection**

Data sources for this study were the free-text Social History Documentation section within the enterprise EHR of Fairview Health Services (FHS). FHS is an integrated healthcare delivery system associated with the University of Minnesota that services both a large metropolitan area and rural parts of greater Minnesota. Social History Documentation from 2011 to 2016 was used for this analysis through the University of Minnesota research Clinical Data Repository (CDR). Each represented patient was used once in this five-year period to avoid oversampling from a single patient. Age groups were defined for pediatric patients according to the NICHD\textsuperscript{15}. The age groups for Neonatal Stage (0-27 days after birth), Infants (28 days after birth to 13 months of age), and Toddlers (13 months of age to 24 months of age)\textsuperscript{15} were combined, due to lower representation in FHS, compared with other age groups. Pediatric age groups for the analysis therefore consisted of: Infant and Toddler (birth to 24 months of age), Early Childhood (2-5 years), Middle Childhood (6-11 years), and Early Adolescence (12-18 years)\textsuperscript{15}. Adult age groups were defined according to the United States Census guidelines\textsuperscript{33} with age groups 19-24 years and 25-44 years were combined into a single Young Adulthood category. Adult age groups for the analysis therefore consisted of: Young Adulthood (19-44 years), Middle Adulthood (45-64 years), and Older Adulthood (65 years and older)\textsuperscript{34}. A representative number of patients within the FHS was used for each age group. Other demographics including race, gender, and ethnicity were also obtained on patients for analysis in the study.
Manual Analysis

To compare the most common social history topics by age, 200 patient entries were randomly selected for each age group from the years 2011 to 2016. An annotation schema based on previous studies with used consisting of twenty-eight major social history topics (Table 2). Annotators were instructed to annotate the text of the entire sentence for as many social history topics as the text contained. For example, the sentence “She lives with Mom and Dad in a single family home” would be annotated as Living Situation, Residence, and Family. An overlapping set of 10% of entries was annotated to perform inter-rater agreement by study investigators (EL, SS). Manual annotation for major social history topics in each age group resulted in a Cohen’s kappa of 0.67 and percentage agreement 0.92.

Automated Topic Analysis

Social history documents extracted from the CDR were pre-processed with an open source biomedical Natural Language Processing (NLP) pipeline to extract sections and split sections into statements. The preprocessed statements were then parsed by Stanford Probabilistic Context-Free Grammars (PCFGs) parser. We observed that multiple social history topics may be contained in a single statement. For example, the social history statement “Mom lives in an apartment and exercises twice a week” includes both information about the living condition and the hobby of the patient. For this reason, we collected all verb phrases of each social history statement separately as inputs for topic modeling. The subject of the verb phrase (e.g., “Mom” or “Siblings”), the verb (e.g., “exercise” or “live”), the verb prepositional conjunction word (e.g. “in”, “at”, or “with”) and sometimes the object of the verb was also collected. Verb frequencies of each group were computed, and those top occurring verbs were examined to decide if the object of the verb should be collected as part of the input for topic modeling. For instance, if a verb denotes no actions (e.g., “feel” or “report”), then the object of the verb (e.g., “safe” or “immunization”) was collected as part of the input for topic modeling. We also normalized language names, country names, relative names (e.g., “grandmother” or “nephew”), occupation name, and month names into single forms (e.g., “January” or “February”). Each word in the input was then normalized based on the SPECIALIST Lexicon.

To decide the optimum topic numbers for each age group, we computed harmonic means of topic models with a sequence of topic numbers from 5 to 70, step by 3. Figure 2 shows the harmonic means change with different topic numbers chosen in topic modeling. In the following step, a topic model was built for each age group with the computed optimum topic number for each group. The R package ‘topicmodels’ was used in this study to compute the optimum topic number and topic model fitting.

Comparison of Major Topics between Corpora

Following annotation, major social history topic occurrences were extracted. The highest frequency topics in the manually validated corpus were compared against topics that arose through topic analysis. A comparative evaluation was performed in order to validate similarities between corpora. Topics generated from topic analysis were grouped according to the annotation schema and coverage was evaluated.

Results

Manual Analysis

For each of the seven age groups, a total of 200 patients were selected at random (Table 1). This provided a total of 1,400 documents that were annotated for twenty-eight social history topics. Table 2 summarizes the representation of social history topics across age groups. A total of 8,335 annotations were made, with Young Adulthood containing the highest number of annotations (1,929) followed by Middle Childhood (1,170), and Early Adolescence (1,158). Sentences were annotated at the sentence level for all present social history topics, so in many cases, there
were multiple annotations made for each sentence. There were slight discrepancies between study investigators in annotations, surrounding a question referring to guns and occupations of guardians. These differences were resolved after the validation stage was completed.

Table 1 summarizes the demographics seen across the manual validation corpus. While a diverse population is represented, patients were selected at random, and therefore, may not be completely representative of populations across the greater FHS system.

**Table 1. Demographic information for patients included in manual validation corpus.**

<table>
<thead>
<tr>
<th>Topic</th>
<th>Infant and Toddler (0-24 months)</th>
<th>Early Childhood (2-5 years)</th>
<th>Middle Childhood (6-11 years)</th>
<th>Early Adolescence (12-18 years)</th>
<th>Young Adulthood (19-44 years)</th>
<th>Middle Adulthood (45-64 years)</th>
<th>Older Adulthood (65+ years)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Patients</td>
<td>200 (14.3%)</td>
<td>200 (14.3%)</td>
<td>200 (14.3%)</td>
<td>200 (14.3%)</td>
<td>200 (14.3%)</td>
<td>200 (14.3%)</td>
<td>200 (14.3%)</td>
</tr>
<tr>
<td>Number of Races Represented</td>
<td>10</td>
<td>10</td>
<td>9</td>
<td>10</td>
<td>10</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>African</td>
<td>14 (7.0%)</td>
<td>25 (12.5%)</td>
<td>24 (12.0%)</td>
<td>14 (7.0%)</td>
<td>9 (4.5%)</td>
<td>3 (1.5%)</td>
<td>7 (3.5%)</td>
</tr>
<tr>
<td>African American</td>
<td>9 (4.5%)</td>
<td>20 (10.0%)</td>
<td>19 (9.5%)</td>
<td>21 (10.5%)</td>
<td>13 (6.5%)</td>
<td>15 (7.5%)</td>
<td>2 (1.0%)</td>
</tr>
<tr>
<td>American Indian or Alaska Native</td>
<td>2 (1.0%)</td>
<td>1 (0.5%)</td>
<td>4 (2.0%)</td>
<td>3 (1.5%)</td>
<td>3 (1.5%)</td>
<td>3 (1.5%)</td>
<td>2 (1.0%)</td>
</tr>
<tr>
<td>Asian</td>
<td>17 (8.5%)</td>
<td>13 (6.5%)</td>
<td>5 (2.5%)</td>
<td>8 (4.0%)</td>
<td>9 (4.5%)</td>
<td>2 (1.0%)</td>
<td>4 (2.0%)</td>
</tr>
<tr>
<td>Hispanic or Latino</td>
<td>3 (1.5%)</td>
<td>2 (1.0%)</td>
<td>6 (1.0%)</td>
<td>1 (0.5%)</td>
<td>3 (1.5%)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Native Hawaiian or Other Pacific Islander</td>
<td>-</td>
<td>-</td>
<td>1 (0.5%)</td>
<td>-</td>
<td>1 (0.5%)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>White</td>
<td>138 (69.0%)</td>
<td>122 (61.0%)</td>
<td>126 (63.0%)</td>
<td>141 (70.5%)</td>
<td>157 (78.5%)</td>
<td>169 (84.5%)</td>
<td>185 (92.5%)</td>
</tr>
<tr>
<td>Some other race</td>
<td>2 (1.0%)</td>
<td>3 (1.5%)</td>
<td>-</td>
<td>1 (0.5%)</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Two or more races</td>
<td>3 (1.5%)</td>
<td>4 (2.0%)</td>
<td>6 (3.0%)</td>
<td>2 (1.0%)</td>
<td>1 (0.5%)</td>
<td>-</td>
<td>1 (0.5%)</td>
</tr>
<tr>
<td>Unknown</td>
<td>2 (1.0%)</td>
<td>2 (1.0%)</td>
<td>-</td>
<td>1 (0.5%)</td>
<td>1 (0.5%)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Choose not to answer</td>
<td>13 (6.5%)</td>
<td>12 (6.0%)</td>
<td>13 (6.5%)</td>
<td>10 (5.0%)</td>
<td>5 (2.5%)</td>
<td>8 (4.0%)</td>
<td>1 (0.5%)</td>
</tr>
<tr>
<td>Number of Ethnicities Represented</td>
<td>16</td>
<td>14</td>
<td>18</td>
<td>19</td>
<td>16</td>
<td>15</td>
<td>12</td>
</tr>
</tbody>
</table>

For all pediatric age groups, Family had the highest frequency of annotations, followed by Occupation and Living Situation. These topics are often seen together in the same text; for example, “She lives at home with Mom, Dad, and 2 siblings.” Occupation can refer to the occupation of an individual, “He works in IT,” or that of a parent or guardian, “Mom works from home, while Dad works in an office.” While annotations for these categories remained high for all age groups, there is a decrease in the frequency of these annotations, and a subsequent increase in the annotations of the Social History (SH) “Other” category. There is a large portion of the Older Adulthood annotations that account for Marital Status; for example, “He is a widower” or “Lives at home with Husband.” Although overall annotations for this Marital Status decrease by Older Adulthood, this type accounts for the third highest annotations for the age group.

**Table 2. Distribution of manual annotations across major social history topics (200 notes per group) 2011-2016.**

<table>
<thead>
<tr>
<th>Topic</th>
<th>Infant and Toddler (0-24 months)</th>
<th>Early Childhood (2-5 years)</th>
<th>Middle Childhood (6-11 years)</th>
<th>Early Adolescence (12-18 years)</th>
<th>Young Adulthood (19-44 years)</th>
<th>Middle Adulthood (45-64 years)</th>
<th>Older Adulthood (65+ years)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alcohol Use</td>
<td>-</td>
<td>1 (0.1%)</td>
<td>1 (0.1%)</td>
<td>5 (0.4%)</td>
<td>17 (0.9%)</td>
<td>18 (1.7%)</td>
<td>19 (2.4%)</td>
</tr>
<tr>
<td>Animals</td>
<td>22 (2.0%)</td>
<td>15 (1.3%)</td>
<td>27 (2.3%)</td>
<td>45 (3.9%)</td>
<td>60 (3.1%)</td>
<td>11 (1.0%)</td>
<td>12 (1.5%)</td>
</tr>
<tr>
<td>Caffeine Use</td>
<td>-</td>
<td>1 (0.1%)</td>
<td>1 (0.1%)</td>
<td>3 (0.3%)</td>
<td>118 (6.1%)</td>
<td>58 (5.5%)</td>
<td>31 (3.9%)</td>
</tr>
<tr>
<td>Diet</td>
<td>10 (0.9%)</td>
<td>8 (0.7%)</td>
<td>-</td>
<td>1 (0.1%)</td>
<td>178 (9.2%)</td>
<td>133 (12.5%)</td>
<td>81 (10.2%)</td>
</tr>
<tr>
<td>Drug Use</td>
<td>5 (0.5%)</td>
<td>1 (0.1%)</td>
<td>1 (0.1%)</td>
<td>4 (0.3%)</td>
<td>11 (0.6%)</td>
<td>6 (0.6%)</td>
<td>9 (1.1%)</td>
</tr>
<tr>
<td>Exposure Other</td>
<td>11 (1.0%)</td>
<td>11 (1.0%)</td>
<td>14 (1.2%)</td>
<td>37 (3.2%)</td>
<td>1 (0.1%)</td>
<td>10 (1.1%)</td>
<td>4 (0.5%)</td>
</tr>
<tr>
<td>Family</td>
<td>286 (26.0%)</td>
<td>319 (28.3%)</td>
<td>273 (23.3%)</td>
<td>217 (18.7%)</td>
<td>161 (8.3%)</td>
<td>10 (1.0%)</td>
<td>94 (11.9%)</td>
</tr>
<tr>
<td>Family History</td>
<td>20 (1.8%)</td>
<td>1 (0.1%)</td>
<td>1 (0.1%)</td>
<td>7 (0.6%)</td>
<td>7 (0.4%)</td>
<td>11 (1.0%)</td>
<td>18 (2.3%)</td>
</tr>
<tr>
<td>Hobby</td>
<td>-</td>
<td>3 (0.3%)</td>
<td>12 (1.0%)</td>
<td>29 (2.5%)</td>
<td>5 (0.3%)</td>
<td>1 (0.1%)</td>
<td>5 (0.6%)</td>
</tr>
<tr>
<td>Hobby Exposure</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Living Condition</td>
<td>7 (0.6%)</td>
<td>6 (0.5%)</td>
<td>16 (1.4%)</td>
<td>41 (3.5%)</td>
<td>2 (0.1%)</td>
<td>1 (0.1%)</td>
<td>-</td>
</tr>
<tr>
<td>Living Situation</td>
<td>155 (14.1%)</td>
<td>178 (15.8%)</td>
<td>181 (15.5%)</td>
<td>159 (13.7%)</td>
<td>143 (7.4%)</td>
<td>4 (4.3%)</td>
<td>46 (5.8%)</td>
</tr>
<tr>
<td>Living Situation Exposure</td>
<td>11 (1.0%)</td>
<td>6 (0.5%)</td>
<td>20 (1.7%)</td>
<td>38 (3.3%)</td>
<td>16 (0.8%)</td>
<td>3 (0.3%)</td>
<td>2 (0.3%)</td>
</tr>
<tr>
<td>Living Situation Other</td>
<td>43 (3.9%)</td>
<td>47 (4.2%)</td>
<td>56 (4.8%)</td>
<td>27 (2.3%)</td>
<td>3 (0.2%)</td>
<td>3 (0.3%)</td>
<td>1 (0.1%)</td>
</tr>
<tr>
<td>Marital Status</td>
<td>86 (7.8%)</td>
<td>103 (9.1%)</td>
<td>98 (8.4%)</td>
<td>56 (4.8%)</td>
<td>134 (6.9%)</td>
<td>106 (10.0%)</td>
<td>88 (11.1%)</td>
</tr>
<tr>
<td>Occupation</td>
<td>189 (17.2%)</td>
<td>184 (16.3%)</td>
<td>223 (19.1%)</td>
<td>201 (17.4%)</td>
<td>141 (7.3%)</td>
<td>103 (9.7%)</td>
<td>72 (9.1%)</td>
</tr>
</tbody>
</table>

1172
Table 2 Continued. Distribution of manual annotations across major social history topics for 200 notes per age group, from the years 2011-2016.

<table>
<thead>
<tr>
<th>Topic</th>
<th>Infant and Toddler (0-24 months)</th>
<th>Early Childhood (2-5 years)</th>
<th>Middle Childhood (6-11 years)</th>
<th>Early Adolescence (12-18 years)</th>
<th>Young Adulthood (19-44 years)</th>
<th>Middle Adulthood (45-64 years)</th>
<th>Older Adulthood (65+ years)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Occupation Exposure</td>
<td>2 (0.2%)</td>
<td>-</td>
<td>2 (0.2%)</td>
<td>1 (0.1%)</td>
<td>1 (0.1%)</td>
<td>1 (0.1%)</td>
<td>1 (0.1%)</td>
</tr>
<tr>
<td>Occupation Other</td>
<td>3 (0.3%)</td>
<td>-</td>
<td>-</td>
<td>1 (0.1%)</td>
<td>3 (0.3%)</td>
<td>2 (0.3%)</td>
<td></td>
</tr>
<tr>
<td>Physical Activity</td>
<td>3 (0.3%)</td>
<td>13 (1.2%)</td>
<td>16 (1.4%)</td>
<td>26 (2.2%)</td>
<td>141 (7.3%)</td>
<td>88 (8.3%)</td>
<td>60 (7.6%)</td>
</tr>
<tr>
<td>Physical Activity Exposure</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Physical Activity Other</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Residence</td>
<td>88 (8.0%)</td>
<td>79 (7.0%)</td>
<td>60 (5.1%)</td>
<td>72 (6.2%)</td>
<td>104 (5.4%)</td>
<td>38 (3.6%)</td>
<td>54 (6.8%)</td>
</tr>
<tr>
<td>Residence Exposure</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Social Support</td>
<td>15 (1.4%)</td>
<td>3 (0.3%)</td>
<td>3 (0.3%)</td>
<td>6 (0.5%)</td>
<td>11 (0.6%)</td>
<td>13 (1.2%)</td>
<td>11 (1.4%)</td>
</tr>
<tr>
<td>Social Support Other</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Tobacco Use</td>
<td>63 (5.7%)</td>
<td>21 (1.9%)</td>
<td>22 (1.9%)</td>
<td>24 (2.1%)</td>
<td>76 (3.9%)</td>
<td>21 (2.0%)</td>
<td>22 (2.8%)</td>
</tr>
<tr>
<td>Travel</td>
<td>1 (0.1%)</td>
<td>2 (0.2%)</td>
<td>2 (0.2%)</td>
<td>4 (0.3%)</td>
<td>5 (0.3%)</td>
<td>6 (0.6%)</td>
<td>6 (0.8%)</td>
</tr>
<tr>
<td>Social History Other</td>
<td>81 (7.4%)</td>
<td>117 (10.4%)</td>
<td>136 (11.6%)</td>
<td>148 (12.8%)</td>
<td>567 (29.4%)</td>
<td>267 (25.2%)</td>
<td>145 (18.3%)</td>
</tr>
<tr>
<td>Total Annotiations</td>
<td>1101</td>
<td>1126</td>
<td>1170</td>
<td>1158</td>
<td>1929</td>
<td>1060</td>
<td>791</td>
</tr>
</tbody>
</table>

**Topic Analysis**

A total of 187,920 patients were included in the automated topic analysis. Table 3 summarizes the demographics of the FHS population served from the years 2011 to 2016 with the Social History Documentation part of the analysis. This population is representative of the individuals serviced in the years specified. The age group for Young Adulthood has the most patients represented. Since adult age groups together contain more years compared to pediatric age groups, more patients are represented in adult age groups.

**Table 3.** Demographic information for patients across FHS Social History Documentation for the years 2011-2016.

<table>
<thead>
<tr>
<th></th>
<th>Infant and Toddler (0-24 months)</th>
<th>Early Childhood (2-5 years)</th>
<th>Middle Childhood (6-11 years)</th>
<th>Early Adolescence (12-18 years)</th>
<th>Young Adulthood (19-44 years)</th>
<th>Middle Adulthood (45-64 years)</th>
<th>Older Adulthood (65+ years)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Patients</td>
<td>5,383 (2.9%)</td>
<td>10,837 (5.8%)</td>
<td>13,103 (7.0%)</td>
<td>14,120 (7.5%)</td>
<td>74,530 (39.7%)</td>
<td>44,355 (23.6%)</td>
<td>25,592 (13.6%)</td>
</tr>
<tr>
<td>Gender – Male</td>
<td>2,781 (51.7%)</td>
<td>5,683 (52.4%)</td>
<td>6,673 (50.9%)</td>
<td>6,599 (46.7%)</td>
<td>18,934 (25.4%)</td>
<td>15,220 (34.3%)</td>
<td>8,979 (35.1%)</td>
</tr>
<tr>
<td>Gender – Female</td>
<td>2,602 (48.3%)</td>
<td>5,153 (47.6%)</td>
<td>6,429 (49.1%)</td>
<td>7,520 (53.5%)</td>
<td>53,504 (71.8%)</td>
<td>29,135 (65.7%)</td>
<td>16,612 (64.9%)</td>
</tr>
<tr>
<td>Gender – Unknown</td>
<td>-</td>
<td>-</td>
<td>1 (0.0%)</td>
<td>1 (0.0%)</td>
<td>1 (0.0%)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Gender – NA</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Number of Races Represented</td>
<td>11</td>
<td>11</td>
<td>11</td>
<td>11</td>
<td>11</td>
<td>11</td>
<td>11</td>
</tr>
<tr>
<td>African</td>
<td>293 (5.4%)</td>
<td>784 (7.2%)</td>
<td>1,121 (6.8%)</td>
<td>704 (5.0%)</td>
<td>3,015 (4.0%)</td>
<td>965 (2.2%)</td>
<td>359 (1.4%)</td>
</tr>
<tr>
<td>African American</td>
<td>394 (7.3%)</td>
<td>932 (8.6%)</td>
<td>1,158 (8.8%)</td>
<td>1,190 (8.4%)</td>
<td>3,846 (5.2%)</td>
<td>1,849 (4.2%)</td>
<td>537 (2.1%)</td>
</tr>
<tr>
<td>American Indian or Alaska Native</td>
<td>73 (1.4%)</td>
<td>152 (1.4%)</td>
<td>125 (1.3%)</td>
<td>184 (1.3%)</td>
<td>710 (1.0%)</td>
<td>395 (0.9%)</td>
<td>120 (0.5%)</td>
</tr>
<tr>
<td>Asian</td>
<td>385 (7.2%)</td>
<td>773 (7.1%)</td>
<td>803 (6.1%)</td>
<td>665 (4.7%)</td>
<td>4,439 (6.0%)</td>
<td>1,541 (3.5%)</td>
<td>682 (2.7%)</td>
</tr>
<tr>
<td>Hispanic or Latino</td>
<td>61 (1.1%)</td>
<td>199 (1.8%)</td>
<td>389 (3.0%)</td>
<td>340 (2.4%)</td>
<td>1,097 (1.5%)</td>
<td>441 (1.0%)</td>
<td>90 (0.4%)</td>
</tr>
<tr>
<td>Native Hawaiian or Other Pacific Islander</td>
<td>16 (0.3%)</td>
<td>31 (0.3%)</td>
<td>28 (0.2%)</td>
<td>34 (0.2%)</td>
<td>140 (2.0%)</td>
<td>54 (1.0%)</td>
<td>17 (0.1%)</td>
</tr>
<tr>
<td>White</td>
<td>2,442 (45.4%)</td>
<td>6,171 (56.9%)</td>
<td>7,892 (60.2%)</td>
<td>9,115 (64.6%)</td>
<td>46,302 (62.1%)</td>
<td>33,761 (76.1%)</td>
<td>21,287 (83.2%)</td>
</tr>
<tr>
<td>Some other race</td>
<td>30 (0.5%)</td>
<td>61 (0.6%)</td>
<td>131 (1.0%)</td>
<td>129 (0.9%)</td>
<td>464 (0.6%)</td>
<td>294 (0.7%)</td>
<td>150 (0.6%)</td>
</tr>
<tr>
<td>Two or more races</td>
<td>56 (1.0%)</td>
<td>218 (2.0%)</td>
<td>410 (3.1%)</td>
<td>224 (1.6%)</td>
<td>535 (0.7%)</td>
<td>122 (0.3%)</td>
<td>26 (0.1%)</td>
</tr>
<tr>
<td>Unknown</td>
<td>1,306 (24.3%)</td>
<td>937 (8.6%)</td>
<td>7,892 (3.7%)</td>
<td>690 (4.9%)</td>
<td>7,987 (10.7%)</td>
<td>2,906 (6.6%)</td>
<td>1,499 (5.9%)</td>
</tr>
<tr>
<td>Choose not to answer</td>
<td>527 (9.8%)</td>
<td>966 (8.9%)</td>
<td>929 (7.1%)</td>
<td>1,138 (8.1%)</td>
<td>4,985 (6.7%)</td>
<td>2,322 (5.2%)</td>
<td>941 (3.4%)</td>
</tr>
<tr>
<td>Number of Ethnicities Represented</td>
<td>54</td>
<td>59</td>
<td>60</td>
<td>63</td>
<td>66</td>
<td>65</td>
<td>63</td>
</tr>
</tbody>
</table>

Table 4 summarizes the ten most frequent topics for each age group identified using the topic analysis methods. These topics were grouped according to the schema of social history topics created for the manual analysis corpus. Many topics are slightly overlapping, causing there to be multiple entries for topics with similar content. For example, Marital Status is mentioned as the second, third, and fourth most occurring topic for Infant and Toddler, but may contain granularity that is not captured by the social history type classification.
Table 4. Representation of ten most frequent topic analysis topics, in order of frequency, with keywords and examples.

<table>
<thead>
<tr>
<th>Topic #</th>
<th>Assigned Name</th>
<th>Keywords</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>SH Other</td>
<td>Seatbelt, Language, Abuse, Helmets</td>
<td>Primary Language Spoken: English. Do you/your family use safety helmets? Abuse: Current or Past (Physical, Sexual, or Emotional) Seatbelts used.</td>
</tr>
<tr>
<td>2</td>
<td>Residence</td>
<td>Location Names, House, Apartment</td>
<td>NAME lives with her parents and sister in PLACE. Environmental History: The family lives in a 6 year old home in a rural setting. Lives with parents and half sister in the upstairs of a house while grandmother and a few cousins live on the first floor of a house.</td>
</tr>
<tr>
<td>3</td>
<td>Living Situation Exposure</td>
<td>Smoke exposure, Safe, Guns</td>
<td>NO: Lead, smokers at home, radon, pool/spa, known TB exposure. Do you feel safe in your home: Yes/No. No guns at home.</td>
</tr>
<tr>
<td>4</td>
<td>Family</td>
<td>Parents, Brother, Sister</td>
<td>Lives at home with mother, grandmother, aunt and 2 older half-siblings. Lives with biological mother and maternal half sister.</td>
</tr>
<tr>
<td>6</td>
<td>Occupation</td>
<td>Work, Part-time/Full-time Company Names</td>
<td>Patient did some part time work as a cashier for COMPANY in the past. Mom is a homemaker. Father is an engineer at COMPANY.</td>
</tr>
<tr>
<td>8</td>
<td>Diet</td>
<td>Calcium, Diet, Food</td>
<td>Calcium intake: eats cheese and drinks a lot of milk. Age solids introduced – 4 months, table food. Balanced diet: Yes</td>
</tr>
<tr>
<td>9</td>
<td>Alcohol Use</td>
<td>Alcohol, Socially Drinks</td>
<td>Alcohol use is &lt; 1 alcoholic drinks per week. Describes intermittent problems with alcohol in terms of excessive drinking in the past. Alcohol use is rare.</td>
</tr>
<tr>
<td>10</td>
<td>Tobacco Use</td>
<td>Cigarettes, Smoking, Exposure</td>
<td>They live in a house with no smoke exposure. Grandmother smokes outside. The patient has 20 yr hx of intermittent pipe and cigar use.</td>
</tr>
</tbody>
</table>

Comparative Evaluation

Table 5 summarizes the comparative evaluation performed to analyze similarities between SH topics that arose through manual analysis and topic analysis. Several topics were consistent through all age groups: Family, Living Situation, Living Situation Exposure, Marital Status, Occupation, Residence, and SH Other. Early Adolescence has the widest variety of social history topics included. Several topics become more prevalent with age, including Physical Activity, Diet, and Caffeine Use.

Table 5. Distribution of topics generated from manual and topic analyses, grouped by manual annotation schema.
Table 5 Continued. Distribution of topics generated from manual and topic analyses, grouped by manual annotation schema.

<table>
<thead>
<tr>
<th>Topic</th>
<th>Infant and Toddler (0-24 months)</th>
<th>Early Childhood (2-5 years)</th>
<th>Middle Childhood (6-11 years)</th>
<th>Early Adolescence (12-18 years)</th>
<th>Young Adulthood (19-44 years)</th>
<th>Middle Adulthood (45-64 years)</th>
<th>Older Adulthood (65+ years)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Living Situation</td>
<td><img src="image" alt="Presence" /></td>
<td><img src="image" alt="Presence" /></td>
<td><img src="image" alt="Presence" /></td>
<td><img src="image" alt="Presence" /></td>
<td><img src="image" alt="Presence" /></td>
<td><img src="image" alt="Presence" /></td>
<td><img src="image" alt="Presence" /></td>
</tr>
<tr>
<td>Living Situation Exposure</td>
<td><img src="image" alt="Presence" /></td>
<td><img src="image" alt="Presence" /></td>
<td><img src="image" alt="Presence" /></td>
<td><img src="image" alt="Presence" /></td>
<td><img src="image" alt="Presence" /></td>
<td><img src="image" alt="Presence" /></td>
<td><img src="image" alt="Presence" /></td>
</tr>
<tr>
<td>Living Situation Other</td>
<td><img src="image" alt="Presence" /></td>
<td><img src="image" alt="Presence" /></td>
<td><img src="image" alt="Presence" /></td>
<td><img src="image" alt="Presence" /></td>
<td><img src="image" alt="Presence" /></td>
<td><img src="image" alt="Presence" /></td>
<td><img src="image" alt="Presence" /></td>
</tr>
<tr>
<td>Marital Status</td>
<td><img src="image" alt="Presence" /></td>
<td><img src="image" alt="Presence" /></td>
<td><img src="image" alt="Presence" /></td>
<td><img src="image" alt="Presence" /></td>
<td><img src="image" alt="Presence" /></td>
<td><img src="image" alt="Presence" /></td>
<td><img src="image" alt="Presence" /></td>
</tr>
<tr>
<td>Occupation</td>
<td><img src="image" alt="Presence" /></td>
<td><img src="image" alt="Presence" /></td>
<td><img src="image" alt="Presence" /></td>
<td><img src="image" alt="Presence" /></td>
<td><img src="image" alt="Presence" /></td>
<td><img src="image" alt="Presence" /></td>
<td><img src="image" alt="Presence" /></td>
</tr>
<tr>
<td>Occupation Exposure</td>
<td><img src="image" alt="Presence" /></td>
<td><img src="image" alt="Presence" /></td>
<td><img src="image" alt="Presence" /></td>
<td><img src="image" alt="Presence" /></td>
<td><img src="image" alt="Presence" /></td>
<td><img src="image" alt="Presence" /></td>
<td><img src="image" alt="Presence" /></td>
</tr>
<tr>
<td>Physical Activity</td>
<td><img src="image" alt="Presence" /></td>
<td><img src="image" alt="Presence" /></td>
<td><img src="image" alt="Presence" /></td>
<td><img src="image" alt="Presence" /></td>
<td><img src="image" alt="Presence" /></td>
<td><img src="image" alt="Presence" /></td>
<td><img src="image" alt="Presence" /></td>
</tr>
<tr>
<td>Physical Activity Exposure</td>
<td><img src="image" alt="Presence" /></td>
<td><img src="image" alt="Presence" /></td>
<td><img src="image" alt="Presence" /></td>
<td><img src="image" alt="Presence" /></td>
<td><img src="image" alt="Presence" /></td>
<td><img src="image" alt="Presence" /></td>
<td><img src="image" alt="Presence" /></td>
</tr>
<tr>
<td>Physical Activity Other</td>
<td><img src="image" alt="Presence" /></td>
<td><img src="image" alt="Presence" /></td>
<td><img src="image" alt="Presence" /></td>
<td><img src="image" alt="Presence" /></td>
<td><img src="image" alt="Presence" /></td>
<td><img src="image" alt="Presence" /></td>
<td><img src="image" alt="Presence" /></td>
</tr>
<tr>
<td>Residence</td>
<td><img src="image" alt="Presence" /></td>
<td><img src="image" alt="Presence" /></td>
<td><img src="image" alt="Presence" /></td>
<td><img src="image" alt="Presence" /></td>
<td><img src="image" alt="Presence" /></td>
<td><img src="image" alt="Presence" /></td>
<td><img src="image" alt="Presence" /></td>
</tr>
<tr>
<td>Residence Exposure</td>
<td><img src="image" alt="Presence" /></td>
<td><img src="image" alt="Presence" /></td>
<td><img src="image" alt="Presence" /></td>
<td><img src="image" alt="Presence" /></td>
<td><img src="image" alt="Presence" /></td>
<td><img src="image" alt="Presence" /></td>
<td><img src="image" alt="Presence" /></td>
</tr>
<tr>
<td>Residence Other</td>
<td><img src="image" alt="Presence" /></td>
<td><img src="image" alt="Presence" /></td>
<td><img src="image" alt="Presence" /></td>
<td><img src="image" alt="Presence" /></td>
<td><img src="image" alt="Presence" /></td>
<td><img src="image" alt="Presence" /></td>
<td><img src="image" alt="Presence" /></td>
</tr>
<tr>
<td>Social Support</td>
<td><img src="image" alt="Presence" /></td>
<td><img src="image" alt="Presence" /></td>
<td><img src="image" alt="Presence" /></td>
<td><img src="image" alt="Presence" /></td>
<td><img src="image" alt="Presence" /></td>
<td><img src="image" alt="Presence" /></td>
<td><img src="image" alt="Presence" /></td>
</tr>
<tr>
<td>Tobacco Use</td>
<td><img src="image" alt="Presence" /></td>
<td><img src="image" alt="Presence" /></td>
<td><img src="image" alt="Presence" /></td>
<td><img src="image" alt="Presence" /></td>
<td><img src="image" alt="Presence" /></td>
<td><img src="image" alt="Presence" /></td>
<td><img src="image" alt="Presence" /></td>
</tr>
<tr>
<td>Travel</td>
<td><img src="image" alt="Presence" /></td>
<td><img src="image" alt="Presence" /></td>
<td><img src="image" alt="Presence" /></td>
<td><img src="image" alt="Presence" /></td>
<td><img src="image" alt="Presence" /></td>
<td><img src="image" alt="Presence" /></td>
<td><img src="image" alt="Presence" /></td>
</tr>
<tr>
<td>Social History Other</td>
<td><img src="image" alt="Presence" /></td>
<td><img src="image" alt="Presence" /></td>
<td><img src="image" alt="Presence" /></td>
<td><img src="image" alt="Presence" /></td>
<td><img src="image" alt="Presence" /></td>
<td><img src="image" alt="Presence" /></td>
<td><img src="image" alt="Presence" /></td>
</tr>
<tr>
<td>F of SH Topics Represented</td>
<td>20</td>
<td>17</td>
<td>22</td>
<td>14</td>
<td>22</td>
<td>16</td>
<td>24</td>
</tr>
</tbody>
</table>

* denotes presence of topic in manual analysis; ✓ denotes presence of topic in topic analysis.

Discussion

Variations Between Topic Analysis and Manual Analysis

The findings of this study demonstrate and validate the hypothesis that social history topics change over the course of an individual’s lifespan. The manual analysis conducted demonstrates the breadth of social history information that is contained within social documentation, while the topic analysis provided further granularity and depth to this information with a larger number of patients. Some of the richer pieces of information from the topic analysis includes references to languages individuals speak both primarily and in the home specifically, abuse factors, access to guns, seatbelt use, and helmet usage.

From twenty-eight potential social history types we have previously described, twenty-four were used at least once. There were no entries for Hobby Exposure, Hobby Other, Physical Activity Exposure, and Physical Activity Other. However, all other topics were observed in multiple age groups. There were minor variations in annotations between study investigators, contributing to the Cohen’s kappa of 0.67. These discrepancies were related to the occupations of parents and guardians and how to annotate the presence of firearms in the home. These discrepancies have also since been resolved at the close of the annotation process. In comparison, the topic analysis contained twenty-two of these topics. No topics were generated for the four types that were not present in the manual analysis. The number of social history topics represented in the topic analysis was fairly consistent across age groups, although Early Adolescence and Young Adulthood contained the best coverage for topics.

The comparative evaluation showed large areas of overlap between the manual analysis corpus and the topic analysis corpus, despite variations in size. This overlap demonstrates the complexity of information characterized within FHS Social History Documentation, and that this analysis is representative of the large clinical population in the Fairview Health Services system, since each patient with Social History Documentation was represented. The presence of SH Other topics in both the manual analysis corpus and the topic analysis corpus demonstrates how broad SDOH for individuals can be, and how difficult it might be to create structured documentation for SDOH that adequately encompasses an individual’s factors.

This work will serve as a basis for further natural language processing efforts, providing more robust tools for examining how social determinants affect individuals as they age. Providing a larger understanding of how SDOH are currently entered into the EHR will support understanding of what providers value and help to further standardization of entry, ultimately aiding in clinical decision-making.
Social Determinant Changes through Lifespan

With respect to social history documentation for pediatric individuals, there is a heavier focus on Living Situation observed prior to adulthood. This is also true for Occupation, possibly due to the focus on occupations of parents and guardians as it relates to a child’s care needs, although occupation continues to be well documented through adulthood and potentially in the structured Occupation fields in the EHR elsewhere for adults. Many social history types may refer to usage of parents, guardians, or exposure. For example, Alcohol Use, Drug Use, and Tobacco Use mentions largely refer to habits of individuals a child lives with in these corpora. Hobbies show the highest occurrence for pediatric individuals in the Middle Childhood and Early Adolescence age groups. Family maintains highest frequency, followed by Occupation, and Living Situation for all age groups throughout childhood.

Family remains prevalent as a subject of documentation through adulthood, but accounts for a smaller portion of annotations than in pediatric age groups. Family is the only social history type that is seen consistently in highest frequency groups. As individuals enter Older Adulthood, mentions of Marital Status account for a larger portion of annotations. The prevalence of Family mentions may be due to the relationship family history and with whom an individual associates with has on health status and outcome.

Following SH Other, the topics that arose with highest frequency from the topic analysis concerned Residence, Living Situation Exposure, Family, and Marital Status. These topics are largely related to one another and very difficult to separate by nature of their juxtaposition within. These findings demonstrate the interdependent nature of these factors within an individual’s SDOH. While Occupation topics were high in frequency for most age groups, the content of this topic was likely to change with age, and referred more directly to work experience rather than education as individuals enter adulthood.

Both corpora show a change in documentation as individuals reach adolescence and adulthood. Specifically, Caffeine, Diet, and Physical Activity show marked increases in documentation as individuals reach Young Adulthood. The largest change documented in these corpora that occurs as individuals enter adulthood is the increase in annotations for SH Other. The prevalence of SH Other topics is reflected through the topic analysis as summarized in Table 4. This category serves as a place to catch any social, behavioral, and environmental factors that are not related to the other types in Table 2. For all adult age groups, SH Other accounts for the most frequent annotations. This could possibly indicate that as we age, standard language around social history becomes harder to capture consistently and potentially more complex. Documenting the SDOH of parents and guardians might correlate to this, prove to be complex, and directly affect the SDOH of children. The complexity and variety of information included in this category will inform extensions to annotation schema for future work.

As patients within FHS reach adulthood, there is a shift in gender prevalence. This was evidenced in the demographics of the randomly selected manual analysis corpus as well as the topic analysis corpus that accessed all patients with Social History Documentation from 2011 to 2016. In pediatric sets, male patients represent a majority in both corpora. At Early Adolescence, genders are fairly comparable, leading up to a major shift at Young Adulthood. As patients enter adulthood, women represent more than 70% of patients in both corpora. This is likely due to a number of factors, but demonstrates that this analysis may be more accurate to the SDOH of adult women and pediatric males. The ratio between men and women becomes slightly more balanced as individuals enter Middle Adulthood and Older Adulthood. The demographics of this patient set also points to a lack of diversity in race. In all age groups, individuals who indicate their race as “White” predominate documentation. This is also subject to a number of factors, but could also point to this analysis more heavily representing the SDOH of those individuals.

Conclusion

Social history documentation in EHR systems will become increasingly valuable to understand for clinical care and other downstream consistent uses of this information. The content of this documentation over different age groups likely in part reflects changes of social history factors affecting individuals throughout their lifespan. The findings of this study point to the changing nature of SDOH as individuals age, and demonstrate the breadth and depth of SDOH that can affect a patient’s health status. Further work to standardize how social, behavioral and environmental factors are documented within the EHR is particularly needed to ensure robust documentation of diverse SDOH topics for different age groups.

Acknowledgements

Research reported in this manuscript was supported in part by the National Library of Medicine of the National Institutes of Health grant R01LM011364 and the University of Minnesota Clinical and Translational Science Institute.
Institute of the National Institute of Health grant 8UL1TR000114. The content and opinions expressed are solely the responsibility of the authors and may not represent the official views of the National Institute of Health.

References

Automatically Detecting Likely Edits in Clinical Notes Created Using
Automatic Speech Recognition

Kevin Lybarger, M.S., Mari Ostendorf, Ph.D., Meliha Yetisgen, Ph.D.
University of Washington, Seattle, WA, US

Abstract

The use of automatic speech recognition (ASR) to create clinical notes has the potential to reduce costs associated with note creation for electronic medical records, but at current system accuracy levels, post-editing by practitioners is needed to ensure note quality. Aiming to reduce the time required to edit ASR transcripts, this paper investigates novel methods for automatic detection of edit regions within the transcripts, including both putative ASR errors but also regions that are targets for cleanup or rephrasing. We create detection models using logistic regression and conditional random field models, exploring a variety of text-based features that consider the structure of clinical notes and exploit the medical context. Different medical text resources are used to improve feature extraction. Experimental results on a large corpus of practitioner-edited clinical notes show that 67% of sentence-level edits and 45% of word-level edits can be detected with a false detection rate of 15%.

Introduction

The use of ASR has increased within the clinical setting as speech recognition technology has matured and the availability of computational resources has increased. The creation of clinical notes using ASR offers system-level benefits, like short document turnaround time; however, note quality is negatively impacted by speech recognition errors, including clinically significant errors, and higher document creation times for practitioners associated with editing. Automatic detection and flagging of likely edits in ASR transcripts through a correction tool could reduce the time required to edit ASR transcripts and improve note quality by reducing the prevalence of uncaught errors. In this work, we investigated the automatic detection of sentences and words that are likely to be edited in the clinical note ASR transcripts by applying data-driven, machine learning detection strategies.

Practitioners may edit ASR transcripts to correct errors and disfluencies, but also to rephrase portions of the transcript. ASR errors are portions of the dictate that are incorrectly transcribed. Disfluencies include dictation that is repeated (e.g. “the patient the patient”), repaired (e.g. “hypertension I mean hypotension”), and restarted (e.g. “heart has abdomen is”). Rephrasing is associated with changes to the transcript that are not ASR errors or disfluencies (e.g. changing “patient got up” to “patient awoke” or changing “nothing by mouth” to “NPO”). Practitioners may also edit the transcripts as a continuation of the note creation process, deleting information that is no longer relevant or correct and inserting additional information such as test results and new plans for patient care. Word sequences and sentences in the ASR transcript that are edited by practitioners during editing are collectively referred to in this paper as transcript edits.

We identified ASR transcript edits within a corpus of clinical notes created through the voice-generated enhanced electronic note system (VGEENS) Project. We applied ASR error detection techniques to the detection of transcript edits within the VGEENS Corpus, utilizing medical domain knowledge, including clinical note structure and medical terminology. ASR error detection identifies discrepancies between what is dictated and what is transcribed, while our investigation of transcript edits focused on identifying differences between what is transcribed and what the practitioner wants. Table 1 contains an example transcript edit, with the ASR transcript text and the corresponding text from the final note. The ASR transcript includes an incorrect categorization of the patient’s cognitive status and a disfluency with the repair word “correction.” In the final note, the cognitive status is corrected and the disfluency is deleted. In this example, our edit detection model correctly identified all of the words in the ASR transcript that should be replaced or deleted (indicated by bold font).

<table>
<thead>
<tr>
<th>Source</th>
<th>Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>transcript</td>
<td>Alert and oriented 4 , pleasant mood , blunted affect correction for affect , thought process is clear</td>
</tr>
<tr>
<td>final note</td>
<td>Alert and oriented x4 , pleasant mood , full affect , thought process is clear</td>
</tr>
</tbody>
</table>

Within the VGEENS corpus, practitioners deleted words and entire sentences from the ASR transcripts. We hypothesized that sentence-level deletions and word level deletions within the ASR transcripts have different
characteristics and split the edit detection task into two tasks: *sentence-level edit detection* and *word-level edit detection*. The primary goals of this investigation were to evaluate the performance of ASR error detection techniques on the broader category of transcript edits at the sentence-level and word-level and explore methods for leveraging medical context and text resources to improve detection. We find that a substantial fraction of the errors can be detected with simple lexical context features but further gains are possible by leveraging medical context.

The rest of this paper is organized as follows. The Related Work section presents relevant ASR error detection work. The Methods section describes the data used in experimentation and the modeling approaches used to automatically detect transcript edits. The Results section presents the performance of the detection models, and the Conclusions section summarizes this investigation and discusses future work.

**Related Work**

There is a significant body of ASR error detection work (also known as ASR confidence estimation). ASR error detection has been approached using a range of discrete models, including the hidden Markov model (HMM), maximum entropy (MaxEnt) model, and conditional random fields (CRF) model, as well as continuous sequence models, including recurrent neural networks (RNN)\(^3\)-\(^11\).

Many studies have explored ASR error detection using the linear chain CRF model, which is a discriminative sequence modeling variant of the general CRF framework\(^12\). Bechet and Favre created a CRF error detection model using ASR posterior probabilities, lexical features (word n-grams and word length), and syntactic features (part of speech (POS) and dependency labels)\(^1\). In their work, the inclusion of both lexical and syntactic features improved error detection performance. Ghannay, Esteve, and Camelin explored the use of the Multi Layer Perceptron (MLP) neural network architecture, as well as a CRF model as a baseline\(^6\). The input features included ASR posterior probabilities, word representations (orthographic for CRF and embeddings for MLP), lexical features (word length and trigram indicator function), and syntactic features (POS tags and dependency labels). The best results were achieved using the MLP.

RNNs are currently a popular neural sequence model for ASR error detection, and several RNN variants have been used. Kalgaonkar, Liu, Yifan, and Yao compared the performance of a MLP model with a standard RNN and an RNN with an output decoder consisting of a two-state (no error/error) bigram language model\(^9\). The input features consisted of acoustic, linguistic, and confidence scores from the speech recognizer. The recurrent approaches outperformed the MLP, and the output decoder provided a small benefit. Ogawa and Hori created ASR error detection models based on CRF and bidirectional RNN frameworks, using acoustic and linguistic features and speech recognizer states and scores\(^10\). The bidirectional RNN outperformed the CRF. Ángel del-Agua, Piqueras, Giménez, Sanchis, and Civera explored speaker-adapted ASR confidence estimation using Naïve Bayes, CRF, and long short-term memory (LSTM) RNN models\(^11\). The models used pre-trained word vectors and speech recognizer-derived features, and the LSTM RNN achieved the best performance.

Much of the work on ASR error detection has focused on constrained domain human-computer interaction or human-directed broadcast news. The creation of clinical notes using ASR differs from other ASR transcription tasks in that the goal is note creation, not faithful transcription of what was dictated. The goal of note creation is to create a medically accurate document that articulates exam findings and plans for patient care and that meets formatting requirements/norms. Edit detection is motivated by this note creation goal and attempts to find a range of edit types (ASR errors, disfluencies, rephrasing, and note continuation) that impact the quality of the clinical notes. Edits within clinical ASR transcripts include sentence-level and word-level edits, where sentence-level edits tend to be a continuation of the note authoring process and word-level edits tend to be associated with disfluencies, speech recognition, and rephrasing. Another important difference with respect to ASR error detection work is that widely used medical dictation systems do not provide the detailed acoustic scores available in research systems, which makes error detection more challenging. On the other hand, in medical dictation, there are more context constraints that can be used to identify errors. Domain contextual constraints includes clinical note structure (e.g. topical sectioning), structured patient data in the Electronic Health Record (EHR), and ranges of numerical values (e.g. drug dosages and vital signs) and motivates the exploration of different features.

There is a relatively small body of work related to ASR error detection within the medical domain. Voll explored the automatic detection of ASR errors in radiology notes using language models, point-wise mutual information, and hand-crafted rules\(^13\)-\(^14\). Schreitter and Trost investigated the correction of medication dosages within ASR transcripts by extracting medications and the associated dosages and then evaluating dosages based on medication databases\(^15\).
Methods

Medical transcript edits comprise more than ASR errors and impact a larger portion of the note. Edit regions within ASR transcripts can be identified by aligning pairs of ASR transcripts and final notes. In clinical settings where practitioners dictate to a recording device without viewing ASR output in real time (noninteractive setting), ASR transcript-final note pairs are created as part of the existing workflow, providing efficient and low cost access to training data for detection models.

Data

The automatic detection of transcript edits was explored through the corpus of free-text clinical notes created through the VGEENS Project, which was conducted at the University of Washington Medical Center and Harborview Medical Center. As part of the VGEENS Project, inpatient progress notes were created by resident and attending internal medicine physicians using ASR through a multi-step process. First, a doctor dictated the note to a recording device during rounds, verbalizing punctuation and topical section headings (e.g. Chief Complaint). Then the dictation was transcribed using a commercial ASR system (Dragon Medical by Nuance Inc.) and automatically post-processed to format section headings. Lastly, the ASR transcript was reviewed and edited by the doctor, and the final note was entered into the EHR. The VGEENS Corpus of clinical notes includes 669 records created by 15 practitioners, where each record consists of an ASR transcript-final note pair.

Transcript edits were identified through the alignment of each ASR transcript-final note pair using Gestalt Pattern Matching. Gestalt Pattern Matching finds the longest sequence of matching tokens and then finds the next longest sequence of matching tokens to the left and right of the longest matching sequence. This process is applied recursively, until all of the matching sequences are identified. Based on the alignment, each word within the ASR transcripts was labeled as keep or delete. The capitalization of tokens was ignored during the alignment of the note pairs.

Sentence boundaries were determined based on the location of colons, periods, and line breaks, rather than using an off-the-shelf sentence boundary detector, because of the structure of the VGEENS notes (verbalized punctuation; section headings, numbered lists, etc. explicitly indicated). Approximately 10% of the sentences within the ASR transcripts were deleted during editing. In the subset of sentences that are not deleted, approximately 9% of the words were deleted during editing. Given the difference between the characteristics of sentence-level and word-level deletions, the detection of transcript edits was split into two tasks: sentence-level edit detection and word-level edit detection. Word-level gold standard labels were determined based on the keep or delete labels from the note alignments. Sentence-level gold standard labels were determined as follows: sentences were labeled as delete when all word-level labels were delete and sentences were labeled as keep when at least one word-level label in the sentence was keep.

The labeled ASR transcripts associated with the VGEENS Corpus were used in model training and testing (80% training/20% testing). Since this data set was relatively small, we explored use of the MedTrans17 and the i2b218 corpora of clinical notes (referred to as the External Corpora) for learning word classes and embeddings and a language model (LM). Table 2 contains a summary of the corpora used. In addition to these corpora, feature extraction utilized a list of medical terms derived from SNOMED CT, RxNorm, and the UMLS SPECIALIST Lexicon.

Table 2. Corpora summary

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Description</th>
<th>Note count</th>
<th>Word count</th>
<th>Sentence count</th>
</tr>
</thead>
<tbody>
<tr>
<td>VGEENS</td>
<td>Clinical notes created using ASR</td>
<td>669</td>
<td>ASR transcripts: 483 k final notes: 695 k</td>
<td>ASR transcripts: 46 k final notes: 57 k</td>
</tr>
<tr>
<td>MedTrans17</td>
<td>Example clinical notes created by human transcriptionists</td>
<td>2.37 k</td>
<td>1.51 M</td>
<td>135 k</td>
</tr>
<tr>
<td>i2b218</td>
<td>De-identified clinical notes, including only unique notes from 2006-2012 competition data sets.</td>
<td>4.32 k</td>
<td>4.75 M</td>
<td>461 k</td>
</tr>
</tbody>
</table>

Detection Models

The goal of the sentence-level edit detection task was to label sentences within the ASR transcripts as keep or delete. Logistic regression, which is a binary discriminative classifier, was selected for the sentence-level edit detection model because it is known to work well for text classification with relatively small amounts of training data when combined with regularization. The goal of the word-level edit detection task was to label words within the ASR transcripts as keep or delete. Word-level edit detection models were only trained and evaluated on sentences from the ASR
transcripts that had sentence-level *keep* labels from the note alignment. Because of the importance of sequential context at the word level, word-level edit detection was explored using the linear-chain CRF modeling framework, which estimates the highest probability sequence of labels given a sequence of observations12. We also experimented with an LSTM RNN in the word-level edit detection task; however, the LSTM RNN did not outperform the CRF, and only the CRF modeling results are presented in this paper.

**Feature extraction**

The VGEENS Corpus only includes the text output of the ASR system and does not include acoustic information or speech recognizer internal states (confidence scores, alternative word sequences, etc.), which are often used in ASR confidence estimation tasks. For both detection tasks, we utilized domain knowledge and unlabeled training data, in order to compensate for the limited information and relatively small size of the VGEENS Corpus. We explored interpretable features, like topical coherence, that may be useful to physicians.

A fixed vocabulary was selected based on the words that occur in the VGEENS Corpus training subset and External Corpora at least four times, resulting in a vocabulary size of 20.7k words. Out of vocabulary (OOV) tokens were mapped to one of seven OOV tokens, depending on whether the token was a medical term, numerical, lower case, upper case, title case, alphabetic, or other.

**Word-based Features**

In both detection tasks, text-based features were created using discrete and continuous word representations. In the sentence-level edit detection task, word-based features were intended to automatically learn relevant sentence attributes, like: numbered lists (e.g. “1. Liver cirrhosis…”), additional information required (e.g. “…waiting for LFT results…” or “Continue to monitor”), or topical headings not conforming to EHR format (e.g. “Cardiovascular:”). In the word-level edit detection task, word-based features were intended to automatically identify frequent ASR errors (e.g. phonetically similar words like “he” and “she”), disfluencies (e.g. repair words like “I mean” and “correction”), and rephrasing (e.g. abbreviating “daily” to “qd”).

Discrete word representations included orthographic and word class forms. Word classes were used to reduce data sparsity by grouping words based on syntactic/semantic similarity. Because of the small amount of VGEENS data, we leveraged external data to learn more reliable classes7,22. Two types of classes were used: manually-defined (e.g. “hypertension” → “<med_term>”)) and automatically learned (e.g. “patient” → “class00101”). Manual classes (rule-based classes) were created indicating punctuation, capitalization, numbers, and medical terms. Automatically learned word classes were created using unsupervised clustering approaches where words that appear in similar context are merged into the same class so as to maximize the mutual information between consecutive words, referred to as Brown clustering22. 500 Brown classes were learned from a merged corpus of the final notes in the VGEENS Corpus training set and the External Corpora.

Words were also represented as continuous word embeddings (vectors of real numbers), in which a sparse (one-hot) representation of words is mapped to a low-dimensional continuous space capturing syntactic, semantic, and topical information. Word embeddings were created using two unsupervised learning approaches. The first method starts with a term frequency-inverse document frequency (TF-IDF) representation of 4k VGEENS note sections and learns a linear transformation of words (and documents) into a 200-dimensional space using non-negative matrix factorization (NMF). (Performance was similar with 100 dimensions.) Word embeddings were also created using the neural word2vec skip-gram model, which is a single-layer neural network that predicts context words given the current word23. The skip-gram embeddings were created using the final notes in the VGEENS Corpus training set and the External Corpora (embedding size 200, context width 10). K-mean clustering was used to create 500 discrete skip-gram classes from the skip-gram embeddings.

**Language Modeling**

We hypothesized that atypical (infrequent) word sequences were more likely to reflect ASR errors or disfluencies (and require editing) than frequent sequences and therefore used word sequence probability as an input feature in both detection tasks. In order to create a LM that did not include the ASR transcript edit regions and did not overfit to the final notes in the VGEENS Corpus, a LM was trained on a subset of the External Corpora that best matched the VGEENS Corpus. The Moore-Lewis data selection approach was used to select the subset of the External Corpora24. An LM was created using the final notes in the VGEENS Corpus training set, and a second LM was created from a random sampling of the sentences within the External Corpora of similar size to the final notes in the VGEENS Corpus training set. The cross entropy of each sentence in the External Corpora was calculated using each LM, and the
difference between the cross entropy scores was used to select the best matching sentences in the External Corpora (approximately half of the corpora selected, 3.3 M words). A LM based on the selected subset of the External Corpora was used in subsequent experimentation. All LMs were trigrams models with Kneser-Ney smoothing.

Topical Coherence

The topical coherence between the target word or sentence and the local context was scored using word embeddings. In the sentence-level edit detection task, topical coherence features were intended to identify sentences that were out of place within the section, potentially belonging in a different section of the note. A sentence-level vector representation was created by averaging the embeddings of each word in the target sentence, and a section-level vector representation was created by averaging the embeddings of each of the remaining words in the note section. The cosine similarity and the vector difference between the representations was then calculated. In the word-level edit detection task, topical coherence features were intended to identify words that did not fit the surrounding context, due to an ASR error. The target word embedding was compared to the averaged vector representation of words in the local context (current sentence without this word and +/- two sentences) using cosine similarity and vector difference.

Task-Specific Feature Sets

Table 3 and Table 4 contain a list of the features for the sentence-level and word-level edit detection tasks, respectively.

**Table 3. Sentence-level edit detection features**

<table>
<thead>
<tr>
<th>Feature category</th>
<th>Feature set</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Structure</td>
<td>length/position</td>
<td>target sentence section, position in line (first, middle, or last), and token count</td>
</tr>
<tr>
<td>Words</td>
<td>words</td>
<td>word occurrence (indicator) vector for target sentence</td>
</tr>
<tr>
<td></td>
<td>Brown classes</td>
<td>Brown class occurrence vector for target sentence</td>
</tr>
<tr>
<td></td>
<td>skip-gram classes</td>
<td>skip-gram class occurrence vector for target sentence</td>
</tr>
<tr>
<td></td>
<td>adjacent words</td>
<td>word occurrence vector for the last three words in the previous sentence and the first three words in the next sentence</td>
</tr>
<tr>
<td>LM</td>
<td>perplexity</td>
<td>average per-word perplexity of the target sentence</td>
</tr>
<tr>
<td>Topical</td>
<td>VGEENS section</td>
<td>cosine similarity and vector difference of averaged TF-IDF/NMF embedding representations of the target sentence and note section</td>
</tr>
<tr>
<td></td>
<td>skip-gram section</td>
<td>cosine similarity and vector difference of averaged skip-gram embedding representations of the target sentence and note section</td>
</tr>
</tbody>
</table>

**Table 4. Word-level edit detection features**

<table>
<thead>
<tr>
<th>Feature category</th>
<th>Feature set</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Words</td>
<td>words</td>
<td>word n-grams (n=1-3) in window size of 5</td>
</tr>
<tr>
<td></td>
<td>manual classes</td>
<td>Manual class unigrams in window size of 5</td>
</tr>
<tr>
<td></td>
<td>Brown classes</td>
<td>Brown class n-grams (n=1-3) in window size 5</td>
</tr>
<tr>
<td></td>
<td>skip-gram classes</td>
<td>skip-gram class n-grams (n=1-3) in window size of 5</td>
</tr>
<tr>
<td>LM</td>
<td>probability</td>
<td>probability of word sequences in window sizes 3 and 5</td>
</tr>
<tr>
<td>Topical</td>
<td>VGEENS context</td>
<td>cosine similarity and vector difference of averaged TF-IDF/NMF embedding representations of the target word and local context</td>
</tr>
<tr>
<td></td>
<td>skip-gram context</td>
<td>cosine similarity and vector difference of averaged skip-gram embedding representations of the target word and local context</td>
</tr>
</tbody>
</table>

**Training and Evaluation**

The edit detection models (logistic regression and CRF) were trained using the VGEENS Corpus training set. Cross validation (three folds) was used to determine the best regularization type (L1-norm, L2-norm) and regularization weight for each feature set. For each feature set, the optimum L1-norm weight was determined with the L2-norm weight set to zero, and the optimum L2-norm weight was determined with the L1-norm weight set to zero. During cross validation, model performance was assessed using the Receiver Operating Characteristic (ROC) area under the curve (AUC). The final model for each feature set was trained on the entire training set using the selected regularization type and weight from cross validation.
The performance of the edit detection models was evaluated using the VGEENS Corpus testing set. Model performance was evaluated through ROC AUC and a performance analysis. In the edit detection tasks, the false detection rate ($P_f$) is the frequency of labeling a target as delete when the true label is keep, and the missed detection rate ($P_m$) is the frequency of labeling a target as keep when the true label is delete. In the performance analysis, $P_f$ was fixed at 15%, and $P_m$ was calculated. Conversely, $P_m$ was fixed at 15%, and $P_f$ was calculated. These $P_f$ and $P_m$ values were selected to understand model performance at different precision-recall operating points.

**Results**

*Sentence-level Edit Detection*

The sentence-level edit detection ROC AUC test results are presented in Table 5. The best performing single feature set was the words feature set, followed closely by the Brown classes. This suggests that the Brown classes capture the salient aspects of word context and syntax. The Brown classes, which are based on bigram occurrences, outperformed the skip-gram classes, which are based on word cooccurrence within a context window of length 10. The length/position feature set, which leverages the section structure of the clinical note, achieved high performance, despite a relatively small number of features (16 features). The TF-IDF/NMF topic modeling (VGEENS section feature set) achieved higher performance than the skip-gram topic modeling (skip-gram section feature set), even though the skip-gram embedding training set was approximately 13 times larger than the TF-IDF/NMF embedding training set. The TF-IDF/NMF approach utilized the structure of the note during topic learning, which may account for the higher performance. Two combined feature sets were tested: combined VGEENS using features based only on the VGEENS data and combined external, which added features based on external data. The combined feature sets outperformed the highest performing single feature, but the features based on external data had no added benefit. Figure 1 presents the error tradeoff curves for a subset of the feature sets evaluated.

**Table 5.** Sentence-level edit detection ROC AUC test results. Features that leverage external text resources are indicated with (*).

<table>
<thead>
<tr>
<th>Feature categories</th>
<th>Feature set</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Structure</td>
<td>length/position</td>
<td>0.75</td>
</tr>
<tr>
<td>Words</td>
<td>words</td>
<td>0.81</td>
</tr>
<tr>
<td></td>
<td>(*) Brown classes</td>
<td>0.80</td>
</tr>
<tr>
<td></td>
<td>(*) skip-gram classes</td>
<td>0.77</td>
</tr>
<tr>
<td></td>
<td>adjacent words</td>
<td>0.65</td>
</tr>
<tr>
<td>LM</td>
<td>(*) perplexity</td>
<td>0.65</td>
</tr>
<tr>
<td>Topic</td>
<td>VGEENS section</td>
<td>0.73</td>
</tr>
<tr>
<td></td>
<td>(*) skip-gram section</td>
<td>0.68</td>
</tr>
<tr>
<td>combined VGEENS</td>
<td>length/position + words + adjacent words + VGEENS section</td>
<td>0.83</td>
</tr>
<tr>
<td>combined external</td>
<td>(*) length/position + words + Brown classes + adjacent words + perplexity + VGEENS section</td>
<td>0.83</td>
</tr>
</tbody>
</table>

**Figure 1.** Sentence-level edit detection error tradeoff curves
Table 6 presents more detailed performance analysis for three systems at three different operating points of the sentence-level edit detector. First, we compare $P_m$ for the different systems at $P_f=15\%$. At this operating point, the $P_m$ of the combined VGEENS feature set was 67% lower than the status quo where all targets are labeled *keep* (no edit detection used). Next, we compare $P_f$ for different systems at $P_m=15\%$. At this operating point, $P_f$ of the combined VGEEN feature set is 57% lower than the status quo. Since many studies assess performance using F-score, we include the results for the operating point with the best F-score for each system. The best performing feature set for all metrics was the combined VGEENS feature set, and the word indicator feature set performed only slightly worse than the combined VGEENS feature set.

<table>
<thead>
<tr>
<th>Feature categories</th>
<th>Feature set</th>
<th>Fixed $P_f$</th>
<th>Fixed $P_m$</th>
<th>Optimized F1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$P_f$</td>
<td>$P_m$</td>
<td>$P_f$</td>
</tr>
<tr>
<td>Structure length/position</td>
<td>15% 46% 59% 15%</td>
<td>0.35</td>
<td>0.44</td>
<td>0.39</td>
</tr>
<tr>
<td>Words words</td>
<td>15% 39% 45% 15%</td>
<td>0.37</td>
<td>0.55</td>
<td>0.44</td>
</tr>
<tr>
<td>combined VGEENS length/position + words + adjacent words + VGEENS section</td>
<td>15% 33% 43% 15%</td>
<td>0.40</td>
<td>0.59</td>
<td>0.48</td>
</tr>
</tbody>
</table>

Table 7 presents section-level edit examples based on predictions from the detection model trained on the combined VGEENS feature set at $P_f=15\%$. The first example, which was correctly labeled as *delete*, illustrates a case where the sentence indicates that additional information is required. We hypothesized that during the time between dictation and editing, new findings or conclusions were available, resulting in the sentence being deleted. The second example, which was incorrectly labeled as *delete*, is similar in that it includes the word “continue;” however, it does not imply additional follow-up. The third example was correctly labeled as *delete* and has a similar format (short phrase followed by a colon) to the topical headings within the notes. This heading does not conform to the section headings defined by the EHR, and the practitioner appears to have deleted this section heading and the associated note content. In example 4, the sentence was incorrectly labeled as *keep*. Similar to the first example, the example 4 implies additional information is needed, but there were insufficient cues for the system to predict a *delete* label. The sentence may have been deleted because the required consultation was performed.

Table 7. Sentence-level edit detection examples

<table>
<thead>
<tr>
<th>No.</th>
<th>Label</th>
<th>Truth</th>
<th>Predicted</th>
<th>Example from ASR transcript</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>delete</td>
<td>Delete</td>
<td>Continue to monitor</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>keep</td>
<td>Delete</td>
<td>Continue rifaximin</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>delete</td>
<td>Delete</td>
<td>Ins and outs :</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>delete</td>
<td>Keep</td>
<td>Discuss with hepatology regarding further management</td>
<td></td>
</tr>
</tbody>
</table>

Word-level Edit Detection

The word-level edit detection test ROC AUC results are presented in Table 8. For the CRF approach, the single feature sets with the highest performance were Brown classes and words. Similar to the sentence-level edit detection task, the discrete features defined by Brown classes outperformed those based on the skip-gram classes. While both word class approaches reduced the size of the feature space significantly, the Brown classes appeared to better capture the salient syntactic aspects of the words. In contrast to the sentence-level edit detection task, the skip-gram topic modeling approach outperformed the VGEENS topic approach for characterizing the word-level context match. This suggests that the external text is useful for characterizing words but not sentences, as is not surprising because of the more controlled format of the VGEENS notes.

The combined VGEENS feature set was created using the words and VGEENS section feature sets. The combined external feature set was created using the best performing word class (Brown classes) and topic model (skip-gram context) feature sets and the remaining feature sets. The combined VGEENS feature set did not outperform the words feature set, but the combined external feature set outperformed the highest performing single feature set and the combined VGEENS feature set by approximately 3%. Figure 2 presents the error tradeoff curves for a subset of the feature sets evaluated.

![Image](image_url)
Table 8. Word-level edit detection test results. Features that leverage external text resources are indicated with (*).

<table>
<thead>
<tr>
<th>Feature categories</th>
<th>Feature set</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Words</td>
<td>words</td>
<td>0.72</td>
</tr>
<tr>
<td></td>
<td>(* Manual classes</td>
<td>0.62</td>
</tr>
<tr>
<td></td>
<td>(* Brown classes</td>
<td>0.72</td>
</tr>
<tr>
<td></td>
<td>(* skip-gram classes</td>
<td>0.68</td>
</tr>
<tr>
<td>LM</td>
<td>(* probability</td>
<td>0.64</td>
</tr>
<tr>
<td>Topic</td>
<td>VGEENS context</td>
<td>0.60</td>
</tr>
<tr>
<td></td>
<td>(* skip-gram context</td>
<td>0.65</td>
</tr>
<tr>
<td>combined VGEENS</td>
<td>words + VGEENS context</td>
<td>0.72</td>
</tr>
<tr>
<td>combined external</td>
<td>(* words + Manual classes + Brown classes + probability + skip-gram context</td>
<td>0.74</td>
</tr>
</tbody>
</table>

Figure 2. Word-level edit detection error tradeoff curves

Table 9 has more detailed performance analysis again at three different operating points for two word-level edit detectors. Comparing $P_m$ for the different systems at $P_f = 15\%$, the $P_m$ of the combined external feature set was 45\% lower than the status quo (assuming all targets are keep). With $P_m$ fixed at 15\%, the combined external feature set $P_f$ was 46\% lower than the status quo. Again, the best performing feature set for all metrics was the combined external feature set, and the words feature set performed only slightly worse than this combined feature set.

Table 9. Word-level edit detection performance analysis

<table>
<thead>
<tr>
<th>Feature categories</th>
<th>Feature set</th>
<th>Fixed $P_f$</th>
<th>Fixed $P_m$</th>
<th>Optimized F1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$P_t$</td>
<td>$P_m$</td>
<td>$P_t$</td>
</tr>
<tr>
<td>Words</td>
<td>words</td>
<td>15%</td>
<td>60%</td>
<td>58%</td>
</tr>
<tr>
<td>combined external</td>
<td>words + Manual classes + Brown classes + probability + skip-gram context</td>
<td>15%</td>
<td>55%</td>
<td>54%</td>
</tr>
</tbody>
</table>

Table 10 presents word-level edit detection examples based on predictions from the model trained on the combined external feature set with $P_f = 15\%$. Example 1 is from the Laboratory results section of the note and includes the word “restaurant,” which is a low probability word in this context and is not topically relevant. At $P_f = 15\%$, the model misses the deletion associated with the word “restaurant;” however the model correctly identifies this deletion at higher $P_f$. Example 2 includes the disfluency repair word “correction,” and the detection model correctly identified the deletion but incorrectly labeled “Temperature” as delete. Example 3 includes a common phrase that is abbreviated, which the model correctly labeled as delete; however, the model incorrectly labels additional words in the sentence as delete. Example 4 appears to include an ASR error (transcription of “technetium” instead of “magnesium”), which the model correctly labels as delete. Similar to examples 2 and 3, a false delete label is also applied within the sentence.
Table 10. Word-level edit detection examples (true delete are **bold strikethrough**, false delete are **bold underline**, false keep are *italics strikethrough*, and true keep are unformatted)

<table>
<thead>
<tr>
<th>No.</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ASR: …heart rate 79 <strong>restaurant rate</strong> 16 blood pressure 121 / 76… Final: …heart rate 79 , blood pressure 121 / 76…</td>
</tr>
<tr>
<td>2</td>
<td>ASR: <strong>Temperature 31 correction</strong> 37.1 , heart rate 90… Final: Temperature 37.1 , heart rate 90…</td>
</tr>
<tr>
<td>3</td>
<td>ASR: Macrocytic anemia , <em>present on admission</em> , <strong>chronic and related</strong>… Final: Macrocytic anemia , POA , chronic and related…</td>
</tr>
<tr>
<td>4</td>
<td>ASR: …Glucose 90 , <strong>calcium</strong> 8.4 , <strong>technetium</strong> 2.1… Final: …Glucose 90 , calcium 8.4 , magnesium 2.1…</td>
</tr>
</tbody>
</table>

Conclusions

In this paper, we applied ASR error detection techniques to the automatic detection of sentence-level and word-level edits within clinical ASR transcripts. The results demonstrate that a substantial number of sentence- and word-level edits can be automatically detected with a small false detection rate. In both tasks, the word and word class feature sets were the highest performing single feature sets, indicating that word classes learned from external medical text resources using unsupervised Brown clustering are effective prediction features. Although the language model and topic-based features achieved lower performance than the word-based features, the results show that these features are relevant for edit detection in that the best performance in both detection tasks was achieved through a combination of features. The high performance achieved in the sentence-level tasks suggest a strong relationship between sentence editing habits and topical coherence within note sections. The best performance in the word-level task was achieved through the incorporation of external data.

This work is limited by the size of the corpus of clinical notes created using ASR and the number of practitioners involved in the creation of this corpus. A larger corpus, created by a larger sample of practitioners, would likely improve detection performance and improve the generalizability of the detection models to notes created with different dictation protocols. The methods used were also constrained by the ASR system configuration; access to alternative recognizer hypotheses or word confidence estimates would lead to further improvements in performance.

This work was motivated by the hypothesis that a correction tool that automatically detects and flags likely edits within ASR transcripts could improve note quality and accuracy. While the performance achieved in this investigation is likely not adequate to create a viable correction tool at this point, this work produced promising results that warrant further exploration, including the procurement of additional training data. Future work to improve performance would likely include incorporating additional, unlabeled data through semi-supervised learning and the inclusion of biomedical knowledge sources. A user study with practitioners is required to assess the required level of performance, determine the appropriate performance metrics and thresholds (precision, recall, etc.), and design the correction tool interface.

Acknowledgements

We thank Eve Riskin, Thomas Payne, and Andrew White for their assistance. The data used was provided through the support of grant number R21HS023631 from the Agency for Healthcare Research and Quality (AHRQ). The content is solely the responsibility of the authors and does not necessarily represent the official views of the AHRQ.

References

17. Transcribed Medical Transcription Sample Reports and Examples - MTSamples [Internet]. [cited 2015 Apr 30]. Available from: http://mtsamples.com/
18. i2b2: Informatics for Integrating Biology & the Bedside [Internet]. [cited 2015 Sep 1]. Available from: https://www.i2b2.org/
Improving Quality of Follow-Up Imaging Recommendations in Radiology

Thusitha Mabotuwana\textsuperscript{1,2}, PhD; Christopher S Hall\textsuperscript{1,2}, PhD; Joel Tieder\textsuperscript{2}, MD, MPH; Martin L. Gunn\textsuperscript{2}, MBChB, FRANZCR

\textsuperscript{1}Philips Healthcare, Seattle, WA, USA, \textsuperscript{2}University of Washington, Seattle, WA, USA

Abstract

Failure of timely follow-up imaging recommendations can result in suboptimal patient care. Evidence suggests that the use of conditional language in follow-up recommendations is associated with changes to follow-up compliance. Assuming that referring physicians prefer explicit guidance for follow-up recommendations, we develop algorithms to extract recommended modality and interval from follow-up imaging recommendations related to lung, thyroid and adrenal findings. Using a production dataset of 417,451 radiology reports, we observed that on average, follow-up interval was not mentioned in 79.4% of reports, and modality was missing in 47.4% of reports (4,819 reports contained a follow-up imaging recommendation for one of the three findings). We also developed an interactive dashboard to be used to monitor compliance rates. Recognizing the importance of increasing precision of follow-up recommendations, a quality improvement pilot study is underway with the goal of achieving a target where follow-up modality and interval are both explicitly specified.

Introduction

Radiology reports often contain follow-up imaging recommendations to monitor stability of potentially malignant findings, to ensure resolution of potentially serious disease, or for further diagnostic characterization [1]. However, failure to comply with imaging follow-up recommendations in a timely manner is common and can lead to delayed treatment, poor patient outcomes, unnecessary testing, lost revenue, and legal liability [1-3].

Follow-up recommendation detection in radiology reports has been an active area of research recently, although much of the focus has been on identifying recommendations associated with specific incidental findings [4, 5]. Incidental findings are those that were unexpected by the ordering provider and incidental to the primary reason for the current exam; for example, a small pulmonary nodule in the lower lobe of the lung may be detected on a CT abdomen and pelvis study that was ordered to evaluate right lower quadrant pain. Other studies have focused on identifying follow-up recommendations for a specific modality, such as CT [2], critical findings [6] or a particular type of finding, such as pulmonary nodules [7] or adrenal masses [8]. In order for follow-up detection algorithms to be more useful in routine practice, there is an opportunity to make algorithms scalable and generic so that recommendations can be identified from all radiology reports irrespective of modality or type of finding.

Imaging follow-up adherence rates have been reported to be low, with over 35% of follow-up imaging recommendations not followed-up [9]. In one study, 12% of cases of potential malignancy were not followed up appropriately [3]. Often, clinicians may determine that follow-up is unnecessary, especially when a follow-up recommendation is made non-applicable by clinical findings that were not available to the radiologist at the time of recommendation [1]. Various other reasons have been attributed to failure to follow-up, including the referring physicians missing the recommendations or losing track while addressing a more acute illness, loss of information during handover between care-teams, the recommendation not being communicated to the patient, and the patient failing to schedule or show-up for the follow-up appointment [10].

Despite various factors that may affect follow-up imaging adherence, one area where radiologists have room for improvement is in the quality and clarity of follow-up imaging recommendations. Referring physicians, especially primary care physicians who may not be as familiar with the latest imaging guidelines, value more explicit follow-up imaging recommendations by radiologists [11]. In fact, in a recent study, imaging follow-up rate was found to drop from 78.8% for no conditional language to 43.8% when conditional language was present [12]. Based on the assumptions that referring physicians prefer more explicit recommendations for follow-up imaging and specific recommendations will in turn improve follow-up compliance rates, in this paper we present a radiology report-processing pipeline that can be used to assess the quality of follow-up imaging recommendations. Further, to be clinically useful as a quality improvement tool, it is often important to identify the anatomy associated with a
follow-up recommendation since some clinical findings are more important to follow up than others. As such, we present a generic methodology to extract the anatomy with a focus on lung, thyroid and adrenal nodules. These three sets of findings have well established guidelines that include mentioning of specific time intervals and imaging modalities. We also present a dashboard that has been developed as part of a quality improvement initiative that can be used to routinely track follow-up recommendation rates by radiology academic section and/or anatomy as well as the quality of the recommendations.

Methods

Dataset
We extracted 417,451 radiology reports generated between 1-January-2015 and 31-May-2016 from the University of Washington radiology information system for three network hospitals. For each report, several meta-data fields were also extracted, including exam date, radiology subspecialty, patient class and modality. The Human Subjects Division at the University of Washington determined that the study was IRB exempt as part of a quality improvement project.

Report processing pipeline: follow-up detection (previous work)
The first step in the process was to identify reports that contained a follow-up imaging recommendation. This was performed using a previously developed follow-up detection algorithm which parses the radiology report to extract sections (e.g., “Clinical Indication”, “Findings” and “Impression” as shown in Figure 1), paragraph headers within each section if any (e.g., “Abdomen” and “Pelvis”) and the sentences within the paragraphs. The algorithm then evaluates the sentences within the “Findings” and “Impression” sections to determine if a sentence contains a follow-up recommendation (e.g., “Given history of malignancy, follow-up CT chest in 3 months is recommended”). Follow-up detection is performed using keyword searches and other heuristics. The output of this first step is a list of follow-up recommendation sentences as shown underlined in Figure 1 (along with meta data, such as whether it is a negated sentence – e.g., “no further follow-up is necessary”). Using 532 reports annotated for follow-up imaging recommendations by a radiologist (senior clinical author MG) as the ground truth, the detection algorithm was evaluated to have 93.2% PPV (95% CI: 89.8-94.5%), 99.5 NPV (95% CI: 98.4-99.9%) and 97.9% accuracy (95% CI: 96.2-98.5%).
Figure 1: Sample radiology report with multiple follow-up imaging recommendations. Underlining is added for emphasis and is not present in the original report.

Report processing pipeline: quality of recommendations

For the purposes of this study, a quality improvement oversight committee composed of multiple clinical and quality stakeholders decided that explicitly mentioning the suggested follow-up duration and modality of the recommended follow-up exam is an important indication of the quality of a follow-up imaging recommendation. For example, we hypothesize that “follow-up with a CT in 3-6 months to assess stability” will be preferred by more referring physicians compared to “follow-up to assess stability”. Due to the nature of specific health conditions, explicit mentioning of time interval and modality is not always possible, and as such, the initial goal of the pilot project was to achieve a reasonably high rate (e.g., 70%) agreed upon by relevant stakeholders.

Once a follow-up recommendation sentence was detected, the next step in the processing pipeline was to determine the modality and time interval associated with the follow-up recommendation sentence. Given the finite number of modalities and the numerical nature of the duration, we used regular expressions to extract this information. Sometimes the interval is explicit (e.g., follow-up in three months) and we accounted for this case as well. We observed that most intervals are specified in months, although in a few cases, “days” was used as well as “annually” (usually when referring to routine screening/monitoring related follow-ups). Therefore, intervals were calculated in
days, and a minimum and a maximum value were extracted (e.g., values 3 and 6 will be the minimum and maximum respectively from sentence “follow-up in 3-6 months”). Minimum and maximum were set to be the same if only one value was specified (e.g., “follow-up in 3 months”).

**Report processing pipeline: anatomy extraction**

Next, to identify the anatomy associated with the follow-up recommendation, we used an ontology based natural language processing engine previously developed [13] along with the publicly available NCBO annotation service [14]. Queries to both services were constrained to extract anatomies as defined by the SNOMED-CT ontology. Results were then merged and unique values selected. This approach was selected to optimize the capabilities of the two systems, for instance, if the text contains “right lower lobe”, the anatomy engine would detect “Structure of right lower lobe of lung” corresponding to SNOMED ID 266005 whereas NCBO would not find a mapping. Conversely, from the sentence “hypervascular liver lesion, MRI follow-up is suggested”, NCBO detected “Liver Structure”, corresponding to SNOMED ID 10200004 whereas the internal engine identified “Lesion of liver” (SNOMED ID 300331000), which is a finding. Since our focus is on identifying anatomy, in this instance, the engine did not find any relevant anatomy since a longer phrase was already matched.

Our follow-up anatomy detection algorithm was developed such that it first attempts to extract anatomy from the follow-up sentence – for instance, concept “Thoracic Structure” corresponding to ID 51185008 will be extracted from “Follow-up CT chest is recommended”. If no anatomy is identified in the follow-up sentence, the algorithm steps backwards from the follow-up sentence, processing one sentence at a time, until at least one anatomy is identified in a sentence. Search was restricted to the section in which the follow-up sentence occurred (which is usually ‘Findings’ and/or ‘Impression’ sections). Once identified, the ‘anatomy context’ becomes the text from the beginning of matched sentence to end of follow-up sentence. This process was repeated for all follow-up sentences when a report contained multiple recommendations. Table 1 shows four examples of extracted anatomy. For each follow-up recommendation, we also keep track of the previous two sentences which is referred to as ‘search context’. This search context can then be queried using regular expressions to detect the type of follow-up (e.g., whether follow-up recommendation is for a pulmonary nodule).

<table>
<thead>
<tr>
<th>Anatomy Context</th>
<th>Extracted Anatomy</th>
<th>SNOMED-CT Description(s) and ID(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>These can be reassessed on CT chest for lung nodule follow-up</td>
<td>chest lung</td>
<td>Thoracic Structure, 51185008 Entire lung, 181216001</td>
</tr>
<tr>
<td>There is a right adrenal nodule which is likely benign and could be further evaluated by CT at the time of lung nodule follow-up</td>
<td>right adrenal lung</td>
<td>Entire right adrenal gland, 281625001 Entire lung, 181216001</td>
</tr>
<tr>
<td>1 cm hypoechoic focal lesion in the mid portion of the left kidney. Although it is possible that it may represent a simple cyst, it is not adequately characterized on this study. Recommend follow up US in 6 months to establish stability.</td>
<td>left kidney</td>
<td>Left kidney structure, 18639004</td>
</tr>
<tr>
<td>Nodular opacities in the right lung may represent infection versus aspiration. Dedicated CT may be helpful.</td>
<td>right lung</td>
<td>Right lung structure, 3341006</td>
</tr>
</tbody>
</table>

**Clinical use case and data visualization**

The stakeholders from the oversight committee decided to focus first on three commonly occurring findings for which published follow-up guidelines exist: lung, thyroid and adrenal nodules. Consequently, the scope of current research identified follow-up recommendations for these three findings. Keywords ‘nodule’, ‘lesion’, ‘tumor’, ‘lump’ and ‘mass’ were included for all three, while several additional descriptors were included at a finding-
specific level. We included ‘opacity’ for lung findings; ‘hypodensity’ and ‘fullness’ for adrenal findings; and ‘hypodensity’ and ‘opacity’ for thyroid findings. We required one of these nodule-related words to be within a 6-word proximity (after removing stop words) of where the anatomy was detected within the anatomy context to ensure the finding was actually related to the detected anatomy. When multiple anatomies are extracted, these results are consolidated in a post-processing step so that follow-up recommendations can be tracked at an exam level. For instance, if there are two follow-up recommendations in a report, one for lung, and one for thyroid, the consolidated anatomy will become “lung and thyroid”.

To accommodate routine monitoring of follow-up rates along with quality compliance rates (in terms of specifying duration and modality), we developed a dashboard that is updated on a monthly basis. An automated report is produced from the radiology information system that contains data for the previous month and the report processing pipeline is executed automatically. The dashboard was developed (using Microsoft Power BI, Redmond, WA) to share monthly quality metrics with specific radiology administrators, including Section Chiefs.

Algorithm validation

To validate our algorithm’s ability to correctly determine the anatomy associated with a follow-up imaging recommendation, we manually selected a total of 200 reports – 50 reports for each of the three follow-up finding types as well as 50 reports that contained a follow-up recommendation but were unrelated to the lung, adrenal or thyroid. This was performed by searching for the specific finding types in the “Findings” and “Impression” sections of randomly selected reports and repeating the process until the required dataset of 200 reports was created. The algorithm performance was 98.7% sensitivity (95% CI: 96.5-98.7%), 100% specificity (95% CI: 93.6-100%) and 99% accuracy (95% CI: 95.8-99%). There were two false-negatives, one related to an adrenal nodule and the other related to a lung nodule. A false-negative was defined as an instance where follow-up detection or anatomy extraction failed. Overall accuracy of 99% was slightly better than 97.9% follow-up detection accuracy reported previously since detection errors were rectified prior to anatomy extraction, which is the focus of the work presented herein.

Results

There were 27,375 (6.6%) reports that had at least one follow-up imaging recommendation sentence, a rate comparable to that observed by other researchers [15]. Of these, 4,819 exams contained at least one of the three specific finding types of interest. Table 2 shows the distribution of the follow-up recommendations by finding type and Table 3 shows the distribution by scanned modality (this is the modality of the performed exam for which the report contained a follow-up imaging recommendation). There were 3,909 CT Chest exams for all three anatomies that contained a follow-up recommendation of which 2,905 (74.3%) were lung related (lung: 2,775; lung and thyroid: 87; lung and adrenal: 43).

Table 2: Exams by type of follow-up recommendation for all modalities

<table>
<thead>
<tr>
<th>Type of Follow-Up Recommendation</th>
<th>Number of Exams (n = 4819)</th>
<th>Percent of Exams with Follow-up Recommendation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lung</td>
<td>3467</td>
<td>71.94%</td>
</tr>
<tr>
<td>Thyroid</td>
<td>890</td>
<td>18.47%</td>
</tr>
<tr>
<td>Adrenal</td>
<td>325</td>
<td>6.75%</td>
</tr>
<tr>
<td>Lung and Thyroid</td>
<td>89</td>
<td>1.85%</td>
</tr>
<tr>
<td>Lung and Adrenal</td>
<td>43</td>
<td>0.89%</td>
</tr>
<tr>
<td>Thyroid and Adrenal</td>
<td>5</td>
<td>0.10%</td>
</tr>
</tbody>
</table>
Table 3: Exams with follow-up findings by scanned modality (‘Others’ include RF, DX and interventional radiology)

<table>
<thead>
<tr>
<th>Modality Associated with Follow-Up</th>
<th>Number of Exams (n = 4819)</th>
<th>Percent of Exams with Follow-up Recommendation</th>
</tr>
</thead>
<tbody>
<tr>
<td>CT</td>
<td>3909</td>
<td>81.12%</td>
</tr>
<tr>
<td>CR</td>
<td>532</td>
<td>11.04%</td>
</tr>
<tr>
<td>MR</td>
<td>131</td>
<td>2.72%</td>
</tr>
<tr>
<td>PT</td>
<td>97</td>
<td>2.01%</td>
</tr>
<tr>
<td>US</td>
<td>72</td>
<td>1.49%</td>
</tr>
<tr>
<td>NM</td>
<td>71</td>
<td>1.47%</td>
</tr>
<tr>
<td>Others</td>
<td>7</td>
<td>0.15%</td>
</tr>
</tbody>
</table>

There were 3,828 (79.4%) lung, thyroid or adrenal related follow-up exams that contained at least one follow-up recommendation sentence where the minimum and/or maximum duration was not specified, and 2,282 (47.4%) reports with at least one follow-up recommendation sentence where the follow-up duration was not specified. 1,973 (40.9%) of the exams contained follow-up imaging recommendations that did not have the follow-up interval or the modality specified. Distribution of these by modality is shown in Table 4.

Table 4: Exams with follow-up interval and/or modality not specified by scanned modality

<table>
<thead>
<tr>
<th>Modality of the original exam</th>
<th>#Exams (n = 4819)</th>
<th>#Exams with Interval Not Specified (n = 3828)</th>
<th>#Exams with Modality Not Specified (n = 2282)</th>
<th>#Exams with Interval and Modality Not Specified (n = 1973)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CT</td>
<td>3909</td>
<td>2974</td>
<td>2013</td>
<td>1716</td>
</tr>
<tr>
<td>CR</td>
<td>532</td>
<td>506</td>
<td>100</td>
<td>98</td>
</tr>
<tr>
<td>MR</td>
<td>131</td>
<td>130</td>
<td>55</td>
<td>54</td>
</tr>
<tr>
<td>PT</td>
<td>97</td>
<td>93</td>
<td>61</td>
<td>60</td>
</tr>
<tr>
<td>US</td>
<td>72</td>
<td>57</td>
<td>31</td>
<td>26</td>
</tr>
<tr>
<td>NM</td>
<td>71</td>
<td>62</td>
<td>22</td>
<td>19</td>
</tr>
<tr>
<td>Others</td>
<td>7</td>
<td>6</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

In order to provide radiology administrators with a quick overview of departmental trends and the ability to monitor the effectiveness of any quality improvement interventions over time, we also developed an interactive dashboard that shows the number and percentage of reports containing follow-up sentences by anatomy, modality as well as by month. Various filters have been provided so that a user can explore the data to understand various trends and opportunities for improvement. Figure 2 shows the number of reports containing follow-up recommendation sentences that are lung, thyroid or adrenal related for all sections across all three hospitals for the 18-month duration. A user can easily explore the underlying data that contributes towards a particular metric, for instance, a user can right-click on the CT bar showing ‘3.9k’ and examine the various report attributes (commonly referred to as a “drill down” capability).
Figure 2: Dashboard showing exams with follow-up recommendations

Note that a given report may be included simultaneously under the ‘Y’ and ‘N’ categories in the ‘Min/Max Duration Specified’ and ‘Modality Specified’ charts (shown in the center of the dashboard). This is because a report can contain multiple recommendation sentences. For instance, a report may contain the sentences “Follow-up recommended to ensure stability” with respect to a suspicious thyroid nodule and “Follow-up CT recommended in 3 months” for an indeterminate pulmonary nodule. This report will be included under ‘Lung and Thyroid’ row in the ‘#F/up Exams by Anatomy’ chart, and contribute to the values under ‘Y’ and ‘N’ in both ‘Min/Max Duration Specified’ and ‘Modality Specified’ charts. As a result, the sum of the ‘Y’ and ‘N’ categories in these two charts will be greater than the number of reports (4819 in this case). The bottom right chart shows the percentage and number of exams (in parenthesis) where duration and/or modality has not been specified as a function of the month of the exam, for instance, 82% is 196/240 for January 2015, and 56% is 135/240 (value 240 can be seen in the top right chart which shows the number of exams with follow-up recommendations by month).

In Figure 2, the ‘Anatomy filter’ (in top left) shows the consolidated list of anatomies for a given report. Since a report can contain multiple recommendations, if the multiple recommendations are not consolidated at a report level (e.g., to “lung and thyroid”), the default dashboard behaviour is to count the report independently under each anatomy (e.g., the same report will be included once under ‘lung’ and again under ‘thyroid’) resulting in an inaccurate total exam count; therefore, after anatomy associated with all follow-up recommendations were extracted for a report, an additional step was performed to consolidate the anatomies associated with the extracted recommendations.

Discussion

In this manuscript, we outlined a generic report processing pipeline that can be used to determine the consistent use of language within follow-up recommendations. Using production data, we have also demonstrated how the pipeline can be used to extract follow-up recommendation sentences associated with lung, thyroid and adrenal nodules for multiple imaging modalities. A key strength of this work is the integration of multiple components to provide an end-to-end solution, starting with a raw data extract from the radiology information system all the way through to automatically updating a dashboard that can be used to support quality improvement initiatives. The technology presented can be used in several ways, including: (1) radiology administrators can use the dashboard to determine
how the number of follow-up recommendations is trending over time as well as the quality of these follow-up recommendations by section; (2) the technology has the potential to be used as a surrogate to identify incidental findings by filtering for exams where the anatomy of follow-up recommendation is different from anatomy of ordered exam; and (3) follow-up detection can be used as a basis to determine follow-up recommendation adherence rates and design appropriate interventions if adherence rates are low.

Despite using a production dataset from three institutions, the current study has several limitations. First, we have only anecdotal evidence that referring physicians value the inclusion of follow up interval and modality when follow-up imaging is recommended. Although follow-up language has been cited as one of the factors that influences a referring physician’s decision to follow-up on imaging recommendations [11, 12] and we believe that more specific recommendations can aid in referring physician to have a more informed conversation with the patient, we did not validate this in the current study. Second, all reports were created using common dictation macros that are shared across the network hospitals and therefore the methods used to parse radiology reports may not be readily generalizable to other institutions. However, follow-up recommendation sentences within these reports did not uniformly use macros. Third, the algorithm performed imperfectly in 2 out of the 200 reports we examined to extract the anatomy. In one of these instances, the algorithm missed the follow-up statement which was mentioned in conjunction with another (“The attenuation coefficient of the left adrenal nodule is about 10 Hounsfield units. Therefore, it cannot be characterized as an adenoma. This could be characterized by CT at the same time as a renal mass protocol”). Although the pipeline failed to recognize this recommendation, it did identify the follow-up recommendation for the renal mass in the previous sentence, likely ensuring that follow-up would occur. In the other failed instance, “Multiple gray nodules are unchanged in size compared to prior, but remain indeterminate. Recommend follow-up CT in 12 months to assess for stability.”, the follow-up recommendation was correctly detected, but the anatomy was not (“gray nodules’ does not match any anatomy concepts in SNOMED). “Gray nodules” was almost certainly a voice recognition mistranscription of “pulmonary nodules.” This also shows some of the limitations of using an ontology-based approach to detect anatomy. Complementing the ontology-based approach with domain-specific dictionaries (e.g., using a text-to-anatomy dictionary where “gray nodule” is a key used to refer to anatomy “lung”) could be one option. We are also looking into generalizing the detection of nodule-related concepts. For instance, filtering for ‘morphologic abnormality’ concepts in SNOMED could be an option instead of specifying variants for ‘nodule’.

There are several potential applications for this algorithm in our department. First, we are planning to use the quality of follow-up imaging recommendations as one of the measures for performance based incentives for radiology seconds as part of a quality improvement pilot project. The goal is to achieve a target where follow up interval and modality are both explicitly specified in an attempt to improve the precision of follow-up recommendations. Incomplete follow-up recommendation sentences can be confusing for ordering clinicians who are unfamiliar with published imaging follow-up guidelines. We are in the process of implementing standardized follow-up macros and ensuring guideline uniformity among radiology sections in our department, and plan to measure their impact using this pipeline. Second, we can use this algorithm to benchmark our own follow-up compliance rate and variability with published clinical and departmental follow-up guidelines. Third, we can use this algorithm to ensure timely follow-up imaging of the appropriate body region using the appropriate duration and modality where applicable. Certain interventions can also be implemented within the department, for instance, by adopting techniques that have shown success in previous studies, such as asking radiologists to dictate certain phrases into the reports and alerting when follow-up is due [16], or by assigning explicit scores to indicate the degree of how suspicious a lesion is for possible malignancy and the need for follow-up [17].

We have demonstrated how a robust pipeline can be developed to quantify the quality of follow-up imaging recommendations, with proof points for lung, thyroid and adrenal related nodules. To improve appropriate follow-up language, it is important to be able to identify and to assess consistent phraseology of follow-up imaging recommendations. Given poor adherence rates with follow-up imaging recommendations, new techniques that are easily extensible without excessive human rework are needed to allow radiology administrators to easily identify opportunities where improvements can be made. With the gradual transition to value-based healthcare, improving precision of follow-up recommendations could be one of the ways radiology can provide more value to the referring physicians and contribute more towards the overall management of the patient.
References

Evidence of Progress in Making Nursing Practice Visible Using Standardized Nursing Data: a Systematic Review

Tamara G. R. Macieira, BSN, PhD student¹, Madison B. Smith, BSN, RN, PhD student¹
Nicolle Davis, BSN, RN, SCRN, PhD student¹, Yingwei Yao, PhD¹, Diana J. Wilkie, PhD, RN, FAAN¹
Karen Dunn Lopez, PhD, MPH², Gail Keenan, PhD, RN, FAAN¹
¹University of Florida, Florida; ²University of Illinois at Chicago, Illinois

Abstract

Nursing care documentation in electronic health records (EHRs) with standardized nursing terminologies (SNTs) can facilitate nursing’s participation in big data science that involves combining and analyzing multiple sources of data. Before merging SNTs data with other sources, it is important to understand how such data are being used and analyzed to support nursing practice. The main purpose of this systematic review was to identify studies using SNTs data, their aims and analytical methods. A two-phase systematic process resulted in inclusion and review of 35 publications. Aims of the studies ranged from describing most popular nursing diagnoses, outcomes, and interventions on a unit to predicting outcomes using multi-site data. Analytical techniques varied as well and included descriptive statistics, correlations, data mining, and predictive modeling. The review underscored the value of developing a deep understanding of the meaning and potential impact of nursing variables before merging with other sources of data.

Introduction

The main frontline providers of care are nurses who also represent the largest category of health workers in the hospital setting. Among the 2.8 million registered nurses currently working in the United States (U.S.), 61% work in hospitals whereas 19% of 297,100 pharmacists and 41.9% of 854,698 physicians in practice work in hospitals. Nurses are responsible 24 hours each day for continuously identifying care issues, implementing and adjusting care prescribed by themselves and other providers to achieve desired patient outcomes. To date, however, it has been difficult to effectively evaluate the impact of nursing on patient outcomes. The growing use of electronic health records (EHRs) to document care now offers the opportunity to use the data captured in practice for discovering knowledge to transform health care. Thus, the documentation entered by nurses into EHRs, for the first time ever, is a potential source for discovering the impact of nursing care on patient outcomes and using the knowledge to improve care. In this article, we report our systematic review of studies that utilized nursing EHRs data to answer a variety of research questions from describing nursing care for a specific population to predicting patient outcomes. The publications reviewed provide a foundation for identifying future paths of inquiry involving nursing and other data retrievable from EHRs.

The use of standardized nursing terminologies (SNTs) to document nursing care enables the easy retrieval and analysis of nursing data while also representing the nurse’s clinical reasoning. The integration of nursing data into large datasets requires the frequent and rapid input of new valid information from EHRs. These can be achieved through the use of controlled vocabularies in EHRs, which helps overcome the major challenges of aggregation, processing and analysis associated with unstructured text data. In nursing, SNTs are controlled vocabularies that represent nursing care as nursing diagnoses, interventions and outcomes. The SNT coded data retrieved from EHRs can be analyzed alone or merged with other EHRs data. The use of SNTs to document nursing practice is a big step toward supporting the aggregation of nursing data to large datasets and big data science.

Different sets of SNTs are used to document nursing care. The American Nurses Association (ANA) recognizes and supports the use of certain nursing terminologies to guide and document care if those have clear and unambiguous concepts, are coded with a unique identifier per concept, and if those terminologies were tested for reliability, clinical usefulness and validity. The following nursing terminologies are recognized by ANA: NANDA-International (NANDA-I); Nursing Interventions Classification (NIC); Nursing Outcomes Classification (NOC); International Classification for Nursing Practice (ICNP); Omaha System; Clinical Care Classification (CCC); and the Perioperative Nursing Data Set (PNDS). While ICNP, Omaha System, CCC and PNDS sets contain diagnoses, interventions and outcomes terms; NANDA-I (diagnoses), NIC (interventions) and NOC (outcomes) are three separate terminologies. Since NANDA-I, NIC and NOC are very often used together, we will refer to them as a terminology set (NNN).

Systematically reporting and analyzing studies that used SNTs nursing data retrieved from EHRs is important to understand the analytic issues related to the complexity and richness of data generated from the use of these
terminologies. Given the growing emphasis on using existing health care datasets, there is a need for new statistical, computational, and visualization methods to analyze EHRs data, given their complexity and volume. For nursing to join this effort, an important first step is to identify and examine the studies that analyze EHR data coded with SNTs.

To date, there are few reviews that report secondary analysis of SNTs nursing data. One recently published systematic review described study focus, sample characteristics and frequency of publications that studied SNTs. The authors also identified a limited set of studies in which SNTs nursing data were being analyzed. However, substantial evaluation and discussion on the analysis of SNTs nursing data were not performed. Another review described characteristics of nursing research data (ranging from patient demographics, social history, medical history, medications, among others) and evaluated if a specific index metadata system represents sufficiently nursing data. Our systematic review differentiates from the previous investigations as we focused only on studies that analyzed nursing data coded with SNTs retrieved from EHRs. The earlier literature review included all types of unstructured nursing data, and different types of studies, such as controlled clinical trials. The authors also did not restrict their selection criteria for nursing data retrieved from EHRs.

The systematic review presented below was conducted to describe and critically analyze the body of studies in which secondary analyses of data coded with the ANA recognized SNTs was performed. These nursing data were documented during the delivery of nursing care and retrieved from EHRs. We believe that the findings from the present systematic review will emphasize the importance of coded nursing data and will encourage a wider use of SNTs that will allow greater participation of nursing in big data science initiatives.

Objective

The objective of this systematic review was to uncover the state of the science related to the use of standardized nursing data (coded with SNTs) retrieved from EHRs to answer research questions, describe the analytical techniques employed, and outline the lessons learned applicable big data science and nursing.

Methods

Search strategy

A comprehensive literature search was conducted to identify publications in which secondary analysis was performed on data extracted from EHRs and documented using terms from the ANA recognized nursing terminologies (or sets). There were two phases used in this process: 1) study selection and 2) data collection process.

Building from the work of Tastan et al. that reported secondary uses of SNTs documentation in studies up to 2011, the databases PubMed and CINAHL were searched using keywords encompassing all ANA recognized SNTs, along with “nursing” and “electronic health records” from 2010 to 2017. Keywords were defined for each database with the help of a librarian, who was a specialist in Consumer Health, Nursing and Health Education and Behavior. The limiters since 2010, abstract available and published in English were used. Potential publications identified in both databases were downloaded into a reference management program (EndNote X7, Thompson Reuters ISI ResearchSoft), in which duplicates were deleted and abstracts were reviewed. Finally, grey literature search was conducted using Google Scholar to identify possible publications not captured by the traditional search methods, including relevant publications not in PubMed or CINAHL, but in computer science databases like IEEE. Name of authors of publications already reviewed and included in this study’s sample were individually searched in Google Scholar. Publications pertinent to the subject were reviewed and included.

Study selection

The studies identified using the specified search strategy were submitted to abstract review according to the following inclusion criteria: 1) publications that conducted secondary analysis of nursing data retrieved from EHRs, 2) nursing data were coded with one of the ANA recognized nursing terminologies and documented at the point-of-care, as part of the institution’s routine, and 3) publications were published in English. Following PRISMA guidelines, Figure 1 shows the number of publications first identified, retained after selection from abstracts, and reviewed as full text.

The team utilized a two-stage selection process. In the first stage, two doctoral students in nursing informatics (TM, ND) independently conducted the literature search in the databases using the same keywords and found the same number of studies. A guide created by the authors was used to systematically review the abstracts. The reviewers first identified if the article contained at least one of the ANA recognized SNTs. If the abstract described the use of a SNT, the reviewer would continue to verify if the data analyzed were retrieved from EHRs. In the case of an affirmative answer, the reviewers would further analyze the abstract to assess if the SNT coded nursing data were documented as
part of routine clinical practice and not solely for research purposes. Publications selected by one reviewer but not the other were discussed until agreement was reached. In the second stage, TM and MS (doctoral students) conducted a full-text review for further eligibility screening.

Data collection process

Data from the eligible publications were independently collected by TM and MS. Tables were created to summarize and analyze the content of the articles. The following information was extracted: study foci; sample characteristics (number of health records); variables (dependent, independent, descriptive); and statistical analysis. Data extracted from the selected articles were also reviewed by two faculty members, one specialist in nursing informatics and one statistician. Consensus agreement was reached. The present study aimed to report the diverse statistical analysis methods utilized and studies foci addressed through the use of SNTs coded nursing data. We did not intend to report the results of those publications.

Quality appraisal

Quality appraisal of publications is common in systematic reviews, however, it is challenging when the publications being reviewed are of the secondary data analysis nature. For the purpose of this review, quality appraisal of articles was evaluated regarding completeness of reporting of information. An adapted version of the STROBE Statement-checklist for observational studies was used, including the following items from the original checklist: study size, clear definition of variables, and description of all statistical methods used. Clear definition of variables was scored as 1 (not defined), 2 (partially defined), and 3 (well defined). These ratings considered if the variables in the publications were described with sufficient clarity to be replicated (collected) in further studies. As the previous stages of the review, authors independently (TM, MS) evaluated the completeness of the items mentioned in rounds until agreement was achieved.

Results

Abstracts of 1,809 publications were examined and 75 publications were retained according to the inclusion criteria. In the data collection process (second stage), of the 75 publications remaining, 40 were excluded due to different reasons as described in Figure 1. A total of 35 publications were included in our systematic review.

The studies included represent considerable diversity in terms of sample sizes, ranging from 29 to 379,601 health records. The largest sample sizes represent more than one unit in a single hospital or multiple hospitals, and more than one primary health care center. In these cases, the same EHR and SNTs were implemented across the institutions. Types of records portrayed in the studies encompass daily entries, or shift entries of nursing information into the EHRs at point-of-care. Episodes of care represent the entire history of a hospitalization, typically consisting of several entries for a unique patient hospitalization. For the purpose of this study, health records were defined as representing unique patients and health records entries were defined as daily/shift documentation of care.

In this review, we found studies for only three of the five terminology sets. There were no studies found for CCC and PNDS. There were 21 NNN articles, 10 Omaha System articles, three articles on ICNP, and one article that evaluated both NANDA-I and ICNP. Although all articles were published in English, we identified studies conducted in countries where English is not an official language. All articles on ICNP were conducted in South Korea, while all articles using the Omaha System were conducted in the US. The studies focusing on NNN were from Spain and U.S.

Descriptive, dependent and independent variables of each study were collected. In studies of descriptive nature, the variables analyzed were the nursing diagnoses, interventions and outcomes relevant to each set of SNTs. For studies in which statistical tests for associations were conducted, the majority of independent variables were the sets of SNTs, along with other factors related to delivery of care and the healthcare environment. Dependent variables in those studies varied significantly, including for example, occurrence of falls, pain, gender, age and medical diagnoses, nursing cost, among others.
Research foci encompassed different questions that could be answered by SNT coded nursing data. Twelve studies focused on characterization of units and patients regarding the most common nursing diagnoses, nursing interventions and nursing outcomes\textsuperscript{28-39} (Table 1). In two studies, authors described potential flaws in the terminologies regarding their content\textsuperscript{40-41} (Table 1).

Study of the nursing care through the use of SNTs included differences between standardized and non-standardized plans-of-care\textsuperscript{38}; comparison of nursing care provided and standard nursing care from guidelines\textsuperscript{32}; and testing different search strategies using nursing interventions to analyze the incidence of a medical complication.\textsuperscript{42} The benefits of using SNTs in EHRs to increase availability, validity, and reliability of data for statistical analyses and other research purposes were also described.\textsuperscript{34}

SNTs NOC and Omaha System were used across different studies, measuring changes in outcomes from admission to discharge.\textsuperscript{43-45} Both terminologies have a grading scale to be used by nurses daily to rate the nursing outcomes for each patient. In the studies\textsuperscript{43-45}, authors compared ratings given at admission and discharge to patients as a way of measuring effectiveness of care. Association of those ratings with other variables from EHR was also performed.\textsuperscript{43} One study focused on demonstrating differences between two statistical methods in describing changes in ratings for the Omaha System terminology.\textsuperscript{46}

**Figure 1.** PRISMA flow diagram illustrating the selection process of publications.
<table>
<thead>
<tr>
<th>SNT</th>
<th>Research focus</th>
</tr>
</thead>
</table>
| NNN  | To describe availability of plans-of-care data<sup>34</sup>  
To describe most prevalent NNN of an SNT-based electronic nursing documentation<sup>37</sup>  
To describe the most frequent NNN documented for hospitalized older patients with a primary discharge diagnosis-related group of pneumonia  
To describe the most frequent NNN documented for hospitalized older patients with a discharge primary diagnosis of heart failure<sup>33</sup>  
To report pain care from admission to discharge or death for end-of-life patients<sup>24</sup>  
To determine statistically significant changes in mean scores of outcomes from admission to discharge for specific medical diagnoses<sup>45</sup>  
To determine changes in outcomes scores from admission to discharge, and describe interventions for pediatric patients with dehydration<sup>44</sup>  
To measure the cost of delivering high surveillance, among hospitalized elders at risk for falling<sup>27</sup>  
To measure association among patient characteristics, unit characteristics, medical, pharmacy, and nursing interventions, and falls for older adults<sup>32</sup>  
To measure the association between high surveillance and low surveillance and the occurrence of failure to rescue<sup>46</sup>  
To establish association between degree of severity of problems and psychiatric diagnosis, number of nursing diagnoses, age, gender<sup>25</sup>  
To determine associations between the variables nursing characteristics and nursing cost per acute care episode<sup>26</sup>  
To identify patient and nurse-related factors associated with nursing outcomes for end-of-life patients with death anxiety<sup>41</sup>  
To report of NOC scores linked with the most frequent nursing interventions, age and length of stay<sup>16</sup>  
To describe the difference between plans-of-care with NNN and plans-of-care without NNN in relation to intermediate health outcomes<sup>9</sup>  
To measure association between the variable nurse continuity and the occurrence of pressure ulcers<sup>36</sup>  
Describe data mining techniques to predict whether a patient would meet the expected pain related outcomes<sup>35</sup>  
To develop predictive models that show how patient and nursing variables impact comfortable death outcome in end-of-life patients<sup>42</sup>  
To predict current NOC rating for each nurse shift based on previous shift ratings and factors related to nurse shift<sup>43</sup>  
To create predictive models that determine whether a patient with pain problems will be re-admitted to a hospital<sup>44</sup>  
To describe infomarkers that mark a shift in pain outcomes from standard care to palliative care among patients that died during hospitalization<sup>45</sup>  
| OS   | To describe client problems, interventions, categories and targets, along with outcomes documented for patients in a low-birth weight program<sup>28</sup>  
To describe all Omaha System data elements across multiple homecare settings using two different EHR vendors<sup>29</sup>  
To describe client risk index, family home visiting interventions between groups of low- and high-risk<sup>30</sup>  
To describe free text entry for the category Signs and Symptoms of the Omaha Classification<sup>31</sup>  
To describe free text entries associated with Omaha System on a software<sup>41</sup>  
To describe ratings of knowledge, behavior, and status on admission and discharge for Latina adolescent and adult mothers with mental problems<sup>43</sup>  
To demonstrate differences between p-values and Cohen’s d in describing clinically meaningful changes in scores for Omaha System problems<sup>46</sup>  
To determine which group of interventions predicts the outcome hospitalization for frail and non-frail elders<sup>47</sup>  
To measure association between patient, support system factors and nursing interventions and improvement in urinary and bowel incontinence<sup>49</sup>  
To determine the association between nurses, Omaha System interventions and client characteristics, and variability in health literacy<sup>40</sup>  
| ICNP | To describe variability in nursing intervention to prevent and treat pressure ulcer patients<sup>32</sup>  
To describe nursing preventive pressure-ulcer interventions and to compare with measures from two published guidelines<sup>35</sup>  
To describe narrative nursing statements of cancer patients treated with cisplatin-based chemotherapy<sup>34</sup>  
To test different computerized search strategies, to analyze the incidence and clinical characteristics of contrast media hypersensitivity<sup>42</sup>  
| N/ICNP | To describe key nursing diagnoses from the two terminologies for patients with heart failure<sup>39</sup>  
|<br>Key = NNN (NANDA-I, NOC, NIC); OS (Omaha System); N/ICNP (NANDA-I and ICNP). |
NOC and NIC were used to estimate nursing cost.\textsuperscript{26-27} Jenkins and Welton\textsuperscript{26} used the NOC rating scale as a measure of nursing intensity, and in turn, nursing cost. A different measure of nursing cost was used by Shever et al.\textsuperscript{27} In their study, NIC was used to measure the cost of delivering the nursing intervention surveillance. Surveillance was used as a measure of intensity of treatment, based on the average number of times a day that it was delivered (low or high). The number of times the intervention was performed was associated with the total hospital cost for the population under study.\textsuperscript{27}

In two studies the researchers examined relationships between nursing interventions and outcomes.\textsuperscript{47-48} Other (n=4) studies in the sample focused on establishing associations among a broader number of variables, such as patient characteristics, support system factors; nursing characteristics, nurse continuity, degree of severity of problems; medical diagnoses, pharmacy and nursing interventions, and nursing outcomes.\textsuperscript{22,25,50-51} Lodhi et al.\textsuperscript{51}, Lodhi et al.\textsuperscript{52}, and Lodhi et al.\textsuperscript{53} identified patient and nurse-related factors associated with meeting expected nursing outcomes.

Almasalha et al.\textsuperscript{23} focused on the use of SNTs to predict whether a patient would meet the expected outcomes by the end of a hospitalization period. Yao et al.\textsuperscript{54} identified specific elements of SNT related to changes in outcomes that could be used as an indicator for the adjustment from standard nursing care to palliative nursing care. Lastly, Lodhi et al.\textsuperscript{55} explored the creation of different predictive models for re-admission of patients that were diagnosed with pain during their last hospitalization.

### Table 2. Progression of nursing science and knowledge.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Year</th>
<th>Sample\textsuperscript{a}</th>
<th>Statistical analyses</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 university hospital, 2 large community hospitals, and 1 small community hospital</td>
<td>2012\textsuperscript{24}</td>
<td>40,747</td>
<td>Descriptive statistics</td>
</tr>
<tr>
<td></td>
<td>2013\textsuperscript{25}</td>
<td>569</td>
<td>Descriptive statistics, Data mining, Chi-square test, t-test</td>
</tr>
<tr>
<td></td>
<td>2013\textsuperscript{26}</td>
<td>596</td>
<td>Chi-square, Wald tests</td>
</tr>
<tr>
<td></td>
<td>2014\textsuperscript{27}</td>
<td>432</td>
<td>Data mining, Chi-square</td>
</tr>
<tr>
<td></td>
<td>2015\textsuperscript{28}</td>
<td>438</td>
<td>Data mining, Logistic regression, Chi-square</td>
</tr>
<tr>
<td></td>
<td>2015\textsuperscript{29}</td>
<td>160</td>
<td>Data mining, Logistic regression, Pearson’s two proportion z-test</td>
</tr>
<tr>
<td></td>
<td>2015\textsuperscript{30}</td>
<td>840</td>
<td>Descriptive statistics, Logistic regression</td>
</tr>
<tr>
<td></td>
<td>2015\textsuperscript{31}</td>
<td>901</td>
<td>Known-group comparative analysis, ANOVA, Tukey’s post-hoc test</td>
</tr>
<tr>
<td></td>
<td>2017\textsuperscript{32}</td>
<td>2,300</td>
<td>Data mining</td>
</tr>
<tr>
<td>3 community hospitals</td>
<td>2007\textsuperscript{33}</td>
<td>29</td>
<td>Descriptive statistics, t-test</td>
</tr>
<tr>
<td></td>
<td>2011\textsuperscript{34}</td>
<td>451</td>
<td>Descriptive statistics</td>
</tr>
<tr>
<td></td>
<td>2013\textsuperscript{35}</td>
<td>302</td>
<td>Descriptive statistics</td>
</tr>
<tr>
<td>1 academic medical center</td>
<td>2008\textsuperscript{36}</td>
<td>7,851</td>
<td>Propensity scores, Regression with generalized estimating equations</td>
</tr>
<tr>
<td></td>
<td>2011\textsuperscript{37}</td>
<td>7,851</td>
<td>Generalized estimating equations</td>
</tr>
<tr>
<td>1 homecare and 1 maternal home visiting program</td>
<td>2010\textsuperscript{38}</td>
<td>3,388</td>
<td>Descriptive statistics</td>
</tr>
<tr>
<td></td>
<td>2011\textsuperscript{39}</td>
<td>61,701\textsuperscript{**}</td>
<td>Descriptive statistics</td>
</tr>
<tr>
<td>1 public health agency</td>
<td>2007\textsuperscript{40}</td>
<td>75</td>
<td>Descriptive statistics</td>
</tr>
<tr>
<td></td>
<td>2011\textsuperscript{41}</td>
<td>486</td>
<td>Descriptive statistics</td>
</tr>
<tr>
<td></td>
<td>2013\textsuperscript{42}</td>
<td>1,016</td>
<td>Paired samples t-test, Effect sizes (Cohen’s d)</td>
</tr>
<tr>
<td></td>
<td>2015\textsuperscript{43}</td>
<td>141</td>
<td>Logistic mixed-effects model</td>
</tr>
<tr>
<td>15 home health agencies</td>
<td>2010\textsuperscript{44}</td>
<td>2,900</td>
<td>Descriptive statistics</td>
</tr>
<tr>
<td></td>
<td>2011\textsuperscript{45}</td>
<td>1,750</td>
<td>Data mining, Logistic regression</td>
</tr>
<tr>
<td></td>
<td>2011\textsuperscript{46}</td>
<td>2,072</td>
<td>Data mining, Logistic regression, Data mining, Chi-square</td>
</tr>
<tr>
<td>1 university hospital</td>
<td>2011\textsuperscript{47}</td>
<td>41,891</td>
<td>Descriptive statistics</td>
</tr>
<tr>
<td></td>
<td>2012\textsuperscript{48}</td>
<td>759\textsuperscript{**}</td>
<td>Descriptive statistics</td>
</tr>
<tr>
<td></td>
<td>2012\textsuperscript{49}</td>
<td>Study 1: 427, Study 2: 355</td>
<td>Study 1: Descriptive statistics, ANOVA, Study 2: Descriptive statistics</td>
</tr>
</tbody>
</table>

Key = * health records; ** health records entries

### Statistical analysis

A total of 15 studies used only descriptive statistics to analyze their data. The following statistical analysis methods were present among the remaining studies: propensity scores\textsuperscript{26-48}, logistic regression and other regression methods\textsuperscript{27-29,48}; generalized estimating equations\textsuperscript{22,25,50-51}; data mining techniques\textsuperscript{22,47,51-53}; chi-square\textsuperscript{22,24,51-53}; t-test\textsuperscript{22,39,44-46};
ANOVA\textsuperscript{35,54}; Pearson correlation\textsuperscript{25,53}; Tukey’s post-hoc test\textsuperscript{4}; logistical mixed-effects model\textsuperscript{50}; effect sizes (Cohen’s d)\textsuperscript{46}; general linear mixed methods models\textsuperscript{43}; and known-group comparative analysis\textsuperscript{54}

Data from same institutions were used across different publications to answer a diverse number of research questions. Table 2 shows the progression of science and knowledge in nursing across the years. Publications are grouped by dataset (each group of institutions used the same EHR). It can be seen (see Table 2) that earlier publications used mostly descriptive statistics and advanced to generalized estimating equations, logistic regressions, and data mining procedures. Sample sizes varied across publications and more specific and complex statistical analysis were related to smaller parts of the datasets.

Quality of publications included

Among the 35 publications included in this systematic review, all described their study size, which was represented by number of health records or health records entries, and listed statistical analyses used. Definitions of variables were scored for the 35 publications; 74\% of the publications\textsuperscript{22-25,27-33,37,39-40,44-46,51-56} scored a 3 (variables well defined), 23\% of the publications\textsuperscript{26,34-35,38,41,42,49-50} scored a 2, and 3\% of the publications\textsuperscript{43} scored a 1. Overall, the studies were rated as good for completeness of reporting of information.

Discussion

The present systematic review provides important information about the types of studies that have been conducted on SNTs coded nursing data retrieved from EHRs. The 35 studies were identified and reviewed through a comprehensive systematic process. The SNTs NANDA-I, NOC, NIC and Omaha System were those implemented most often across different EHRs. Specifically, NNN was represented more in the hospital setting, while Omaha System was present in the primary level of healthcare. Worth noting however is that both of these terminologies can be used interchangeably between hospitals and primary health care centers.

There was great diversity in the questions addressed in each study and the variables analyzed. Characterization of nursing care through the most common nursing diagnoses, interventions and outcomes was the aim of some studies. Others focused on determining significant changes in outcomes from admission to discharge, measure cost of nursing interventions, establishing associations using variables inherent to nursing care and other variables found in EHRs. The creation and use of predictive outcomes models were also studied. This range of study foci show the potential of obtaining and translating new knowledge from nursing data coded with SNTs and the holistic approach of nursing care. It also brings the attention of researchers to the possibility of asking new questions aimed at improving patient outcomes, nursing practice and the healthcare system.

Among the studies included in the sample, most investigators either utilized descriptive analysis (mean, standard deviation, frequency, percentage, etc.) to describe terminology usage in different healthcare settings or relied on classical tests (e.g., ANOVA, Tukey’s Post Hoc, Chi-square, t-test, Mann Whitney U test, correlation test) to determine bivariate associations. Although of limited nature, descriptive studies are a first step to understand a dataset and the target population under study. Descriptive statistics are also very efficient and helpful in the identification of predictors with low variability that could be excluded in further analysis.\textsuperscript{7} The descriptive studies included in this systematic review provide a foundation for the nursing scientific community to move forward to the use of more sophisticated statistical methods.

The association analyses were those between the adoption of SNTs and patients’ outcomes, between nursing diagnosis and episode length, and between nursing interventions and patient outcomes. A number of articles conducted multivariate analysis utilizing regression analysis (e.g., generalized linear mixed regression, generalized estimating equations, logistic regression, etc.) incorporating techniques such as propensity score and clustering to examine the effects of patient, nurse, and care characteristics on patient outcomes. Finally, there were a few studies\textsuperscript{51-53,56} using data mining methods to identify hidden patterns among nursing care data or to construct predictive models for patient outcomes. Interesting results were reported in these studies. For example, the pain management, medication, management, and positioning interventions were collectively associated with pain relief among end-of-life patients.\textsuperscript{13} Continuity of nursing care was not associated with occurrence of pressure ulcers\textsuperscript{56} and predictive modeling on large EHR SNTs data can be used as foundation in nursing decision support.\textsuperscript{57}

Our systematic review revealed strengths of using SNTs to code nursing data. Data stored in digital format that are unstructured or not standardized pose difficult challenges to data processing and analysis.\textsuperscript{19,57} Unstructured data may obscure linkages among elements, adversely impact the validity of the data collected, and inadvertently fuel misunderstandings and errors. Words that describe care might have different meanings from unit to unit, and between
institutions, due to the fact that standards of care and cultures can vary considerably. The processing of unstructured data is also more time-consuming and labor-intensive when compared to SNT coded nursing data.

Nursing data coded with SNTs facilitates comparisons of patient care within and across institutions. The evidence generated in turn can enhance the decision-making process and demonstrate the impact of nursing care making nursing practice visible. The use of valid standardized nursing terminologies makes extraction of nursing data easier, reduces the cost of reporting quality measures, and produces interoperable data.

There are some limitations of this systematic review. The search focused on two databases, which excluded research studies that could have been published in books or sources not indexed by the selected databases, although some articles from other databases were identified through grey literature search. Also, the performed search strategies and the search terms used to uncover the literature may have excluded publications on the subject that used different search terms or keywords. That the focus of our review was on study methods rather than study results limits the overall conclusions that are informed by this review and point to the need for a review of study results.

A limitation identified among the studies included in this systematic review is the scarce integration and analysis of SNT coded nursing data with other parts of EHRs. In our sample, only four studies included other variables from EHRs in their analysis, such as medical diagnoses, pharmaceutical treatments and pressure ulcer rating scale. To be a part of big data science, nursing data not only need to be coded with SNTs, but also should be combined with other parts of the EHRs in order to enable big data analysis. An example of big data analysis in the healthcare field would be to examine relationships among nursing diagnoses, medical conditions and laboratory results and their impact on patients’ outcomes.

A main lesson learned for nursing and big data science is the extraordinary potential of coding nursing care with SNTs and using the data to understand the impact of nursing and continuously expand nursing science. Our systematic review highlights the importance of SNTs to enable faster retrieval and processing of nursing data and the application of big data analysis. Big data science can benefit from the integration of nursing data to other datasets. The care provided by nurses is unique and an important contribution to patient outcomes. Thus, including the nursing component in analyses involving health care outcomes is essential to learning the improvements nurses can make to enhance the overall patient outcomes achieved. In the absence of analyzable nursing care data, the impact of nursing will continue to be elusive. This review provided a lens into the uses and potential benefits of SNTs now and in the future.

Conclusion

The use of SNTs to code nursing data enables the aggregation of this fundamental part of EHRs to big data datasets. Secondary data analyses of nursing data using methods such as data mining and clustering techniques should be used more often as a way of finding more meaningful results that could change positively the nursing practice and impact directly the care received by patients. This systematic review underscored the value of developing a deeper understanding of the meaning and potential impact of nursing variables before merging with other sources of data. Further research is needed to analyze nursing data along with other parts of the EHRs, revealing possible associations among care variables from different health professions.

References


Detection of Adverse Drug Reactions using Medical Named Entities on Twitter

Andrew MacKinlay, PhD, Hafsah Aamer, BA, Antonio Jimeno Yepes, PhD
IBM Research, Carlton, Victoria, Australia

Abstract

Adverse Drug Reactions (ADRs) are unintentional reactions caused by a drug or combination of drugs taken by a patient. The current ADR reporting systems inevitably have delays in reporting such events. The broad scope of social media conversations on sites such as Twitter means that inevitably health-related topics will be covered. This means that these sites could then be used to detect potentially novel ADRs with less latency for subsequent further investigation. In this work, we investigate ADR surveillance using a large corpus of Twitter data, containing around 50 billion tweets spanning 3 years (2012-2014), and evaluate against over 3000 drugs reported in the FAERS database. This is both a larger corpus and broader selection of drugs than previous work in the domain. We compare the ADRs identified using our method to the FDA Adverse Event Reporting System (FAERS) database of ADRs reported using more traditional techniques, and find that Twitter is a useful resource for ADR detection up to 72% micro-averaged precision. Micro-averaged recall of 6% is achievable using only 10% of Twitter, indicating that with a higher-volume or targeted feed it would be possible to detect a large percentage of ADRs.

Introduction

Pre-marketing clinical trials are used to identify adverse events (such as side effects) of drugs before they are introduced in the market. However such trials will inevitably fail to find many of the Adverse Drug Reactions (ADRs) associated with a drug due to the size and time constraints of such trials. In particular, they are unlikely to find rarer ADRs, those which only occur in combination with other drugs or those which only appear after prolonged usage. The increasing numbers of ADR-related deaths indicate that post-marketing ADRs are an important public health problem, indicating the need for continuous automated ADR surveillance.

There are existing mechanisms to collect adverse drug events, which are typically managed by government agencies. In the US, the Food and Drug Administration (FDA) provides the FDA Adverse Event Reporting System (FAERS) while similar databases exist in other countries. Clinicians and consumers can use FAERS to report serious reactions that they encounter when taking a drug. However, there is considerable friction for users of these systems – consumers may not even be aware of them, and for both clinicians and consumers, the overhead of printing out, filling in and mailing or faxing a multiple-page PDF will act as a disincentive to reporting ADRs. Given that the only motivation for submitting these reports is altruism, it is likely that many ADRs are underreported, particularly those that are not life-threatening. While a more streamlined web-based reporting system could help alleviate this, there will likely always be a barrier to entry to such a formal reporting system which will continue to deter reporting of ADRs. There is also a large delay in these reports, partially due to the pharmaceutical companies which may receive the initial report. Examination of the FAERS data shows that the delay between an event being first encountered and the FDA receiving the first report is on average 1.1 years, with a standard deviation of 6.6 years.

Post-marketing surveillance has been shown to be effective using Yahoo query logs. Other recent studies used Natural Language Processing to extract pharmacovigilance related information from online resources including DailyStrength, Twitter, MedHelp, SIDER and others. In this paper, we concentrate on Twitter as a resource for ADR surveillance. With its high volume of 500 million tweets per day, Twitter is a promising data source to for ADR surveillance from a research and monitoring perspective, as the data is readily available using a simple API and less encumbered by legal and privacy issues than data derived from medical records. For consumers who may report health information including ADRs, awareness of the platform in general is high – even if none would consider it to be an ADR reporting platform – and, crucially, the effort required to actually post information is very low. We investigate whether this lower barrier and corresponding data volume offsets the fact that Twitter is not a platform targeted at ADRs. In contrast to

---

1. The first two authors contributed equally to this work
2. See https://www.fda.gov/Safety/MedWatch/HowToReport/default.htm
previous work, our experiments have been run on a large number of tweets – around 50 billion, spanning 3 years (2012-2014) – and investigated over 3000 drugs reported in the FAERS database, thus it is the largest study around pharmacovigilance using Twitter. We automatically compare identified adverse events in Twitter to those reported in the FAERS database.

**Methods**

To perform Twitter-based pharmacovigilance, we needed to automatically process our Twitter database to identify ADRs related to a given drug. We first detect relevant tweets for ADR, then aggregate tweet information from the set of relevant tweets to identify ADRs related to a given drug. This is summarised in Figure 1, along with the processing of the FDA FAERS database, which is described in more detail below.

A reference symptom list is needed to map symptoms from FAERS and Twitter, so information from both sources can be compared. We propose three symptom lists that make different assumptions in their development. To develop these we use the UMLS (Unified Medical Language System) version 2015AA.

**Symptom lists**

A Twitter user may not mention symptoms as technical terms such as *insomnia*, instead using phrasing such as *can’t sleep*. Relatedly, the informal language can include a wide range of terminology, so that *vomiting* may be described as *throwing up, chucking up or puking* for example. To attempt to incorporate the informal language of symptoms, the UMLS was used as source for synonyms FAERS uses MedDra as controlled vocabulary for adverse events, so it can easily be mapped to entries in the UMLS. Since the UMLS incorporates deliberately less formal terminologies such as the Consumer Health Vocabulary (CHV), the extracted synonyms are more likely to encompass the language used by non-medical users on Twitter although it only includes moderately colloquial language: the moderately informal *throwing up* appears in CHV but highly colloquial *chucking up* does not.

We considered three symptom lists to map the symptoms identified from Twitter and the FAERS database, so it was possible to compare them. These lists are designed to cover different subsets of symptoms to allow for a better understanding of the ADRs covered in Twitter. The first, Wiki, is designed to cover the symptoms that a lay person might use to describe their symptoms, which is further described below. We generated a layperson list of symptoms automatically by combining Wikipedia’s List of Symptoms with those from the Yom-Tov and Gabriovich. The second, UMLS, covers the set of symptoms identified under the *Sign or Symptom* UMLS semantic type, while the third, UMLS2, extends this list with additional UMLS semantic types – specifically T184 (*Sign or Symptom*), T048 (*Mental or Behavioral Dysfunction*) and T033 (*Finding*). The UMLS lists where generated automatically using a local database installation of the UMLS.

Table 1 shows statistics on the symptom lists, including the number of unique symptom terms or synonyms and the average number of synonyms per symptom. Wiki has fewer symptoms but more synonyms. UMLS2 includes more
Table 1: Synonym statistics per symptom list. The number of symptoms (Symptoms), number of unique synonym terms (Unique Synonyms) and the average number of synonyms per symptom (Synonyms/Symptom) are shown.

<table>
<thead>
<tr>
<th>Data set</th>
<th>Symptoms</th>
<th>Unique Synonyms</th>
<th>Synonyms/Symptom</th>
</tr>
</thead>
<tbody>
<tr>
<td>WikiList</td>
<td>183</td>
<td>2,016</td>
<td>11.74</td>
</tr>
<tr>
<td>UMLS1</td>
<td>2,733</td>
<td>8,654</td>
<td>3.99</td>
</tr>
<tr>
<td>UMLS2</td>
<td>68,720</td>
<td>105,721</td>
<td>1.66</td>
</tr>
</tbody>
</table>

specific symptoms, which more often have only a single synonym, with 1.66 synonyms on average.

Drug name expansion

Different names might be used for drugs that have equivalent compositions, i.e. compounds might be distributed under several trade names. For instance, *alprazolam* is an anxiolytic that can be distributed under different names including *Xanax*. We used the UMLS Metathesaurus to find all trade names related to an equivalent active ingredient available from RxNorm. We start with an input drug name, then we use the RxNorm portion of the UMLS database (which provides names for all drugs available in the US) to map these to generic names before expanding out again to known trade names matching the generic name, then include all trade names and generic names as synonyms. For example, the trade name *Tylenol* would map to the generic name *Acetaminophen* which would in turn be expanded out to include brand names for other Acetaminophen preparations such as *Panateleve*, and all would be included as synonyms. This expansion process is used for matching drug names both from Twitter and the FAERS database, accounting for the possible variations in describing drugs in each case.

Extracting adverse drug reactions from the FAERS database

We downloaded all the files available up to October 2016 from the FAERS site and processed them following the workflow in Figure 1. FAERS contains adverse events reported much before 2012 (a broader time span compared to our Twitter data set). We extracted all the adverse reaction terms for drugs that were the primary suspect in a report. Drug names were filtered using RxNorm and a set of 3,050 drugs were selected and were used as the reference list of drugs for our study, discarding drugs not mappable to RxNorm. Examples of drugs not in RxNorm include *diclofenac potassium*, but its trade name *Cataflam* was included, even though it is not possible to determine this relationship automatically using the UMLS Metathesaurus.

In FAERS, adverse events are codified using the Medical Dictionary for Regulatory Activities (MedDra) terminology. MedDra terms have been mapped to the symptom lists mentioned above, so adverse events in FAERS can be compared to the adverse events identified from Twitter. Using the drug name expansion described above and the symptoms, the adverse events for a given drug are aggregated, so for each drug it is possible to find the number of reports in which an adverse event is mentioned.

Twitter database

We collected tweets over a three year period from 2012 to 2014 from GNIP Decahose, which provides a random sample of 10% of Twitter posts. We processed this data set of approximately 50 billion tweets by filtering out retweets and non-English tweets – yielding around 13 billion tweets – then automatically annotated them with named entities (NE) in the categories DISEASE, SYMPTOM and PHARMSUB (Pharmacological Substance) using a previously trained linear conditional random field (CRF) model on a manually annotated set of tweets. Tweets containing medical entities – approximately 230 million – were discarded for further processing into our Twitter database while others (the overwhelming majority) were discarded.

---

3[http://support.gnip.com/apis/firehose/overview.html#Decahose](http://support.gnip.com/apis/firehose/overview.html#Decahose)
Filtering of tweets for ADRs

The next step of this process was to filter these 230 million tweets containing medical entities to identify those that are relevant for ADR. This section describes the methods we used for this, using a pipeline of rule-based heuristic filtering followed by a second stage of filtering using machine learning.

Rule based filtering

The first step of the filtering was to select the tweets which mention a drug of interest generated from the FAERS database as explained above. The drug expansion name procedure described above was used for each drug in the FAERS-generated list to match the drugs in Twitter.

We extracted 18.9 million tweets (8.4% of total) with mentions of a drug listed in UMLS RxNorm database. We then filtered the tweets further to include only tweets with a Disease or Symptom entity as well as a PharmSub entity. We included both Disease and Symptom entities in our relevant tweet set because it is difficult to reliably distinguish between the two without considering the context, giving 533,000 tweets.

The automatically-annotated PharmSub entities contain many false positives, particularly due to drug names that coincide with words likely to be used in English. For example, Stay Awake is an alertness medication. Many tweets contained the phrase stay awake as a verb phrase with no relationship to the drug. Other such examples included fml, at 10 and my way. We removed tweets where the drug name was in this list which we produced by manual examination, yielding 393,000 tweets.

We also observed that many tweets mentioned drug names for advertisements, news and other purposes not related to first-hand experience with a drug and which were thus out of scope. For example: RT @BetterHealth1O1: Low fat yogurt contains tons of protein and calcium, which can fight hunger, stimulate weight loss by burning fat. This mentions calcium and weight loss but it is not indicating an ADR. We therefore filtered out the tweets with text that contained http assuming that these linked to advertisements, spam or news articles. Also, since much of the spam contained the word fact, due to usernames such as @AcneFacts, @thegoogleFact, @WhatTheFacts and @FactBook, such tweets were heuristically removed as well. These simple filters greatly reduced the volume of irrelevant tweets with relatively little effort, although more sophisticated relevance detection techniques could be considered.

Finally, there were entities that were annotated as PharmSub but they were too general to be taken as a cure for a disease and not specific enough for our study such as caffeine, cough syrup, vitamin d, zinc and pain killer. We manually prepared a stopword list based on the most frequently occurring mentions of these entities. After filtering out all the irrelevant tweets, we were left with 192,000 tweets potentially containing information about ADRs.

Feasibility of filtering using machine learning

By examining tweets in which a PharmSub co-occurs with a Disease or Symptom, we can find tweets which may describe ADR events. However, it would also be useful to determine whether the Disease or Symptom is the primary target of the drug – the condition which the drug is designed to treat – or an ADR. For example, if allergy is mentioned with a drug name, was the user taking the drug to treat existing allergy symptoms, or did allergy symptoms appear as a side effect of taking the drug? We attempted to answer this question using machine learning, by training a supervised classifier to identify relevant tweets for pharmovigilance.

We manually labelled 2,917 random tweets as either relevant or irrelevant for ADR detection. As tweets are short with limited context, differentiating symptoms which the drug treats from ADR symptoms is not easy for a human to do accurately. We also evaluated the results of merging our data with 4,799 ADR labelled tweets made available by Sarker et al.18, giving us 7,716 labelled tweets for training. We used Weka’s SVM implementation and a bag-of-words model, and evaluated the classifier using 10-fold cross-validation. The results of this are shown in Table 2. Our intention had been to use this classifier as an extra filtering layer for the 192,000 tweets remaining after the previously-

\[\text{This is likely due to the underlying NE tagger using features derived from a UMLS drug name database}\]

\[\text{The human annotator did not look up history for the user to add context, as this information was not available in our automatic classification scenario either}\]
described rule-based filtering, however in the resulting classifiers, the recall was below 36%. Incorporating this classifier into a filtering pipeline would unsurprisingly improve the overall precision, but the large loss in recall would be such that the classifier would be detrimental on balance, especially as achieving good recall is the biggest challenge. Such classification could still have value for precision-focused applications of ADR (for example if we wanted to apply these methods to the entirety of Twitter with lower risk of false positives), but for the remainder of this paper, we report on results without this filtering step.

<table>
<thead>
<tr>
<th>Data set</th>
<th>Total</th>
<th>Relevant</th>
<th>F1</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Our Data Set</td>
<td>2917</td>
<td>285</td>
<td>0.398</td>
<td>0.453</td>
<td>0.354</td>
</tr>
<tr>
<td>Sarker et al.18</td>
<td>4799</td>
<td>523</td>
<td>0.396</td>
<td>0.481</td>
<td>0.337</td>
</tr>
<tr>
<td>Combined</td>
<td>7716</td>
<td>808</td>
<td>0.392</td>
<td>0.450</td>
<td>0.348</td>
</tr>
</tbody>
</table>

Table 2: The distribution of classes in the training data and the performance achieved.

Results

One of the potential benefits of our approach is timeliness of information compared to government-controlled databases which can be slow to respond to new information. As such, the most informative analysis would evaluate how often the ADRs detected using our method preceded the first mention in FAERS. Our Twitter data set has tweets from 2012, but since many of the drugs considered in this study were released before 2012, it would have vastly reduced the number of drugs that we could have considered, which is a key aspect of this study. The downloadable FAERS data contains information about several dates, including the event date, the date the manufacturer was notified and dates when FAERS received the first notification; this information could be considered in a follow up study to quantify how timely the Twitter ADR detection can be.

Instead, we ignored time information, and compared our ADRs derived from Twitter with those reported in the FAERS database as explained above. We calculated precision, recall and F-score as both macro-averages (aggregated by drug, ignoring drugs which have no positives produced by our method and thus undefined precision) and micro-averages (including all drug-ADR combinations which came from FAERS).

This evaluation is suboptimal in another important way – our ground truth derived from FAERS is imperfect. In particular many ADRs may not have ended up in there due to aforementioned reporting delays or underreporting. So, if our method produces a purported false positive according to this evaluation, it may be a true but hitherto unknown ADR. A manual examination of random selection of tweets relating to false positive ADRs would indicate the plausibility off this, however this is left as an area for future work.

Table 3 shows the micro-averages and the macro-averages when comparing the results obtained from Twitter with the 3,055 drugs reported in FAERS. Note that not all of the drugs have a mapping to the list of ADRs that we have selected. The average numbers of ADRs per drug is significantly lower in our Twitter dataset compared to the number of ADRs reported per drug in FAERS. The approach clearly shows promise but there is much room for improvement in the recall. This work uses the GNIP Decahose, which has a large data volume but still comprises just 10% of Twitter. It is difficult to access the remaining 90% but it is likely that if we could apply these methods to the whole of Twitter (which is computationally very feasible) that the recall would be dramatically improved by detecting more ADRs from the long tail (although perhaps not linearly with the data set size, as many common drug-ADR combinations would be detected repeatedly). It is less clear whether or not expanding the data set would maintain the precision.

To get an idea of the trajectories of the precision and recall figures when changing the data set size, we have plotted learning curves using smaller subsets of the data, shown in Figure 2. The precision does decrease slightly as the data set size increases, however this is more than offset by the improved recall, and as a result the F-score shows a steady improvement as the data set size increases. We cannot justifiably claim that we could simply extrapolate linearly to

---

18Interestingly this was largely unchanged by the addition of the extra training data, perhaps due to different assumptions underlying the other data set meaning there was little extra new and consistent information for the classifier to learn from
Table 3: Results selecting drugs for which FAERS ADRs could be mapped to the symptom lists. Number of drugs from FAERS (FDrugs), number of drugs from Twitter (TDrugs), average number of ADRs per drug in FAERS (FADRs) and Twitter (TADRs) and performance measures are shown. Performance measures are reported for micro and macro-averages: miPr (micro precision), miRe (micro recall), miF1 (micro F1), maPr (macro precision), maRe (macro recall) and maF1 (macro F1). Macro-averages are aggregated by drug name, and exclude results with no predicted positives (and therefore undefined precision).

<table>
<thead>
<tr>
<th>Symptom List</th>
<th>FDrugs</th>
<th>TDrugs</th>
<th>FADRs</th>
<th>TADRs</th>
<th>miPr</th>
<th>miRe</th>
<th>miF1</th>
<th>maPr</th>
<th>maRe</th>
<th>maF1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wiki</td>
<td>2784</td>
<td>975</td>
<td>38.92</td>
<td>9.67</td>
<td>0.7203</td>
<td>0.0624</td>
<td>0.1149</td>
<td>0.7153</td>
<td>0.1128</td>
<td>0.1948</td>
</tr>
<tr>
<td>UMLS1</td>
<td>2750</td>
<td>1003</td>
<td>53.30</td>
<td>17.23</td>
<td>0.4530</td>
<td>0.0533</td>
<td>0.0953</td>
<td>0.5493</td>
<td>0.0942</td>
<td>0.1609</td>
</tr>
<tr>
<td>UMLS2</td>
<td>2919</td>
<td>1029</td>
<td>138.02</td>
<td>23.60</td>
<td>0.4882</td>
<td>0.0294</td>
<td>0.0555</td>
<td>0.5417</td>
<td>0.0561</td>
<td>0.1017</td>
</tr>
</tbody>
</table>

Figure 2: Learning curves for different proportions of Twitter used for the detection of ADRs. Micro-averages and macro-averages are shown for the Wiki and UMLS1 symptom lists.

extend these curves to the right for hypothetical larger data sets, however, it is almost certain that both recall and overall F-score would continue to improve as we added more Twitter data if such data were available.

Discussion

The analysis presented in the Results section summarises results on 3,050 drugs. In this section, we provide examples of the set of ADRs identified in Twitter mapped to FAERS and the ones missed by FAERS or Twitter.

Table 4 shows examples of the most frequent ADRs matched in Twitter and FAERS. There are three results, one per symptom list. The ADRs are similar in the three lists with some lexical variation for the same ADR, which only affects presentation but not performance. The number of distinct ADRs is larger for the UMLS2 ADRs list due to the larger
set of symptoms in this list.

<table>
<thead>
<tr>
<th>Count</th>
<th>Wiki</th>
<th>UMLS1</th>
<th>UMLS2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>headache 478</td>
<td>headache 478</td>
<td>headache 478</td>
</tr>
<tr>
<td>2</td>
<td>pain 477</td>
<td>pain 477</td>
<td>pain 477</td>
</tr>
<tr>
<td>3</td>
<td>nausea 339</td>
<td>nausea 339</td>
<td>pain adverse event by ctcae anatomic descriptor 467</td>
</tr>
<tr>
<td>4</td>
<td>itching 337</td>
<td>itching 337</td>
<td>adverse event associated with pain 467</td>
</tr>
<tr>
<td>5</td>
<td>dizziness 297</td>
<td>dizziness 295</td>
<td>nausea 339</td>
</tr>
<tr>
<td>6</td>
<td>vomit 294</td>
<td>emesis 294</td>
<td>itching 337</td>
</tr>
<tr>
<td>7</td>
<td>cough 280</td>
<td>cough 280</td>
<td>dizziness 295</td>
</tr>
<tr>
<td>8</td>
<td>insomnia 270</td>
<td>insomnia 270</td>
<td>emesis 294</td>
</tr>
<tr>
<td>9</td>
<td>rash 252</td>
<td>rash 252</td>
<td>cough 280</td>
</tr>
<tr>
<td>10</td>
<td>fever 248</td>
<td>fainting 239</td>
<td>insomnia 270</td>
</tr>
</tbody>
</table>

Table 4: True positives: 10 most frequent ADRs in FAERS that could be predicted using Twitter for a given drug. ADRs are shown by symptom list and the row Count shows the number of unique ADRs.

Table 5 shows the most frequent ADRs missed by Twitter. Some of the ADRs are able be identified at least for some drugs in Twitter as they appear in the true positives list in Table 4, and yet are still false negatives for individual drugs (e.g. nausea, vomit and dizziness). These false negatives are the largest category, and thus represent most of the divergences between FAERS and Twitter. Some ADRs such as diarrhea or tachycardia may be difficult to find in Twitter due to a lower likelihood of people mentioning these symptoms in such a medium due to taboos or specificity. Most are likely due to the latter: for any given drug, the majority of false negative ADRs (beyond what is presented in the table) are infrequently reported and often highly specific, representing the “long tail” of ADRs (with little validation beyond a small handful of reports), such as breast skin orange peel texture and edema eyelid. Capturing such events, the extent that is possible based on Twitter, could likely be improved by increasing the data volume even further, which we discuss further below.

<table>
<thead>
<tr>
<th>Count</th>
<th>Wiki</th>
<th>UMLS1</th>
<th>UMLS2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>vertigo 1699</td>
<td>vertigo 1699</td>
<td>vertigo 1699</td>
</tr>
<tr>
<td>2</td>
<td>dyspnea 1680</td>
<td>breathlessness 1680</td>
<td>breathlessness 1680</td>
</tr>
<tr>
<td>3</td>
<td>asthenia 1633</td>
<td>asthenia 1633</td>
<td>dizziness adverse event 1680</td>
</tr>
<tr>
<td>4</td>
<td>malaise 1601</td>
<td>malaise 1601</td>
<td>asthenia 1633</td>
</tr>
<tr>
<td>5</td>
<td>tachycardia 1537</td>
<td>tachycardia 1537</td>
<td>vomiting adverse event 1627</td>
</tr>
<tr>
<td>6</td>
<td>dizziness 1516</td>
<td>nausea 1505</td>
<td>malaise 1601</td>
</tr>
<tr>
<td>7</td>
<td>nausea 1505</td>
<td>emesis 1491</td>
<td>blood pressure low 1598</td>
</tr>
<tr>
<td>8</td>
<td>vomit 1491</td>
<td>dizziness 1479</td>
<td>tachycardia 1537</td>
</tr>
<tr>
<td>9</td>
<td>diarrhea 1449</td>
<td>diarrhea 1449</td>
<td>nausea 1505</td>
</tr>
<tr>
<td>10</td>
<td>confusion 1420</td>
<td>erythema 1434</td>
<td>emesis 1491</td>
</tr>
</tbody>
</table>

Table 5: False negatives: 10 most frequent ADRs in FAERS that could not be predicted using Twitter for a given drug. ADRs are shown by symptom list and the row Count shows the number of unique ADRs.

There are also ADRs that are discovered using our method from the Twitter data but are not present in FAERS. Examples of the most frequent ADRs from this category are present in Table 6. Some are clearly false positives. For example, there are some potential ADRs that are too generic to be collected by FAERS, e.g. sick, illness (finding). There are also some ADRs that seem very similar in meaning, e.g. asthenia and weak or pain and ache, which would require extra analysis to group into the same ADR. Some, such as hangover effect or black out, are sometimes used
colloquially, so it is not clear whether they would match to their literal interpretation if we more diligently mapped them to FAERS symptoms. On the other hand, there are some potential ADRs, for instance stomach ache or pyrosis for certain drugs, which may be hitherto unknown side effects for these drugs.

<table>
<thead>
<tr>
<th>Count</th>
<th>Wiki</th>
<th>UMLS1</th>
<th>UMLS2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>sick</td>
<td>377</td>
<td>ache</td>
</tr>
<tr>
<td>2</td>
<td>stomach ache</td>
<td>240</td>
<td>sore throat</td>
</tr>
<tr>
<td>3</td>
<td>cramp</td>
<td>237</td>
<td>black out</td>
</tr>
<tr>
<td>4</td>
<td>pyrosis</td>
<td>191</td>
<td>hunger</td>
</tr>
<tr>
<td>5</td>
<td>light-headed</td>
<td>142</td>
<td>muscle cramp</td>
</tr>
<tr>
<td>6</td>
<td>weak</td>
<td>130</td>
<td>sore to touch</td>
</tr>
<tr>
<td>7</td>
<td>thirst</td>
<td>110</td>
<td>sigh</td>
</tr>
<tr>
<td>8</td>
<td>pain</td>
<td>83</td>
<td>stomach upset</td>
</tr>
<tr>
<td>9</td>
<td>headache</td>
<td>65</td>
<td>ache stomach</td>
</tr>
<tr>
<td>10</td>
<td>flatulence</td>
<td>64</td>
<td>hangover effect</td>
</tr>
</tbody>
</table>

Table 6: False positives: 10 most frequent predicted ADRs from Twitter which did not appear in FAERS for a given drug. Symptoms are shown by symptom list and the row Count shows the number of unique symptoms.

We have further examined specific drugs, selecting Aleve and Vicodin – some of the top occurring drugs in FAERS – for discussion. Table 7 shows true positive, false positive and false negative ADRs when comparing Twitter vs FAERS, using Wiki as the reference ADRs. True positives are ADRs found in both Twitter and FAERS, false positives are ADRs found in Twitter but not in FAERS and false negatives are ADRs found in FAERS only. Note that we show only the top 10 most frequent symptoms in both FAERS and Twitter.

<table>
<thead>
<tr>
<th>Drug</th>
<th>True Positives</th>
<th>False Positives</th>
<th>False Negatives</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aleve</td>
<td>headache, pain, back ache, back pain, toothache, cough, fever, sleepy, vomit, dizziness</td>
<td>cramp, stomach ache, sick, pyrosis, light-headed</td>
<td>vertigo, insomnia, itching, diarrhoea, rash, dyspnea, urticaria, dyspepsia, constipation, hematochezia</td>
</tr>
<tr>
<td>Vicodin</td>
<td>pain, nausea, back ache, back pain, toothache, nausea, sleep, itching, fainting, dizziness</td>
<td>cramp, sick, light-headed, stomach ache, bruise, weak</td>
<td>drug overdose, vertigo, malaise, urticaria, insomnia, dyspnea, hallucination, anxiety, hyperhidrosis, tachycardia</td>
</tr>
</tbody>
</table>

Table 7: ADRs matched, missed or falsely detected from the top ten set from Twitter against FAERS

Aleve is an over-the-counter (OTC) pain relief drug. Over 10 of the ADRs predicted by Twitter were matched for Aleve. We anticipate that since Aleve is an OTC drug it may not have extreme side effects. We also observed that the true positives included the primary targets of the drug: fever, headache or toothache could easily be the reason for using Aleve rather than a side effect. However, headache and pain also appear amongst the frequent side effect for Aleve in SIDER20 (Side Effect Resource database’s), showing the challenge of differentiating between primary target symptoms and ADRs. Vicodin is a prescription drug used for moderate to severe pain. Most of the side effects predicted by Twitter match the ADRs in FAERS. The false positives in some cases are similar to false negatives (light-headed vs vertigo), suggesting, as seen before, that additional work on normalization of symptoms might help.

Our Named Entity Recognition (NER)-based method is highly scalable to large datasets such as Twitter. However, it misses particular classes of information. In particular, some classes of ADRs are not often encoded as simple medical named entities. For example drug ineffective is a frequent MedDra reaction listed in FAERS, including for the drug Advil. And indeed, when Advil is mentioned in Twitter, it is often in ways which indicate that users are experiencing a lack of effectiveness (but which our method is not able to detect), e.g. I still have a headache, Advil get your shit together or I must have taken like 8 Advil today and I still feel and look like hammered shit #sosore.

4http://sideeffects.embl.de
Our method also has no knowledge of the context surrounding a particular NE mention. Often this is not a problem but in particular cases it greatly affects the semantics of an entity. For example, sleepy was usually annotated as a symptom by our NE tagger, however in many of these cases it was modified by not, e.g. I’m just not sleepy tonight (and there could also be more complex forms of phrasal negation such as never very sleepy these days). This negation changes the symptom from sleepiness to insomnia, suggesting we need to incorporate modification such as negation in our annotations. Finally we could doubtless improve the normalization of the symptoms, so that vertigo could be mapped to light-headed to account for vocabulary differences between people. Word embeddings may be a useful starting point for this but we have not yet investigated this; these could potentially also help with capturing language too informal to be covered in the CHV such as the previously-mentioned chucking up example.

**Conclusion**

In this paper, we extracted relevant posts from Twitter containing mentions of drugs and symptoms from a three year time period (2012-2014) and processed them automatically using natural language processing to identify adverse drug reactions. The method shows promise as an additional source of ADR information to traditional reporting methods, with respectable precision of 42-75% depending on the dataset (or even higher if our ML-based filtering method is used). However, the low recall shows that this method still suffers from Twitter being a data set without any specific medical focus, but as we hypothesised, the volume of tweets available helps to partially offset this. Using 10% of all tweets over only 3 years, recall is only 3-6% but our learning curves support the fact that this would increase substantially with more data, either over a longer time period or with a higher relative proportion of Twitter.

We suspect the ADRs being uncovered using Twitter are at least somewhat complementary to those in FAERS (out of all our false positives, it is likely that some are actually novel ADRs), but this is difficult to assess without further investigation. It would be possible for a small team at a pharmaceutical company or public health authority to manually examine each false positive (ie potentially novel ADR) in close to real time, as after all of the filtering, the data volume is relatively manageable. Over the three year period, we uncovered around 200,000 possibly relevant tweets, of which around 50,000–100,000 would be false positives, which is around 50-100 tweets per day to manually examine. Many could be rapidly discarded (or they could be further pre-filtered using our trained ML model described above to improve the precision) and the remainder could be flagged for further investigation. While there is much room for refinement of our techniques, the primary gains would be by expanding to a larger data set to yield an effective technique for ADR using Twitter.

**Acknowledgments**

We would like to thank Elad Yom-Tov for the support to generate a symptom list based on Elad’s previous work.

**References**


Usability and Learnability of RxUniverse, an Enterprise-Wide App Prescribing Platform Used in an Academic Tertiary Care Hospital

Sonya Makhni, MD/ MBA1, Rivelle Zlatopolsky2, Farah Fasihuddin, MPH3, Misael Aponte4, Jason Rogers5, Ashish Atreja MD/ MPH.6
1-6Icahn School of Medicine at Mount Sinai, New York, NY

Abstract

The objective of this study was to assess the usability of RxUniverse, a novel platform that enables health care providers to directly disseminate proven, evidence-based mobile health apps to patients. Among five pilot clinical sites, 40 physicians and front-line providers were trained on the RxUniverse platform. They were educated on the platform’s functionality and instructed how to prescribe apps to their patients. The well-validated System Usability Score (SUS) was used to assess the usability of the platform. The adoption goal was set as 100 prescriptions of relevant apps within an 8-week pilot period. Within the pilot period, over 2000 apps were prescribed. Nineteen responses were received from the System Usability Score survey, and the platform received a usability score of 84.2, which is in the 96th percentile across all systems. The pilot study outcomes demonstrate the high adoption and usability of the RxUniverse platform.

Introduction

Digital medicine is the subset of new technologies and mobile applications that demonstrate, or aim to demonstrate, positive impacts in the treatment and management of disease. Several hundred thousand digital medicine tools, such as mobile health (mHealth) applications, have been developed and released. In fact, there are approximately 259,000 mHealth apps available across the major app stores, and these are produced by an estimated 58,000 mHealth publishers. These numbers are increased from 2015 and years prior, and as the clinical community continues to seek solutions that further engage and empower patients, we can expect continued growth within the mHealth space in the years to come.

Such extensive proliferation of mobile health technologies can in part be explained by the increasing body of research that correlates more engaged and activated patients with improved clinical, behavioral, and utilization outcomes. Mobile health applications can potentially serve as tools to empower patients to become more actively engaged in their care, and thus more effective self-managers of their own health. For these reasons, developers are creating mHealth solutions ranging widely from remote monitoring tools, to chronic disease management platforms, to behavioral modification apps.

The rapid proliferation of mHealth apps has led to two notable unintended challenges. The first is that providers have no standardized way of ascertaining which apps are best suited to their patients. Evaluation of clinical interventions traditionally relies on rigorous study design and peer-review, but these are lacking for the vast majority of mHealth apps. Many developers advertise the potential health benefits of a given mHealth app to would-be patients and prescribers, and consumer reviews for some products do exist, but it is difficult to ascertain how much of these claims are substantiated by rigorous, evidence-based proof. As a result, health care providers find it difficult to objectively assess an app’s efficacy and usability.

Second, providers have no streamlined way of directly disseminating the appropriate mHealth apps to patients. Current practices are bulky and place the burden on patients to locate and download the appropriate app, and on providers to accurately and efficiently direct them to the accurate app. There is a lack of effective filtering and organization of the mHealth apps within any given app store, and this translates to worsened execution in finding and disseminating a specific mHealth tool. As such, the benefits of mobile health technologies can neither be fully felt nor properly evaluated.

To address these concerns, RxUniverse (http://rxuniverse.com) was developed at the Mount Sinai AppLab. It is a digital medicine-centric care delivery platform that features a catalogue of apps that have been curated based off of published, evidence-based reviews regarding their efficacy and usability; additionally, it incorporates user-centered
features that enable efficient, direct dissemination to patients. The purpose of RxUniverse is to enable physicians to rapidly adopt digital medicine technologies and strategies within clinical care, to facilitate the effective dissemination of these technologies to patients, and to provide an ecosystem that promotes the evidence-based review of various digital medicine technologies.

A pilot study was conducted at the Mount Sinai Health System to assess the usability and adoption of RxUniverse, the first standardized platform to facilitate both curation and dissemination of mHealth apps.

**Methods**

*Platform Description and Specifications*

The platform assessed in this study was the RxUniverse mobile health platform, which was developed by Mount Sinai researchers at the Mount Sinai AppLab. RxUniverse integrates into any web browser, Android, or iOS device-based workflow. Development environment produced a Node.js web application built using AngularJS 1.4.3 framework hosted on an Azure back-end with a data model persisted in Microsoft SQL Server. Communication on the backend is done through full RESTful API for a data layer running on Azure SQL server. All data was stored on a HIPPA secure database that could only be securely accessed by members of the RxUniverse research team.

The application solves discovery challenges by allowing providers to find healthcare apps using an advanced filter relevant to their specialties, disease, type of app, or functionality, and to prescribe apps directly from their mobile device or within a linked EHR. RxUniverse addresses the evidence challenge by allowing providers to rank apps and share their rating with their peers, to find apps which are FDA approved, and to read a brief peer-reviewed synopsis about the app and links to evidence published in literature.

*Clinical Sites and Participants*

Five outpatient clinical sites within the Mount Sinai Health System were selected to participate in the eight-week pilot study: two primary care clinics, one pulmonology clinic, one gastroenterology clinic, and one cardiology clinic. Clinic sites were chosen based off the following general criteria: outpatient primary care or specialty clinic, standardized work-flows, clinician interest in integrating use of at least one mHealth app into clinical care. Project procedures were carried out as part of standard of care.

*User Training on RxUniverse Platform*

Among these sites, 40 physicians and front-line providers, consisting of medical assistants and receptionists, were trained on the use of the RxUniverse platform. They were individually instructed by one of two researchers using a standardized protocol of instruction, lasting approximately 5-7 minutes. Each instruction session consisted of introduction to the platform’s purpose, a demonstration of its functionality in both the web and mobile interfaces, and an observed trial process of prescribing an app. Upon completion of the training, each new user was given a unique username and password.

Users were instructed to direct message via text or email from the platform’s interface, or “prescribe”, a mobile health app to each patient seen. Eight to ten institutionally approved apps were pre-selected for use during the pilot study and were available to the users on the platform to prescribe to patients. Careful consideration was given to the specific operational workflows of each clinical site, and both office managers and user volunteers were consulted in devising the optimal implementation plan for RxUniverse at each pilot site so as to minimize time burden on the volunteer users. Three clinics primarily relied on front desk receptionists to disseminate mHealth apps upon patient check-in. Two clinics utilized medical assistants to prescribe apps, either during patient intake or while patients waited in the waiting area. Two of these five clinical sites used a hybrid method that integrated both medical assistants and front-desk receptionists.

*Measures*
The System Usability Scale (SUS) is a very reliable tool\(^4,5\) designed to obtain subjective feedback on overall usability and user satisfaction. According to Nielsen, usability is a “quality attribute that assesses how easy user interfaces are to use”\(^3\). Learnability is considered to be a component of usability and refers to the ease by which users are able to learn to use the technology\(^3\). This is a 10-item questionnaire with a 5-point Likert scale, with response options ranging from 1 (Strongly disagree) to 5 (Strongly agree). Items 1, 3, 5, 7, and 9 are positively worded and items 2, 4, 6, 8, and 10 are negatively worded. The SUS is able to effectively discern both good and bad usability features even with small sample sizes (<10). Ratings for SUS scores are as follows: 0–64 is unacceptable, 65–84 is acceptable, 85–100 is excellent, with a score of 82 representing the likelihood to recommend (LTR) threshold. Users are likely to recommend a product that has an average SUS score of 82, whereas users would not recommend a product that has an average SUS score of 67\(^6\). Using factor analysis, the SUS is able to provide additional information via two sub-scales: an 8-item “Usability” and 2-item “Learnability” scale\(^7,8\).

Analysis

Success of the RxUniverse platform was measured in terms of number of apps prescribed and system usability. The number and type of apps prescribed by each user was tracked on a daily basis. The prescription of 100 relevant app prescriptions within the 8-week pilot period was set as the adoption goal. Adoption targets were based off of a prior study conducted by Mount Sinai AppLab researchers in determining the impact of a novel mobile health app, HealthPROMISE, in improving health and satisfaction outcomes. Weekly progress updates were sent to volunteer users and their respective clinic managers and physician sponsors.

Following the completion of the pilot study, the well-validated System Usability Score (SUS) was used to assess the usability of the platform. Descriptive analysis of demographic and individual SUS questionnaire items were conducted in excel to examine measures of central tendency and variability as well as level of use of the users. User responses were then converted to 0–100 percentiles as per SUS guidelines\(^9\) and scored by comparing them to standard rating scales\(^10\). The mean SUS score for all participants was calculated. Likelihood to Recommend (LTR) was determined by comparing overall mean SUS score to industry thresholds.

Results

Adoption

Within the 8-week pilot period, over 2000 apps were prescribed across all users among the five clinical sites. Of the 40 providers trained on the RxUniverse platform, 26 prescribed >5 apps during the trial period. Of these 26 individuals, 18 prescribed >20 apps, 14 prescribed >50 apps, and 5 prescribed >80 apps (Table 1). 58% of users reported frequent use (weekly or daily) of the platform.

<table>
<thead>
<tr>
<th>Table 1. RxUniverse Survey Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Apps RX</td>
</tr>
<tr>
<td>Time Period</td>
</tr>
<tr>
<td>People Trained</td>
</tr>
<tr>
<td>Provider prescribed greater than 5 apps</td>
</tr>
<tr>
<td>Provider prescribed over 20 apps</td>
</tr>
<tr>
<td>Provider prescribed over 50 apps</td>
</tr>
<tr>
<td>Provider prescribed over 80 apps</td>
</tr>
</tbody>
</table>

Demographics

A total of 19 care providers / users completed the SUS questionnaire (response rate: 47.5%). Three respondents had more than 2 items with missing data on the SUS portion of their survey, they were still reported (Table 3), however
were excluded from the tabulation of the SUS scoring. In all, 63.16% of included respondents (n=12) reported their age as being between 36-45 and between 46-55 respectively. 73.68% of respondents were female (n=14). The sample was also diverse in terms of frequency with which users interacted with RxUniverse. In all, 57.9% (n=11) of the participants reported using RxUniverse “daily” or “weekly”, with the remaining 26.32% (n=5) only using the system a few times per month (Table 2, 3).

Table 2. Demographic Variables of SUS Respondents

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Number</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>5</td>
<td>26.32%</td>
</tr>
<tr>
<td>Female</td>
<td>14</td>
<td>73.68%</td>
</tr>
<tr>
<td>Total</td>
<td>19</td>
<td>100%</td>
</tr>
<tr>
<td>Age</td>
<td></td>
<td></td>
</tr>
<tr>
<td>18-29</td>
<td>2</td>
<td>10.53%</td>
</tr>
<tr>
<td>30-35</td>
<td>4</td>
<td>21.05%</td>
</tr>
<tr>
<td>36-45</td>
<td>6</td>
<td>31.58%</td>
</tr>
<tr>
<td>46-55</td>
<td>6</td>
<td>31.58%</td>
</tr>
<tr>
<td>55+</td>
<td>1</td>
<td>5.26%</td>
</tr>
<tr>
<td>Total</td>
<td>19</td>
<td>100.00%</td>
</tr>
<tr>
<td>Frequency</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Never</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td>Rarely</td>
<td>3</td>
<td>15.79%</td>
</tr>
<tr>
<td>Few times per Month</td>
<td>5</td>
<td>26.32%</td>
</tr>
<tr>
<td>Weekly</td>
<td>6</td>
<td>31.58%</td>
</tr>
<tr>
<td>Daily</td>
<td>5</td>
<td>26.32%</td>
</tr>
<tr>
<td>Total</td>
<td>19</td>
<td>100.00%</td>
</tr>
</tbody>
</table>

Table 3. Summary of SUS Questionnaire Results for Overall Sample

<table>
<thead>
<tr>
<th>SUS Items¹</th>
<th>Mean</th>
<th>95% CI</th>
<th>SD</th>
<th>Variance</th>
<th>Count</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1</td>
<td>3.79</td>
<td>3.27-4.31</td>
<td>1.06</td>
<td>1.11</td>
<td>19</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Q2</td>
<td>2.06</td>
<td>1.32-2.80</td>
<td>1.39</td>
<td>1.94</td>
<td>17</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Q3</td>
<td>4.61</td>
<td>4.19-5.03</td>
<td>0.83</td>
<td>0.68</td>
<td>18</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>Q4</td>
<td>1.4</td>
<td>0.82-1.98</td>
<td>1.02</td>
<td>1.04</td>
<td>15</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Q5</td>
<td>4.12</td>
<td>3.58-4.66</td>
<td>1.02</td>
<td>1.04</td>
<td>17</td>
<td>1</td>
<td>5</td>
</tr>
</tbody>
</table>

¹| SUS Items: | I think that I would like to use RxUniverse frequently. | 3.79 | 3.27-4.31 | 1.06 | 1.11 | 19 | 1 | 5 |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>I found RxUniverse unnecessarily complex.</td>
<td>2.06</td>
<td>1.32-2.80</td>
<td>1.39</td>
<td>1.94</td>
<td>17</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>I thought RxUniverse was easy to use.</td>
<td>4.61</td>
<td>4.19-5.03</td>
<td>0.83</td>
<td>0.68</td>
<td>18</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>I think I would need the support of a technical person to be able to use RxUniverse.</td>
<td>1.4</td>
<td>0.82-1.98</td>
<td>1.02</td>
<td>1.04</td>
<td>15</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>I found the various functions in RxUniverse were well integrated.</td>
<td>4.12</td>
<td>3.58-4.66</td>
<td>1.02</td>
<td>1.04</td>
<td>17</td>
<td>1</td>
<td>5</td>
</tr>
</tbody>
</table>
Q6 I thought there was too much inconsistency in the RxUniverse platform.  | 1.67 | 1.02-2.32 | 1.14 | 1.29 | 15 | 1 | 5
Q7 I would imagine that most people would learn to use RxUniverse very quickly.  | 4.44 | 4.02-4.87 | 0.83 | 0.69 | 18 | 3 | 5
Q8 I found the RxUniverse platform very cumbersome to use.  | 1.82 | 1.19-2.46 | 1.2 | 1.44 | 17 | 1 | 5
Q9 I felt very confident using RxUniverse.  | 4.67 | 4.29-5.05 | 0.75 | 0.56 | 18 | 2 | 5
Q10 I needed to learn a lot of things before I could get going with the RxUniverse platform.  | 1.5 | 1.06-1.94 | 0.79 | 0.63 | 16 | 1 | 5

* Items 2, 4, 6, 8, and 10 are negatively worded. Lower Means for these items represent higher perceived satisfaction.

SUS: System Usability Scale; CI: confidence interval; SD: standard deviation.

**System Usability and Satisfaction**

Overall, users felt that RxUniverse performed well. The group mean for overall SUS score was 84.2, an “Excellent” rating based on standard SUS¹. RxU met the industry benchmark SUS score of at least 80 for users to likely promote your product. The mean score for the “Usability” sub-scale was 82.7 and the mean score for the “Learnability” sub-scale was 90. Individual item means are reported in Table 4. The majority of users had a favorable opinion about RxUniverse in terms of how confident they felt using the system (Q9) (4.67 with 95% CI of 4.29-5.05) and ease of use (Q3 and Q7). The “I found RxUniverse unnecessarily complex” question (Q2) indicates the respondents who felt mild concern.

A raw SUS score of 84.2 has a higher SUS score than 96.19% of all products. We can be 95% confident the population SUS score is between 78.28 and 90.06. With a mean SUS of 84.2 and a standard deviation of 10.6 – compared to global SUS population benchmark of 68 and standard deviation of 12.5 – a t-statistic was calculated. The sample standard deviation was selected as a point of comparison, since it is more specific to this technology. The resulting t-score yields a p-value below 0.005 (0.001), allowing us to confirm the statistical significance of this pilot sample (Figures 1, 2). RxUniverse scored highest in the learnability category of the SUS score, with an average score found to be 90 (Figure 1).

![Figure 1. SUS Questionnaire Results](image-url)
As the pace of innovation continues to accelerate, health care providers will need to quickly integrate new digitally-based tools into their workflows, and patients will need to be able to easily and readily access these tools. It is equally important, however, to ensure that the technologies being disseminated to providers and patients pass performance and quality standards. RxUniverse not only provides the necessary mechanisms, user-friendly interface, and EHR integration functionality to accomplish these tasks, but it also surpasses industry standards in terms of usability and learnability.

The total number of apps prescribed on this platform surpassed 2000, which far exceeded the initial target of 100 apps. The type of mHealth apps prescribed varied widely, but the most common were MyChart, an app to help patients connect to their personal health records, a Mount Sinai Health Systems app, and two internally developed apps targeted Inflammatory Bowel Disease and Heart Failure. Additionally, the platform also scored an 84.2 (p = 0.001) on the System Usability Score, which is a score greater than 96.19% of all products and is considered to fall in the highest usability category of “excellent” (Table 4). According to industry standards, a score of 68 is considered to be average across all systems. By comparison, other health apps considered to be of respectable usability have reported scores of 77.5\(^10\). Additionally, RxUniverse received a high score within the learnability component of the SUS, which demonstrates the ease by which users were able to learn the technology utilized by the platform (Table 3). Research has shown that SUS scores provide reliable measures of usability and user satisfaction, and the high performance of RxUniverse on this scale demonstrates high achievement in these areas\(^8\). These outcomes demonstrate the high adoption and usability of the RxUniverse platform, an important platform that can be used to prescribe the latest technologies directly to patients.

Table 4. SUS Questionnaire Percentile and Subscales

<table>
<thead>
<tr>
<th>User</th>
<th>Percentile</th>
<th>SUS</th>
<th>Usability</th>
<th>Learnability</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>81.20%</td>
<td>77.5</td>
<td>81.3</td>
<td>62.5</td>
</tr>
<tr>
<td>2</td>
<td>97.20%</td>
<td>85</td>
<td>87.5</td>
<td>75</td>
</tr>
</tbody>
</table>
### Table

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>100%</td>
<td>92.5</td>
<td>90.6</td>
<td>100</td>
</tr>
<tr>
<td>4</td>
<td>100%</td>
<td>95</td>
<td>93.8</td>
<td>100</td>
</tr>
<tr>
<td>5</td>
<td>48.40%</td>
<td>67.5</td>
<td>59.4</td>
<td>100</td>
</tr>
<tr>
<td>6</td>
<td>48.40%</td>
<td>67.5</td>
<td>59.4</td>
<td>100</td>
</tr>
<tr>
<td>7</td>
<td>99.80%</td>
<td>90</td>
<td>87.5</td>
<td>100</td>
</tr>
<tr>
<td>8</td>
<td>64.80%</td>
<td>72.5</td>
<td>68.8</td>
<td>87.5</td>
</tr>
<tr>
<td>9</td>
<td>88.10%</td>
<td>80</td>
<td>78.1</td>
<td>87.5</td>
</tr>
<tr>
<td>10</td>
<td>100.00%</td>
<td>97.5</td>
<td>96.9</td>
<td>100</td>
</tr>
<tr>
<td>11</td>
<td>88.10%</td>
<td>80</td>
<td>75</td>
<td>100</td>
</tr>
<tr>
<td>12</td>
<td>100%</td>
<td>95</td>
<td>93.8</td>
<td>100</td>
</tr>
<tr>
<td>13</td>
<td>97.20%</td>
<td>85</td>
<td>81.3</td>
<td>100</td>
</tr>
<tr>
<td>14</td>
<td>81.20%</td>
<td>77.5</td>
<td>87.5</td>
<td>37.5</td>
</tr>
<tr>
<td>15</td>
<td>100.00%</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td><strong>Mean</strong></td>
<td><strong>96.20%</strong></td>
<td><strong>84.2</strong></td>
<td><strong>82.7</strong></td>
<td><strong>90</strong></td>
</tr>
</tbody>
</table>

This study may have some potential limitations. Surveys were offered to all users, but some bias may exist with respect to which users chose to fill out the survey or to participate in the pilot study. Additionally, the patient perspective was not assessed; the primary focus of this study, however, relates to the usability from the prescriber standpoint. Future iterations of this study will more closely involve the patient perspective.

### Conclusion

RxUniverse is the first platform that creates an ecosystem that facilitates a standardized process for dissemination and curation of mobile health apps. The goals of this pilot study were to evaluate the RxUniverse platform and to compare to industry standards of usability and learnability. To further validate this platform, it will be important to conduct additional studies both health system-wide, as well as in partnership with other health systems across the United States. The platform is now being expanded for multisite collaborators. Future plans include full integration within the Mount Sinai health system’s EHR, release of additional mHealth apps and content, assessment of usability from the patient perspective, and continued platform development and modification.

It is our belief that RxUniverse can serve as a valuable tool in connecting health technology innovation to the end users of clinicians and patients, as well as enable evidence-based review of mHealth solutions.
References


Deep Learning Meets Biomedical Ontologies: Knowledge Embeddings for Epilepsy

Ramon Maldonado, BS¹, Travis R. Goodwin, MS¹, Michael A. Skinner, MD¹,², Sanda M. Harabagiu, PhD¹

¹The University of Texas at Dallas, Richardson, TX; ²The University of Texas Southwestern Medical Center, Department of Surgery, Dallas, TX

Abstract

While biomedical ontologies have traditionally been used to guide the identification of concepts or relations in biomedical data, recent advances in deep learning are able to capture high-quality knowledge from textual data and represent it in graphical structures. As opposed to the top-down methodology used in the generation of ontologies, which starts with the principled design of the upper ontology, the bottom-up methodology enabled by deep learning encodes the likelihood that concepts share certain relations, as evidenced by data. In this paper, we present a knowledge representation produced by deep learning methods, called Medical Knowledge Embeddings (MKE), that encode medical concepts related to the study of epilepsy and the relations between them. Many of the epilepsy-relevant medical concepts from MKE are not yet available in existing biomedical ontologies, but are mentioned in vast collections of epilepsy-related medical records which also imply their relationships. The evaluation of the MKE indicates high accuracy of the medical concepts automatically identified from clinical text as well as promising results in terms of correctness and completeness of relations produced by deep learning.

Introduction

Over the past two decades, the biomedical research community has increased its efforts to produce ontologies encoding biomedical knowledge, justified by the steady increase in biological and biomedical research and the growth of data that is being collected in all areas of biology and medicine. Not only is the number of ontologies increasing and their size growing, but their relevance in biomedical research is also rising as they contribute to the interpretation of the biomedical data and enable complex inference from their encoding. The BioPortal* of the National Center for Biomedical Ontology (NCBO) is the most comprehensive repository of biomedical ontologies in the world (as of this writing it includes 541 ontologies, with almost 8 million classes and almost 40 million indexed records). Many of the ontologies available from the BioPortal became widely used resources, e.g. the Gene Ontology¹ (GO), one of the most important resources available in genomics research. The survey published in Huang et al. (2009)² discusses 68 bioinformatics enrichments tools informed by GO, that have played a very important and successful role contributing to the gene functional analysis of large gene lists for various high-throughput biological studies, evidenced by thousands of publications citing these tools. Moreover, Lependu et al. (2011)³ showed that it is possible to create reference annotation sets for enrichment analysis † using other ontologies than the GO, still available from BioPortal, for example, the Human Disease Ontology (DO). As reported in Noy et al. (2009)⁴, the ontologies in the BioPortal are publicly available in several formats, including OWL, RDF, OBO format or the Protege frame language. As such, they follow the principles of the OBO Foundry⁵, forming graph-theoretic structures, with concepts connected by edges representing relations such as ’Is-A’ or ’Part-Of’ or others from the OBO Relation Ontology (RO), generating well-principled ontologies for many biomedical domains.

Recently, a new ontology was added in the BioPortal, namely the Epilepsy Syndrome and Seizure Ontology‡ (ESSO), encoding 2,705 classes with an upper ontology targeting epilepsy as a disease and designed to be machine readable and to allow for federated queries across distributed databases and patient data capturing systems. The availability and development of the ESSO ontology answers the recommendations of the Institute of Medicine report§ for promoting the understanding of epilepsy by increasing the power of data in comprehensive, timely, and accurate epilepsy surveillance. Because epilepsy affects an estimated 2.2 million people in the United States, it is one of the most common neurological disorders. However ontological resources for this disorder are only now starting be become available to biomedical researchers. Nevertheless, large clinical datasets relevant to epilepsy are also becoming available. For example, the Temple University Hospital (TUH) EEG Corpus⁷ assembles over 25,000 sessions of electroencephalography (EEG) of 15,000 patients collected over 12 years. Clinical electroencephalography is an electrophysiological monitoring method used to record electrical activity of the brain.

*http://bioportal.bioontology.org
†Gene set enrichment (also functional enrichment analysis) is a method to identify classes of genes or proteins that are over-represented in a large set of genes or proteins, and may have an association with disease phenotypes.
‡http://bioportal.bioontology.org/ontologies/ESSO
§http://biopedia.snomed.org
representing the most important investigation in the diagnosis and management of epilepsies. As expected, EEG reports contain a wealth of epilepsy-related knowledge, derived from clinical practice. This knowledge is expressed by clinical language used in the reports and it explicitly mentions many of the concepts that can be linked to the ESSO ontology. Moreover, many implicit relations between these concepts can be inferred from the EEG reports. While biomedical ontologies have traditionally been used to guide the identification of concepts or relations in biomedical data, recent advances in deep learning were able to capture knowledge from textual data enabling an alternative knowledge representation.

This alternate knowledge representation, known as knowledge embeddings (KE), incorporates deep learning to model the interactions between concepts and relations and generate graphical knowledge structures. Knowledge embeddings are defined as multi-dimensional continuous vector representations of concepts and their relations. The KE methods were inspired, as reported in Weston et al. (2013)\(^8\), by the work of Craven et al. (1999)\(^9\), which matched the Yeast Protein Database with PubMed abstracts.

In this paper, we aim to investigate the medical knowledge embeddings (MKE) automatically learned from the TUH EEG corpus, encoding multiple EEG findings (e.g. EEG events and activities), associated medical problems, and treatments. Unlike the top-down methodology used in the generation of ontologies, which starts with the principled design of the upper ontology, the bottom-up methodology enabled by deep learning observes the likelihood that concepts share certain relations, as evidenced by data. Specifically, whereas the edges in an ontology graph represent hand-coded “hard” relations between entities, the edges in the knowledge graph are “softer”, i.e. probabilistic in nature. Unlike concepts and relations encoded in the BioPortal ontologies, the MKE associate relations between medical concepts with a probability or likelihood, enabling a probabilistic representation of biomedical knowledge. Thus, the MKE are able to account for the variability and inconsistencies in the way this knowledge is expressed in natural language by assigning more plausible relations a higher likelihood. While previous KE were generated from human curated knowledge bases, our work is unique in that we automatically extract entities and relations from free text in a data-driven approach. To the best of our knowledge, this is the first report of an the development of an embedded medical knowledge graph using free text clinical records.

We learned MKE representing 1,195,927 instances of binary relations between epilepsy-related concepts. These relations involved 2,442 instances of medical concepts. We evaluated the MKE by (1) the quality of the medical concepts identified in EEG reports; (2) assessing the plausibility of the potential relations discovered in EEG reports as well as (3) measuring the knowledge completeness as a form of link prediction\(^10\). We believe that the MKE encode medical knowledge that is complimentary to the knowledge available in traditional ontologies and can be used (1) to provide data-driven knowledge that can be linked to ontologies from BioPortal, and (2) as a potential mechanism for enriching existing ontologies using the learned concepts, relations, and probabilities.

**Background**

Recently, qualified medical knowledge graphs (QMKGs) automatically discerned from medical records have been used successfully in a system designed for patient cohort identification\(^11,12\). As reported in Goodwin & Harabagiu (2013)\(^13\) the QMKG was generated using big-data techniques applied to a large set of clinical records available to the participants in the TREC Medical Records track (TRECMed), a task developed in 2011 and 2012 as an Information Retrieval challenge pertinent to real-world clinical medicine and evaluated in the annual TExT Retrieval Conference (TREC) hosted by the National Institute for Standards and Technology (NIST). In another TREC special track on Clinical Decision Support (TREC-CDS), the system reported in Goodwin & Harabagiu (2016)\(^14\) used a knowledge representation as a Clinical Picture and Therapy Graph (CPTG) which was automatically acquired from the MIMIC-III\(^15\) clinical database. The TREC-CDS has addressed the challenge of retrieving bio-medical articles relevant to a medical case when answering one of three generic medical questions: (a) “what is the diagnosis?”; (b) “what test(s) should be ordered?”; and (c) “which treatment(s) should be administered?”. The system described in Goodwin & Harabagiu (2016)\(^14\) answered these types of questions by relying on a medical knowledge representation as a factorized Markov network\(^16\), suited ideally for answer inference.

Medical knowledge embeddings (MKE) enable a new probabilistic knowledge representation which is differs from the QMKG and the CPTG because (1) the relationships are not informed only by cohesive properties of texts, but by patterns of interactions between medical concepts, as captured by deep learning methods; and (2) similar medical concepts and relations share the same neighborhoods in the multi-dimensional space enabled by the knowledge embeddings. The latter property resolves semantic heterogeneity which arises when disparate terminology is used to refer to the same concepts or relations while identical terms may refer to distinct concepts. As noted in Sahoo et al. (2014)\(^17\) a seizure with alteration of consciousness
may be referred to as complex partial seizure, dialetic seizure or focal dyscognitive seizure by different epilepsy experts. An MKE representation should place all these expressions in a similar location of the multi-dimensional space, as it learns that they are involved in the similar relations with other epilepsy-relevant concepts. Thus, unlike the Epilepsy and Seizure Ontology\(^7\) (EpSO), the MKE representation does not require reconciliation of semantic heterogeneity, while being used for retrieving patient cohorts from medical records.\(^8\)

Data

In this work, we used the EEG reports publicly available from the Temple University Hospital (TUH), comprising over 25,000 EEG reports from over 15,000 patients collected over 12 years. Following the American Clinical Neurophysiology Society Guidelines for writing EEG reports, the reports from the TUH EEG Corpus start with a clinical history of the patient, including information about the patient’s age, gender, and conditions prevalent at the time of the recording followed by a list of the medications the patient is currently taking that might modify the EEG (e.g. “Keppra”, “Lamictal”). Both initial sections depict the clinical picture of the patient, containing a wealth of medical concepts, including the medical problems (e.g. “seizures”), signs, and symptoms (e.g. “loss of consciousness”) as well as significant medical events which may be temporally grounded (e.g. “2 years ago”). The following sections of the EEG report target mostly information related to the EEG techniques, findings and interpretation. The introduction section describes the techniques used for the EEG (e.g. “digital video routine EEG”, “standard 10-20 electrode placement system with additional anterior temporal and single lead EKG”), as well as the patients conditions prevalent at the time of the recording (e.g. fasting, sleep deprivation), level of consciousness (e.g. “during wakefulness”), and possible activating procedures that were performed (e.g. “hyperventilation”). The description section is the mandatory part of the EEG report, and it provides a complete and objective description of the EEG signal, noting all observed activity (e.g. “frontocentral beta activity”), patterns (e.g. “K-complexes”) and events (e.g. “eye opening”). The impression section states whether the EEG test is normal or abnormal (i.e. indicating some form of cerebral dysfunction). If the impression is abnormal, then the abnormalities are listed in order of importance. The clinical correlation section explains what the EEG findings mean in terms of clinical interpretation (e.g. “findings are consistent with idiopathic generalized epilepsy”).

Methods

Bottom-up knowledge acquisition methods rely on the automatic identification of concepts and relations from data to enable (i) the population of the knowledge representation and (ii) linking the acquired knowledge to existing ontologies. In learning medical knowledge embeddings (MKE) from EEG reports we do not only perform bottom-up acquisition of medical knowledge from EEG reports, but we also represent the knowledge probabilistically in a multi-dimensional space and perform inference on it. To do so, we followed a methodology which involves the following four steps:

**STEP 1:** Decide which medical concepts and which relations between them are expressed in the EEG reports;
**STEP 2:** Automatically generate the Knowledge Graph by extracting medical concepts and relations from the EEG reports;
**STEP 3:** Learn Medical Knowledge Embeddings (MKE) from the associated Knowledge Graph;
**STEP 4:** Perform inference with MKE.

It is to be noted that the the MKE represent only knowledge available from the EEG reports, which do not discuss the taxonomic organization of medical concepts or their partonymy relations. These forms of relations are encoded in medical ontologies, thus the MKE provide complementary knowledge to medical ontologies. However, many of the concepts represented in the MKE are also encoded in existing medical ontologies, providing a simple mechanism of linking the MKE to various ontologies available in BioPortal. For example, the clinical history and the medication list of EEG reports mention multiple medical concepts already encoded in the Unified Medical Language System (UMLS)\(^9\) ontology:

Example 1: **CLINICAL HISTORY:** This is a 20-year-old female with history of seizures described as generalized tonic-clonic with loss of consciousness for a few minutes. Last seizures occurred 2 years ago. **MEDICATIONS:** Keppra and Lamictal.

Medical problems such as seizures, and treatments such as “Keppra”, “Lamictal” are encoded in UMLS while concepts such as idiopathic generalized epilepsy will be linked both to UMLS and the ESSO ontology. However, these ontologies do not capture relations between such concepts that are implied in the EEG reports, e.g. which brain activities evidence some epilepsy-specific medical problems. Our four-step methodology aims to capture and represent such relationships, while also providing their probabilistic likelihood, learned automatically from the medical practice evidenced in the large corpus of EEG reports.

1228
**Figure 1:** Medical concepts and relations considered for Medical Knowledge Embeddings (MKE)

**STEP 1:** Decide which medical concepts and relations between them are expressed in EEG reports

In addition to medical problems and treatments that describe the clinical picture and therapy of a patient, EEG reports mention EEG events, which represent stimuli that activates the EEG (e.g. hyperventilation) and EEG activities, representing brain waves or sequences of waves. The section of the EEG reports describing the EEG record mention a multitude of EEG activities and events recognized by the neurologist from the analysis of the EEG signal. The following example illustrates mentions of EEG events such as photic stimulation and eye opening, while mentions of EEG activities are beta activity and polyspike discharges:

Example 2: DESCRIPTION OF THE RECORD: …the alpha rhythm was 9-10 Hz in frequency seen in the occipital region, which attenuates with eye opening. …Photonic stimulation was performed at multiple flash frequencies and results in a symmetric driving response without any photoparoxysmal response.

EEG activities are also mentioned in the impression section and in the clinical correlation section. Thus we decided to encode in the MKE four types of medical concepts: (1) EEG events; (2) EEG activities; (3) medical problems and (4) treatments. Whenever these concepts are also encoded in other ontologies, we linked to them. For example, medical problems such as idiopathic generalized epilepsy, when identified in an EEG report, with methods developed in the STEP 2 of our methodology, shall be linked to UMLS through its concept unique identifier (CUI). In addition to these four types of concepts, we decided to discern four types of binary relations that are implicit in the EEG reports. Each of these relations operates between a source argument and a destination argument. The relations along with examples of the four types of medical concepts are illustrated in Figure 1. The four binary relation types that we considered were motivated by discussions with several practicing neurologists and surgeons, corresponding to the implicit knowledge they discern from EEG reports. As shown in Figure 1, the EVIDENCES binary relation always has a medical problem as its destination concept, which is always mentioned in the clinical correlation section of the EEG report. The following example shows how the medical problem idiopathic generalized epilepsy, is evidenced by findings such as polyspike discharges, which is a mention of an EEG activity, in the impression section:

Example 3: IMPRESSION: This is an abnormal EEG recording capturing wakefulness through stage II sleep due to generalized spike and wave and polyspike discharges seen during wakefulness.

CLINICAL CORRELATION: The above findings are consistent with idiopathic generalized epilepsy.
Table 1: Examples of the Relations and Concepts expressed in EEG reports.

<table>
<thead>
<tr>
<th>Evidences</th>
<th>Evokes</th>
</tr>
</thead>
<tbody>
<tr>
<td>(seizures, EVIDENCES, idiopathic generalized epilepsy)</td>
<td>(photic stimulation, EVOKES, photic driving response)</td>
</tr>
<tr>
<td>(polyspike discharges, EVIDENCES, idiopathic generalized epilepsy)</td>
<td>(hyperventilation, EVOKES, slowing)</td>
</tr>
<tr>
<td>(facial grimacing, EVIDENCES, psychogenic seizure)</td>
<td>(seizures, EVOKES, periodic lateralized epileptiform discharge)</td>
</tr>
<tr>
<td>(toxoplamosis, EVIDENCES, degenerative brain disorder)</td>
<td>(shaking, EVOKES, rhythm)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Treatment For</th>
<th>Occurs With</th>
</tr>
</thead>
<tbody>
<tr>
<td>(lumical, TREATMENT-FOR, idiopathic generalized epilepsy)</td>
<td>(keppra, OCCURS-WITH, lumical)</td>
</tr>
<tr>
<td>(depakote, TREATMENT-FOR, generalized anxiety disorder)</td>
<td>(encephalopathies, OCCURS-WITH, occipital lobe epilepsy)</td>
</tr>
<tr>
<td>(dilantin, TREATMENT-FOR, hematoma, subdural, chronic)</td>
<td>(cerebral dysgenesis, OCCURS-WITH, recurrent convulsions)</td>
</tr>
<tr>
<td>(ampicillin, TREATMENT-FOR, infection of foot)</td>
<td>(spike and slow wave complex, OCCURS-WITH, polyspike complex)</td>
</tr>
</tbody>
</table>

Figure 2: Deep Learning Architectures used for Recognizing Qualified Medical Concepts from EEG Reports.

As shown in Figure 1, the EVIDENCES relation considers EEG events, EEG activities, treatments, and medical problems as providing evidence for the medical problem from the clinical correlation section of the EEG report. The EVOKES binary relation always has an EEG activity as a destination concept, as it attempts to capture the medical concepts that evoke the respective EEG activity. Those medical concepts can be either EEG events, or other EEG activities, medical problems or treatments followed by the patient. The third relation, namely OCCURS-WITH constraints both its arguments to be of the same type, e.g. either EEG activities, medical problems or treatments. The TREATMENT-FOR relation captures the treatments prescribed for certain medical problems. Table 1 illustrates examples of each of the four relations we considered, involving medical concepts illustrated in Figure 1, which lists all the EEG events and EEG activities that we decided to encode in the MKE, while providing several examples of medical problems and treatments, along with their UMLS CUIs. We used the vocabularies of EEG Activities and EEG Events from Maldonado et al. (2017) based on the International Federation of Clinical Neurophysiology’s glossary of terms.

STEP 2: Automatically generate the Knowledge Graph by extracting medical concepts and relations from the EEG reports.

The extraction of medical knowledge from EEG reports consists of (1) automatic identification of medical concepts and (2) binary relation detection. Medical concept identification aims to recognize all the four types of concepts mentioned in EEG reports, along with their inferred polarity and modality. For identifying polarity of medical concepts in EEG reports, we considered that each concept can have either a negative or a positive polarity, depending on whether the medical concept was negated or not in the text. The recognition of the modality, as in Maldonado et al. (2017) uses the modality values of factual, possible, and proposed to indicate that medical concepts mentioned in the EEG reports are actual findings, possible findings and findings that may be true at some point in the future, respectively. Through the identification of modality and polarity of the clinical concepts, we aimed to capture the neurologists beliefs about the clinical concepts mentioned in the EEG report. Thus our medical concept identification method needed also to qualify the concepts by their polarity and modality.

Medical Concept Identification was performed by taking advantage of our existing active deep learning methodology, which is illustrated in Figure 2. This methodology first uses two stacked Long Short-Term Memory (LSTM) networks for detecting the boundaries of medical concepts in the text of the EEG report. The task of identifying spans of text that correspond to mentions of medical concepts is called boundary detection. We relied on one stacked LSTM to identify the boundaries of EEG activities and another stacked LSTM to identify the boundaries of EEG events, medical problems and treatments. The
motivation for using two separate stacked LSTM networks is determined by the fact that different features are used to identify the boundaries of EEG activities than boundaries of the other three types of medical concepts, as detailed in Maldonado et al. (2017). Once the boundaries of each medical concepts are known, two Deep Rectified Linear Networks (DRLNs) are used to (a) identify the medical concepts and (b) discern their polarity and modality. EEG activities are identified only with their polarity, as their mentions are always factual, as illustrated in Figure 2. Moreover, medical problems and treatments are both normalized into UMLS concepts using MetaMap Lite.

Detecting Relations between Medical Concepts was possible when pairs of medical concepts identified in the same EEG report were considered. Specifically, we established the four types of relations illustrated in Figure 2 by considering: (1) a potential EVIDENCES relation between any medical concepts from an EEG report and a medical problem identified in its clinical correlation section; (2) a potential EVOKES relation between any medical concept and an EEG activity, provided that the treatments were not identified in the clinical correlation section, as they may indicate possible or recommended treatments; (3) a potential OCCURS-WITH relation between pairs of EEG activities, medical problems and treatments that are identified in the same section of the EEG report; and (4) a potential TREATMENT-FOR relation between any treatment and a medical problem identified in the history section of the EEG report. We discard potential relations involving medical concepts with “negative” polarities and “possible” modalities since these medical concepts, while mentioned, were not actually observed. All these potential relations are indicative of implied relations, that are not always directly stated in the text of the EEG report.

Taken together, the set of medical concepts extracted from the entire TUH EEG corpus along with the collection potential relations between them constitute a Knowledge Graph, \( G = \{ V, E \} \) where \( V \) is the set of graph vertices and \( E \) is the set of graph edges. In our knowledge graph, \( V \) is the set of medical concepts and \( E \) is the set of relations between them. The medical knowledge embeddings learned in the following step will provide the likelihood of any example of one of these relations.

STEP 3: Learning medical knowledge embeddings (MKE) from the associated knowledge graph

Learning MKE is made possible by relying on the TransE method, widely used for representing multi-relational data corresponding to concepts and relations by modeling concepts as points in a continuous vector space, \( \mathbb{R}^N \), called the embedding space, where \( N \) is a parameter indicating the dimensionality of the embedding space. In our use of the TransE framework, relations between medical concepts are represented as translation vectors, also in \( \mathbb{R}^N \), that connect the two points representing the two medical concepts in the embedding space. TransE learns an embedding, \( \vec{c}_i \), for each concept \( c_i \) and an embedding, \( \vec{r} \), for each relation type \( r \) such that the relation embedding is a translation vector between the two concept embeddings representing its arguments. This means that for any medical concept \( c_i \), the concept most likely to be related to \( c_i \) by the relation \( r \) should be the medical concept whose embedding is closest to \( (\vec{c}_i + \vec{r}) \) in the embedding space. By modeling the medical concepts as points in the embedded space and the relations between them as translation vectors, we can measure the plausibility of any potential relation between any pair of concepts using the geometric structure of the embedding space. The plausibility of a relation between a source medical concept and a destination medical concept, represented as a triple, \( \langle c_s, r, c_d \rangle \), is inversely proportional to the distance in the embedding space between the point predicted by our model \( (\vec{c}_i + \vec{r}) \) and the point in the embedding space representing the destination argument of the relation, i.e. \( \vec{c}_d \). In this work, we use Manhattan Distance as our distance function:

\[
f(c_s,r,c_d) = ||\vec{c}_s + \vec{r} - \vec{c}_d||_1
\]

where \( || \cdot ||_1 \) is the L1 norm. Using this distance function, plausible triples have low value of \( f \) (since \( \vec{c}_s + \vec{r} \approx \vec{c}_d \) for plausible triples) and implausible triples have a high value of \( f \).

Neural Network Architecture for learning MKE. To learn the optimal points and translation vectors, we use a neural network that will in fact produce the MKE. Formally, let \( C \) be the set of medical concepts found in the EEG reports and \( L \) be the set of relation types. Let \( X = \{ x^1 = \langle c_s^1, r^1, c_d^1 \rangle, \ldots, x^m = \langle c_s^m, r^m, c_d^m \rangle \} \) be the set of \( m \) relation triples extracted from the corpus of EEG reports at Step 2; where each \( c_s^j, c_d^j \in C \) is a medical concept and each \( r^j \in L \) is a relation type. The embedding, \( \vec{c}_i^j \), for a concept \( c_i^j \), is calculated by first generating a one-hot vector representation of \( c_i^j \) given by \( v(c_i^j) \) which is a \( |C| \)-dimensional vector of zeros with a one in the dimension corresponding to the index of the concept \( c_i^j \) in the set of concepts \( C \). The embedding \( \vec{c}_i^j = v(c_i^j) \cdot \text{E} \) is derived by multiplying the one-hot vector \( v(c_i^j) \) with the embedding matrix \( \text{E} \in \mathbb{R}^{|C| \times N} \). Each row of \( \text{E} \) corresponds to a medical concept embedding and the operation \( v(c_i^j) \cdot \text{E} \) corresponds to selecting the \( v(c_i^j) \)th row of \( \text{E} \). Likewise, the embedding for a relation type \( r^j \) is given by \( \vec{r}^j = w(r^j) \cdot \text{R} \) where \( w(r^j) \) maps \( r^j \) to a one-hot vector of size...
To learn useful embeddings we must also define a training objective that encodes useful relationships. Inspired by the work of Bordes et al. (2011)\textsuperscript{26}, we use the following training objective: if either the source argument or destination argument from a training triple is removed, the model should be able to correctly predict the correct medical concept. For example, the model should ensure that the value of $f(\text{keppra}, \text{TREATMENT-FOR}, \text{idiopathic generalized epilepsy})$ is less than the value of $f(\text{morphine}, \text{TREATMENT-FOR}, \text{idiopathic generalized epilepsy})$ since keppra is a treatment for idiopathic generalized epilepsy, but morphine is not. Formally, we wish to learn the values of $E$ and $R$ such that for any training triple $x_i = (c_s^{i}, r^{i}, c_d^{i})$, the following two constraints are met:

$$f(c_s^{i}, r^{i}, c_d^{i}) < f(c_s^{i}, r^{i}, c_d^{i}), \forall j: (c_s^{i}, r^{i}, c_d^{i}) \not\in X$$

$$f(c_s^{i}, r^{i}, c_d^{i}) < f(c_s^{i}, r^{i}, c_d^{i}), \forall j: (c_s^{i}, r^{i}, c_d^{i}) \not\in X$$

To learn the optimal embedding matrices $E$ and $R$, we optimize the objective defined by the constraints outlined in Equations 3-4 by iterating the following process:

1. Randomly select a training triple $x^i = (c_s^{i}, r^{i}, c_d^{i})$ from $X$.
2. Create a corrupted version of the triple $x^{i}_{\text{neg}}$ by selecting a medical concept $c^{\text{neg}}$ at random from the set of medical concepts $C$ and randomly replacing either $c_s^{i}$ or $c_d^{i}$ in $x^i$ such that $x^{i}_{\text{neg}} \not\in X$.
3. Update $E$ and $R$ by backpropagating the ranking margin loss\textsuperscript{23}, $\max(0, \gamma + f(x_i) - f(x^{i}_{\text{neg}}))$, where $\gamma$ is the margin parameter that determines how much of a margin should exist between triples in the training set and triples not in the training set.
4. Normalize each row $e$ of $E$ (i.e. $e := \frac{e}{||e||}$).

This process is repeated for each triple in $X$ a fixed number of iterations (200,000 in this work). Our collection of 1,195,927 relation triples extracted from the TUH EEG corpus consisted of $|X| = 138,369$ unique relation triples. It is important to note that, as reported in Bordes et al. (2013)\textsuperscript{10}, the normalization in the fourth step prevents the model from trivially minimizing the loss by artificially increasing entity embedding norms.

### STEP 4: Performing Inference with the MKE graph

Inference from a knowledge base can be viewed as answering questions using its encoded knowledge. Answering questions like (Q1) “what is the most likely treatment for idiopathic generalized epilepsy?”, (Q2) “what EEG activity is most likely to occur with polyspike discharges?”, and (Q3) “what is the likelihood that a patient with background slowing is diagnosed with cerebral dysfunction?” requires the ability to perform probabilistic inference. The MKE can be used to perform probabilistic inference by (1) representing the question as a relation triple $q$ and (2) measuring the plausibility of $q$ using equation 1 with the embeddings matrices $E$ and $R$ automatically learned from the TUH EEG corpus. We estimated the probability of $q = (c_s^{q}, r^{q}, c_d^{q})$ in terms of the geometric structure of the embedding space. Formally:

$$P(c_s^{q}, r^{q}, c_d^{q}) = 1 - \frac{f(c_s^{q}, r^{q}, c_d^{q})}{\sum_{(c_s^{i}, r^{i}, c_d^{i}) \in X} f(c_s^{i}, r^{i}, c_d^{i})}$$

For example, answering (Q1) is the result of $\hat{c}_s = \arg\max_{c_s \in C} P(c_s, \text{TREATMENT-FOR, idiopathic generalized epilepsy})$; answering (Q2) is the result of $\hat{c}_d = \arg\max_{c_d \in C} P(\text{polyspike discharges, OCCURS-WITH, c_d})$; and answering (Q3) is the result of $P(\text{background slowing, EVOKES, cerebral dysfunction})$.

### Experiments

To evaluate the MKE, we measure (a) the quality of the medical concepts that were extracted from the EEG reports as well as (b) the quality of the relations learned between them. When evaluating the medical concepts, we relied on the latest performance of our active deep learning annotation methodology\textsuperscript{21} and found that the quality of boundary detection of EEG activities had an F1-score of 0.9154 while the F1-score for detecting the boundaries of the other three forms of medical concepts was 0.9421. The identification of the medical concept type was performed with an F1-score of 0.9532 and the polarity was detected with an accuracy of 0.978 while the modality was recognized with an accuracy of 0.973.
Table 2: Quality of relations encoded in the MKE, measured using Pairwise Plausibility Accuracy (PPA), Mean Reciprocal Rank (MRR), Precision at 10 (P@10), Hits at 10 (H@10) and Hits at 100 (H@100).

The relations represented in the MKE were evaluated in terms of (a) their plausibility; and (b) their completeness. The plausibility of relations encoded in MKE was assessed in three ways, measuring how well MKE rank triples from a test set $T$, of 1,000 relation triples held out from the data used to train the MKE. For each triple $t$ in the test set, we randomly remove either the source or destination argument and produce a set of candidate triples by replacing the removed argument with every medical concept $c \in C$. We rank the candidate triples in ascending order according to the distance function $f$. This allows us to calculate the following metrics using the rankings produced from every triple in the test set:

- **Mean Reciprocal Rank (MRR)** is a standard ranking evaluation that measures how high the first correct triple is ranked according to the model. $MRR = \frac{1}{|T|} \sum_{i=1}^{|T|} \frac{1}{\text{rank}_i}$ where $\text{rank}_i$ refers to the rank of the first correct triple in the ranking, where a correct triple is defined as any triple from any of the training, validation, or tests sets.

- **Precision at 10 (P@10)** is another standard ranking evaluation that measures the percentage of the top 10 ranked triples are correct. As with MRR, correct triples are defined as any triple from any of the training, validation, or tests sets. The Precision at 10 evaluation shows how well the MKE ranks the triples about which the model is most confident.

- **Hits at K (H@K, H@100)** is a standard evaluation used for knowledge graph embeddings for evaluating link prediction. Hits at K measures how often the specific test triple $t$ occurs in the $K$ highest ranked triples, as opposed to precision which measures how often any correct triple occurs in the $k$ highest ranked triples. We report both Hits at 10 and Hits at 100 to illustrate how often $t$ is ranked among the most plausible triples, and how often $t$ is ranked in the top 5% of triples.

The evaluation of completeness of the relations from the MKE also used the test set, $T$. We evaluated how well the MKE can infer new knowledge in the form of new relations from the test set. To measure how well the MKE can model relations of the held out triples from the test set, we consider each test triple, $t \in T$, and a corrupted version of the test triple, $z$, created by randomly replacing either the source argument or destination argument with a random medical concept and compute the **Pairwise Plausibility Accuracy (PPA)**. The PPA measures the percentage of test triples for which the plausibility, $P(c^t_s, r^t, c^t_d)$, of the test triple $t$ is higher than plausibility, $P(c^z_s, r^z, c^d)$, of the corrupted triple. PPA demonstrates how well the MKE can differentiate between a correct, $t$, and an incorrect triple, $z$, even if the model had never encountered $t$. For these evaluations, the MKE were learned from 137,369 training triples automatically extracted from the TUH EEG corpus as described in the Methods section. We selected the dimension of the embedding space $N = 50$ from [25,50,100,200] and the margin parameter $\gamma = 1.0$ from $[0.1,1.0,5.0,10.0]$ using grid search on a validation set of 500 relation triples.

Table 2 presents these results. The results for the Pairwise Plausibility Accuracy show that the MKE can correctly distinguish between relations that occur in the data (but that the model has not seen during training) and corrupted relations 88.95% of the time. The micro-averaged Mean Reciprocal Rank of 83.33% indicates that for the majority of triples in the test set, the top ranked candidate triple is correct. While the MRR of the OCCURS-WITH relation is the lowest (62.3%), it should be noted that, on average, there is at least one correct candidate triple ranked in the top two. The Precision at 10 metrics show that 66.73% of the top 10 ranked triples were correct, in general. It is interesting to note that the results for the Hits at 10 metric have the most variability between relation types. For the OCCURS-WITH relation, test triple, $t$, only occurs within the top 10 ranked triples 27.77% of the time. In contrast, for the EVOKES relation, $t$ occurs within the top 10 ranked triples 84.91% of the time. In general, the Hits at 100 results show that the MKE correctly ranks $t$ in the top 5% of candidate triples 81.3% of the time.

**Discussion**

To analyze the correctness of medical knowledge distilled from EEG reports in the MKE, we manually inspected the 30 most plausible triples for each relation type. Specifically, for each triple, we determined whether that triple is consistent with established medical knowledge. Many of the triples in the MKE encode general knowledge which is difficult to judge. For example,
consider the triple \(\langle\text{Dilantin}, \text{TREATMENT-FOR}, \text{disease}\rangle\). Determining whether or not \text{dilatinin}\ is a treatment for \text{disease}, necessitates considering additional context specifying the disease. In general, we found the \text{EVOKE}\ relation type to have the highest percentage of correct triples, highlighting the ability of the MKE to capture neurological experience from EEG reports. By contrast, the MKE successfully identified a number of unexpected \text{OCCURS-WITH} relations, including \(\langle\text{hypothyroidism}, \text{OCCURS-WITH}, \text{turner syndrome}\rangle\), and \(\langle\text{infantile spasms}, \text{OCCURS-WITH}, \text{MELAS Syndrome}\rangle\). Whereas the coincidence of hypothyroidism and Turner Syndrome is fairly well known, the relationship between infantile spasms and MELAS syndrome is relatively obscure. Infantile Spasms, also known as West syndrome, is an exceedingly rare condition, with an estimated incidence in the United States of about 0.25-0.4 per 1000 live births\(^{27}\). The MELAS syndrome is an even rarer inherited disorder of mitochondrial function which may be responsible for 8% of cases of infantile spasms\(^{28}\). That the MKE recognized the connection between these two very rare conditions is quite interesting, and suggests that knowledge graph embedding holds promise for the elucidation of unusual concepts and relations from EEG reports in particular, and perhaps in medical reports more generally.

Owing to the data-driven nature of our technique, we generated some incorrect triples, as might be expected when using noisy free text data. For example, we observed two common types of errors when evaluating the \text{EVIDENCES} relation: (1) relation inversion, inverting the source and destination arguments of the relation; and (2) relation confusion, confusing one relation type with another. Consider the following example of a triple exhibiting relation inversion: (E1) \(\langle\text{liver cirrhosis}, \text{EVIDENCES}, \text{encephalopathies}\rangle\). As defined in Figure 1, the source argument of the \text{EVIDENCES} relation is a medical concept suggesting or supporting the diagnosis listed in the destination argument. By contrast, it could be argued that, for triple (E1), the destination argument \text{encephalopathies} more commonly evidences the source argument \text{liver cirrhosis}. We believe these types of error could be addressed by incorporating semantic attributes (e.g. temporal information) to contextualize or constrain the arguments allowed for each relation type. Relation confusion is exemplified by the triple (E2) \(\langle\text{rifaximin}, \text{EVIDENCES}, \text{brain diseases, metabolic}\rangle\). The source argument \text{rifaximin} is an antibiotic used in the management of the encephalopathy (i.e. the destination argument \text{brain diseases, metabolic}) related to severe liver failure. Thus, whereas there is a biologically plausible explanation for (E2), the \text{EVIDENCES} relation clearly does not accurately describe the relation; instead, the relation \text{OCCURS-WITH} may be preferred. This type of error could be mitigated in future work by introducing constraints into the knowledge embedding framework, as reported in Guo et al. (2015)\(^{23}\). Finally, there were rare cases in which the MKE assigned a high plausibility to triples in which the source argument contradicts the destination argument, i.e. \(\langle\text{insulin}, \text{TREATMENT-FOR}, \text{Diabetes Mellitus, Non-Insulin-Dependent}\rangle\). We believe that these types of error may be resolved by incorporating knowledge from existing ontologies to enforce consistency.

**Conclusion**

In this paper, we presented the medical knowledge embeddings (MKE) automatically learned from clinical text in EEG reports. Unlike traditional ontologies which encode curated knowledge, the MKE infers probabilistic knowledge by extracting a large number of potential relation triples. Experimental results demonstrate the promise of this approach and highlight the potential of the MKE for bridging the knowledge gaps of existing neurological ontologies. The MKE presented in this paper showcase the way in which deep learning techniques applied to large collections of medical records can supply medical knowledge derived from clinical practice to complement the knowledge already encoded in existing biomedical ontologies. By encoding the plausibility of medical knowledge, the MKE also enable probabilistic reasoning on its knowledge. Future work will consider techniques for learning plausibility thresholds that will allow MKE to be considered for curation and acceptance in existing, expert and community-validated biomedical ontologies.

**Acknowledgements**

Research reported in this publication was supported by the National Human Genome Research Institute of the National Institutes of Health under award number 1U01HG008468. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

**References**


Extracting Healthcare Quality Information from Unstructured Data

Shervin Malmasi, PhD1,2; Naoshi Hosomura, DDS, DMSc1,2; Lee-Shing Chang, MD1,2; C. Justin Brown1; Stephen Skentzos, MA3; Alexander Turchin, MD, MS1,2
1Brigham and Women’s Hospital, Boston, MA; 2Harvard Medical School, Boston, MA; 3Kiip, Inc., San Francisco, CA

Abstract

Healthcare quality research is a fundamental task that involves assessing treatment patterns and measuring the associated patient outcomes to identify potential areas for improving healthcare. While both qualitative and quantitative approaches are used, a major obstacle for the quantitative approach is that many useful healthcare quality indicators are buried within provider narrative notes, requiring expensive and laborious manual chart review to identify and measure them. Information extraction is a key Natural Language Processing (NLP) task for discovering and mining critical knowledge buried in unstructured clinical data. Nevertheless, widespread adoption of NLP has yet to materialize; the technical skills required for the development or use of such software present a major barrier for medical researchers wishing to employ these methods. In this paper we introduce Canary, a free and open source solution designed for users without NLP and technical expertise and apply it to four tasks, aiming to measure the frequency of: (1) insulin decline; (2) statin medication decline; (3) adverse reactions to statins; and (4) bariatric surgery counselling. Our results demonstrate that this approach facilitates mining of unstructured data with high accuracy, enabling the extraction of actionable healthcare quality insights from free-text data sources.

1 Introduction

Improving the quality of healthcare is a fundamental but difficult task in any healthcare system. The first step to such improvements is the identification of specific performance deficits. This can be achieved through measuring healthcare quality by using and/or developing performance measures that can serve as indicators of quality. Such metrics range from simple measures (e.g., patient waiting time) to more sophisticated ones (e.g., time to administration of antibiotics to a patient with pneumonia).

Research methods in healthcare quality can be broadly categorized as qualitative and quantitative in nature. “Quality” is not an easily defined concept, but is a much more complicated and sophisticated issue. Consequently, the research questions that may arise do not always have quantifiable answers. Qualitative research has proven to be an effective way to answer some of these complex questions using approaches based on interviews, observation and data analysis.

Quantitative methods, on the other hand, aim to measure phenomena and statistically evaluate them. Common measures include prevalence, incidence, frequency and severity. This research generates numeric data by applying specific methods to preselected data. It also bears mentioning that mixed methods – those combining both qualitative and quantitative approaches – continue to gain wider currency in the research community.1 Regardless of the research paradigm, informatics plays an important role in obtaining the right data.

One issue with quantitative research is that beyond statistics derived from structured data (e.g., patient records or laboratory results), the development of measures for more sophisticated questions has proven to be more difficult.2 A major hindrance is that while some of the target information is readily available as structured data, the great majority of it is stored as unstructured data, such as free text written by care providers. This use of free text is borne out of the flexibility required by professionals in describing their observations, diagnoses and treatment strategies. However, the unstructured natural language data that they produce is not directly usable in large-scale quantitative analyses. To use this narrative data, it would need to be manually abstracted via labor-intensive chart review. Indeed, retrospective chart review is a commonly used method for identifying issues not well documented by other methods.3 While it is widely considered to be a laborious process, it has also been said that “chart review is more difficult than it appears on the surface.”4 A demonstrative example of these issues is seen in the work of Pivovarov et al.,5 who state:

“Indicators used to assess quality of care are often buried within patient records. To accurately abstract these quality indicators, specially trained nurses manually comb through patient records to locate relevant information. Our 2,600-bed institution employs 35 full-time data abstraction specialists dedicated to reporting quality metrics for 30 databases covering 13 disease states and processes of care.”

Given that these free-text notes contain vital clinical data and that their manual review is a costly process, this has led to the development of computational methods to process and mine them for information of interest.6 This is often done
via information extraction (IE), the task of identifying and extracting relevant fragments of text from a larger, unstructured document. Researchers have been developing such methods to extract the information that they require from their own data sources. Once extracted, this information is used for quantitative research, clinical decision support, evidence-based medicine or further processing.

The overarching aim of the present paper is to highlight how unstructured data can be utilized by researchers for quantitative healthcare research. We approach this by introducing a new platform for extracting information from free-text data, and demonstrate its application in four different projects focused on healthcare quality.

The introduction of information extraction software is an important facet of facilitating more sophisticated quantitative healthcare quality research. This is because despite the importance, cost and laborious nature of this task, as we highlighted above, no readily usable tools for biomedical researchers are available. Ideally, such tools should be easy to use for researchers, even those without any NLP or software development experience. However, no such free or open-source solution exists. In this paper we present Canary, an NLP-based information extraction platform designed to meet these criteria. Canary has been developed for processing clinical documents to support the extraction of data using user-defined information discovery parameters and vocabulary. The various components that form the Canary NLP pipeline and the system for capturing text fragments are described, followed by an empirical evaluation on clinical data from four projects.

2 Background: Approaches to Information Extraction

In this section we describe a number of different approaches that have been used for information extraction, highlighting some of their advantages as well as disadvantages. In doing so we aim to position Canary within the wider context of the information extraction literature.

Simple Text Matching: The most elementary and straightforward approach to IE is based on defining specific words or sequences of words to be matched and output. While this can work in the simplest of cases, such as identifying particular drugs, it is impractical for more complex tasks.

Pattern Matching: Some shortcomings of the string matching method can be addressed by using more expressive and powerful pattern matching methods. Regular expressions are a common technique commonly employed for this purpose. They can be useful in extending patterns to match variations (e.g., different patterns of expressing drugs and dosage) or to account for other patterns like typographical errors. Disadvantages include the large number of rules required to capture all possible variations as well as difficulties in maintaining and updating the rules. They are also unable to capture structure, as we will discuss in the next section.

Language Parsing: The methods described thus far can be considered as “shallow” text processing techniques, relying solely on the words as they appear in text. Another approach involves using a “deeper” understanding of the text by parsing it to produce syntactic representations of the data; this can include constituency or dependency parsing. The parsing process can add linguistic information such as part-of-speech tags for each word as well as structural information such as noun phrases or prepositional phrases. This information can then be incorporated into the IE rules to produce more accurate and generalizable rules. These methods have been successfully applied in recent studies. For example, Wang et al. showed that such parser-based methods can be useful for automatic extraction of substance use information from clinical notes. One disadvantage of this approach is that the parsing process can be slow.

Supervised Learning: A more recently developed family of techniques based on statistical analysis of text is supervised machine learning. This involves the use of labelled training data to train a learning algorithm to identify elements of interest. While it does not involve manually engineering information extraction rules like the previous approaches, the costs of creating labelled data for supervised training are also significant. Moreover, learning algorithms require large amounts of data for effective training, along with the relevant expertise in tuning them. Nevertheless, a number of successful and popular machine learning-based toolkits for clinical information processing have been developed. The Automated Retrieval Console (ARC) is one such tool that attempts to eliminate rule creation via supervised learning. The Apache cTAKES system has also gained wide currency among clinical researchers in recent years. However, there is a population of clinical researchers without the prerequisite NLP and/or computer science skills that are not able to make use of these solutions. In this regard, the solution that we present here is complementary and designed to aid the aforementioned researchers in conducting their investigations in a self-sufficient manner.
3 Methods

3.1 Design

We designed a method for information extraction and developed NLP software implementing it. This platform was evaluated by conducting four quantitative healthcare quality experiments using unstructured clinical notes.

3.2 A Hybrid Information Extraction Method Based On User-Defined Parameters

The Canary software employs a hybrid approach for the IE task, combining the pattern matching and parsing approaches in order to address their shortcomings. The approaches are not mutually exclusive and can be complementary, as we show.

Regular expressions can be very useful for recognizing words or chunks of text, but they are not designed for capturing structure within text. While they can be helpful for identifying text in a pre-specified format, such as dates, numbers, email addresses or measurements, they are not suitable for more sophisticated structures with greater variance in their composition. For example, writing a regular expression to match text fragments describing a body part can be a cumbersome and error-prone task, given the large number of possible variations. Let us consider the following example phrases of interest:

<table>
<thead>
<tr>
<th>(1) left hand</th>
<th>(2) anterior cruciate ligament</th>
</tr>
</thead>
<tbody>
<tr>
<td>(3) lower back region</td>
<td>(4) lateral ankle ligament</td>
</tr>
<tr>
<td>(5) right hand’s index finger</td>
<td>(6) upper left abdominal quadrant</td>
</tr>
</tbody>
</table>

A key shortcoming of regular expressions is their inability to capture recursive structures, such as nested components in a tree structure. For example, they are unable to match nested brackets within a string.

Phrase (5) above is an example of such a nested structure in which a single entity is composed of two smaller entities: right hand and index finger. Capturing such nested constructs requires the definition of recursive rules, something of which regular expressions are inherently incapable.

On the other hand, detecting nested constructs can be achieved through a parsing-based approach in which recursive grammatical rules can be defined. While all aspects of regular expressions can be implemented through parsing, only some of the parsing functionality can be performed using regular expressions.

Another disadvantage of using regular expressions is that the rules encompass both words and their possible ordering. However, given the size of the target vocabulary that researchers use (particularly in the medical domain), it would be advantageous to separate the lexical entries and the ways in which they can be combined. That is to say, split the rules into a vocabulary (a set of recognized words and their categories) and a grammar (rules defining how words can be combined). This can be easily addressed by a parse-based solution by defining a vocabulary of terms and grammatical rules that define how they can be combined.

A related issue with regular expressions is that they can quickly grow in complexity, resulting in an unwieldy set of cryptic rules that can be very hard to comprehend. Updating such rules is also fraught with difficulties. Researchers have noted that the modification and documentation of regular expressions are a source of difficulty in their work. The isolation of vocabulary and grammar rules can help address these issues, as we will demonstrate.

The use of a complete language parser would require that users define their rules on top of the linguistic rules governing the target language (i.e., first categorize words into their grammatical categories and then create subsets of interest). An alternative approach, employed by Canary, is to allow users to create a simple, customizable grammar that enables them to model their target information, e.g., body parts or something broader that includes body parts. Complete language parsing is also a more computationally intensive task.

We now turn to a concrete example to illustrate how this approach works. The first step is to define the vocabulary or lexicon, which is a set of words organized into word classes. A class is a grouping of words from the same semantic category. For our body part example, we can define two classes referring to body parts and anatomical adjectives, which could be defined as:

- ANATOMICAL → (bi)?-?~(lateral)?, anterior, caudal, upper, lower, left, right, [...]
- BODYPART → (gastro)?-?intestinal, (gastro)?-?esophageal, (musculo)?-?skeletal, abdom.?, abdomen, [...]

1238
Users may define as many classes as needed, and the words belonging to each class can be matched using a regular expression, as seen in the above example. This allows users to create customizable ontologies to meet their needs.

The second step involves defining grammatical rules that define how these word classes can be combined for form phrases. A phrase can be a single word or a combination of words, as allowed by the grammar.

$$BODYPARTPHRASE \rightarrow BODYPART$$

$$BODYPARTPHRASE \rightarrow ANATOMICAL BODYPART$$

The above rules state that a body part phrase can be a single body part or an anatomical adjective followed by a body part. These rules are then processed by a parser to match all fragments of a text that match any of the supplied rules. Words that are not found in the vocabulary are ignored. Some examples matched by this simple grammar are shown in Figure 1.

![Figure 1](image1.png)

**Figure 1.** Example fragments of text describing body parts matched by our simple grammar.

We can also extend the grammar to match nested body parts by simply adding a recursive rule:

$$BODYPARTPHRASE \rightarrow BODYPARTPHRASE BODYPARTPHRASE$$

A recursive rule expresses that a phrase may include a sub-phrase of the same type as its own constituent, allowing us to capture the recursive property of natural language. This is important because, for example, English nouns and sentences can be infinitely recursive. In the above example, this extension allows the capture of one or more adjacent body part phrases. We illustrate this with the examples in Figure 2.

![Figure 2](image2.png)

**Figure 2.** Examples of recursive body part phrases that contain other body parts (marked in blue) as constituents.

We can see that this simple extension allows the capture of the more complex phrases. This is something not possible with regular expressions. A parser, on the other hand, can find an arbitrary number of such nested elements.

Furthermore, these constituency grammar rules can be used to build more sophisticated phrases that contain multiple recursive elements, as we show in the next section.

### 3.2.1 Creating Longer Phrases

Researchers often need to extract complex phrases that capture more information than just a body part. To this end, the simple rules we have seen so far can serve as the building blocks for forming longer, more sophisticated phrases. Building on the previous example, we now extend the rules to detect a medical condition involving a body part. We first add additional entries to our vocabulary:

- **ARTICLE**: the, a, an
- **PREPOSITION**: of, in, [...]  
- **POSSPRONOUN**: his, her, your, my
- **CONDITION**: (hepato)?(to)?xicity, aches?, aching, [...]
We also add some grammar rules to capture a phrase indicating a condition involving a body part:

\[
\text{CONDITIONPHRASE} \rightarrow \text{BODYPARTPHRASE CONDITION} \\
\text{CONDITIONPHRASE} \rightarrow \text{CONDITION PREPOSITION BODYPARTPHRASE}
\]

Applying these rules, we can match the phrases shown in Figure 3. We can also observe that the body part phrase in the left example has three levels of nested phrases. This example demonstrates that an advantage of the grammar-based approach is that the items in our vocabulary can be easily referenced and used to build larger and more meaningful phrases, something that is not possible with regular expressions.

![Figure 3. Two parse trees for condition phrases that include recursive body part phrases.](image)

### 3.3 Canary: An NLP-based Information Extraction Platform

In order to facilitate the mining of text documents by researchers, we created a free software program called Canary,\(^\text{12}\) as shown in Figure 4 below. The software, which can be downloaded for free (http://canary.bwh.harvard.edu/), was developed to enable information extraction using the approach discussed in section 3.2. The software also includes numerous sample projects that demonstrate how the grammar-based approach described above is used in practice. A number of factors were considered in designing the Canary platform, some of which we highlight here.

**Ease of Use:** Canary was designed for users without software development or engineering experience, allowing the vocabulary and grammar rules to be defined through a unified graphical user interface (GUI). The output is generated in plain text format for easy analysis.

**Easy Setup:** Canary is made available as an off-the-shelf software solution. It can be installed in a local folder without administrator privileges and can even run off a flash drive. It was designed to work out of the box.

**Security and Privacy Concerns:** Researchers in health informatics often work with data that includes protected health information, requiring compliance with the appropriate legal and security measures to guard this information (e.g., HIPAA compliance). These data may be stored on local networks and behind firewalls, with strict rules governing their transmission. Although a number of cloud-based data processing and information extraction solutions have been proposed, the abovementioned restrictions may rule out their use for many clinical researchers. In such cases, the use of self-contained software packages that can be run on local machines may be a better option. This is one of the motivating reasons that underlies Canary’s design.

### 4 Insulin Decline

The first task we demonstrate is the quantitative assessment of insulin decline. It is anecdotally known that patients frequently decline medications that are recommended by their healthcare providers. However, little systematic data are available on this phenomenon. It is not known how commonly patients decline medications and how frequently they ultimately receive medications they initially declined. This information, if extracted, would be of great value to healthcare quality and outcome research.

Insulin is thought to be one of the medications that are especially frequently declined by patients. Many patients are reluctant to start injectable medications; others express fear that “once you start insulin, you can’t get off it.” Studies show that patients whose diabetes is poorly controlled on oral medications take a very long time to be started on
insulin; insulin decline by patients could be one of the reasons for this. However, data on insulin decline remains extremely limited.

One reason for the paucity of research in this area is that information on patients declining medications is not easily available. As these patients declined the medication before any prescription was written, no trail is generated in the data sources that are typically used to study medication prescribing, such as pharmacy insurance claims or EMR medication records. Instead, medication decline is primarily recorded in narrative notes, requiring labor-intensive chart review. To this end, NLP software like Canary holds great promise for allowing clinical researchers to access the valuable pieces of relevant information locked away among millions of unstructured health records. Accordingly, we evaluate our software on data for this task, using it to extract this information and conduct a study to assess the prevalence of patients who decline insulin.

![Canary Software](http://canary.bwh.harvard.edu/)

**Figure 4.** An overview of the Canary software.
4.1 Data Collection

Data for this study comes from the clinical records of all adult patients with diabetes treated in primary care practices affiliated with Massachusetts General Hospital and Brigham and Women’s Hospital between 2000 and 2014. The number of notes reviewed for each task varied between 600 and 50,000.

4.2 Insulin Decline Language Model

A manual review of a subset of the collected data containing 50,000 notes was conducted to identify instances of insulin decline by patients. This task required the largest set of notes due to the extremely low prevalence of the information. Next, Canary vocabulary and information extraction criteria were created by a clinical researcher with no formal training in NLP or software development. They were designed to detect the language used to document insulin decline in manually identified instances with maximum accuracy and generalizability. After iteratively refining the language model, it resulted in a set of 148 word classes and 284 rules.

4.3 Evaluation

The primary aim of our evaluation is to assess the accuracy with which instances of insulin decline can be detected using Canary. This evaluation was conducted against a held-out gold standard dataset of 1,501 provider notes which were randomly selected and independently annotated by trained pharmacy and medical students. The reviewers marked all sentences describing patients who refuse to take insulin. We then compare the reviewers’ annotation to output generated by Canary. We conducted our evaluation against this annotated gold standard at two levels of granularity:

- **Note level:** detecting notes that contain any mention of insulin decline anywhere in the document.
- **Sentence level:** detecting sentences that mention insulin decline, across all notes.

The sentence-level assessment is the more challenging task. At the note level, sensitivity (recall), specificity and positive predictive value (PPV/precision) were calculated. Specificity has no meaning at the sentence level because of the arbitrary nature of the tokenization process (i.e. token boundaries in free text are implementation specific), so only sensitivity and PPV were calculated. Results from this experiment will be used to estimate the prevalence of insulin decline in the unstructured data belonging to patients with diabetes.

4.4 Results

**Table 1:** Evaluation results on the test set of 1,501 manually annotated gold-standard notes. The 95% confidence interval is provided in parentheses.

<table>
<thead>
<tr>
<th></th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>PPV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Note-level</td>
<td>100.0% (76.8–100.0)</td>
<td>99.9% (99.6–100)</td>
<td>93.3% (68.0–99.8)</td>
</tr>
<tr>
<td>Sentence-level</td>
<td>100.0% (82.4–100.0)</td>
<td>N/A</td>
<td>95.0% (74.4–99.9)</td>
</tr>
</tbody>
</table>

The manually annotated test set included a total of 19 sentences mentioning insulin decline across only 14 notes. This is a note-level prevalence of 0.93%, highlighting the difficulty associated with identifying this information through manual review. It should also be noted that given the length of the average note, this prevalence is significantly lower at the sentence level, i.e., less than 1% of sentences.

We next applied Canary to the same data. The results for this evaluation are listed in Table 1. We observe that at the note level, Canary achieved a sensitivity of 100.0% and a PPV of 93.3%, with even better results at the sentence level.

5 Statin Decline

The study of medication decline, as outlined in the previous section, can be extended to other classes of drugs. Cardiovascular disease is the number one cause of death both in the United States and worldwide, and hypercholesterolemia is the most common risk factor. HMG-CoA reductase inhibitors (statins) reduce the risk of cardiovascular events in patients with hypercholesterolemia. Nevertheless, many patients at high cardiovascular risk are not taking statin therapy, likely leading to thousands of preventable deaths. The reasons for this are not fully understood, but it is thought that medication decline may play a role.

Our methodology can be used to empirically study this phenomenon and analyze initial statin decline rates, how frequently patients ultimately start statin therapy after initially declining it, and whether the provider who eventually prescribes statin is likely to be different from the one whose statin recommendation was initially refused in the population of patients at high cardiovascular risk.
To this end we conducted a preliminary study to assess the specificity and PPV of our method in detecting instances of statin decline.

5.1 Data Collection

Data for this study comes from the clinical records of all adult patients with coronary artery disease treated in primary care practices affiliated with Massachusetts General Hospital and Brigham and Women’s Hospital between 2000 and 2013.

5.2 Statin Decline Language Model

A set of 8,800 notes were used by a clinical researcher with no formal training in NLP or software development to create a Canary model to detect instances of statin decline. Canary vocabulary and grammar rules were created to match the text fragments with maximum accuracy and generalizability. This resulted in a model with 97 semantic word classes and 88 structures.

5.3 Evaluation and Results

A set of 4,000 held-out notes were used for evaluation. The model achieved sensitivity (true positive rate) of 88% and PPV (precision) of 92%. These preliminary results highlight the utility of our approach to studying this healthcare issue. We are now in the process of creating additional resources to measure the specificity of our models, in order to apply them for answering the research questions laid out in section 5.

6 Adverse Reactions to Statins

EMRs are widely used in documenting adverse reactions to medications. Clinical decision support systems that analyze previous allergies and reactions have been shown to dramatically decrease prescription errors, making this a widely-researched area. However, a substantial portion of these reactions are not documented in a structured format but instead persist as free text. To this end, we applied our methodology to quantitatively extract this information and conduct a study to assess the prevalence of adverse reaction information that is not present in a structured format. More specifically, we studied this for 3-hydroxy-3-methyl-glutaryl-CoA reductase inhibitors (i.e., statins), as they are thought to have high rates of adverse reactions reported by patients.

6.1 Data Collection

This study was conducted at Partners HealthCare System, a healthcare delivery network in eastern Massachusetts. Partners Healthcare maintains a network-wide repository of medication allergies called the Partners Enterprise Allergy Repository (PEAR). The EMR system used at Partners allows data entry into PEAR. Data for our study came from all patients who were recorded as having been prescribed a statin between 2000 and 2010. This resulted in a set of 4.7 million provider notes.

6.2 Statin Side Effects Language Model

Trained pharmacy students manually reviewed a set of 3,175 narrative provider notes, annotating instances of adverse statin reactions. These notes were randomly selected from a set of all notes written on the data a statin was noted as being discontinued in the EMR system. After the annotation, Canary vocabulary and grammar rules were created to match the text fragments with maximum accuracy and generalizability.

6.3 Evaluation and Results

The evaluation follows the same scheme as described in section 4.3 above, with the exception that a held-out gold standard dataset of 242 provider notes were randomly selected and independently annotated by two trained pharmacy students. The results for the first stage of evaluation are listed in Table 2. We observe that at the note level, Canary achieved a sensitivity of 87.4% and a PPV of 99.4%. As expected, detection at the sentence level is more challenging, with slightly lower results.

Table 2: Evaluation results on the test set of 242 manually annotated gold-standard notes. The 95% confidence interval is provided in parentheses.

<table>
<thead>
<tr>
<th></th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>PPV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Note-level</td>
<td>87.4% (83.0–91.8)</td>
<td>98.3% (96.5–100)</td>
<td>99.4% (98.2–100)</td>
</tr>
<tr>
<td>Sentence-level</td>
<td>80.6% (75.4–84.9)</td>
<td>N/A</td>
<td>98.6% (96.9–100)</td>
</tr>
</tbody>
</table>
Having validated the model and demonstrated that it can achieve high levels of accuracy, we proceeded to the second stage of our evaluation. Canary was used to process the full set of 4.7 million notes collected for the study. This resulted in the identification of 224,421 patients who were prescribed a statin during the study period, with 31,531 of these being flagged by our software as having had an adverse reaction to a statin side. However, only 9,020 (28.6%) of the patients had a statin reaction recorded in structured format (in PEAR). This result indicates that the great majority of care providers record drug reaction information only in unstructured data.

7 Bariatric Surgery Counselling
Bariatric surgery is the single most effective treatment for significant and sustained weight loss in obese patients and significantly improves numerous obesity-related comorbidities, including cardiovascular risk, hypertension, myocardial infarctions, strokes and cardiovascular deaths.

A critical step in a patient’s decision to undergo bariatric surgery is clinician discussion and recommendation of bariatric surgery to the patient. Little systematic data are available on the epidemiology of bariatric surgery recommendation. It is not known how commonly physicians discuss and recommend bariatric surgery to obese patients who are surgical candidates. One of the reasons for the paucity of research in this area is that information on physician recommendation of bariatric surgery is not easily available. This information is typically not reflected in either administrative or structured electronic clinical data, as no prescription or insurance claims data are generated. Instead, recommendation of bariatric surgery is primarily recorded in narrative notes, requiring labor-intensive chart review.

The methodology we have used thus far can be used to study this issue empirically. In attempting to quantify the recommendation of bariatric surgery within notes, any approach must be able to distinguish between discussion and mentions of prior procedures. Consequently, the identification of notes containing these two categories of information the aim of our preliminary project.

7.1 Data Collection
Data for this study comes from the clinical records of all adult patients with body mass index (BMI) $\geq 35$ kg/m$^2$ treated in primary care practices affiliated with Massachusetts General Hospital and Brigham and Women’s Hospital between 2000 and 2014.

7.2 Bariatric Surgery Counselling Language Model
A set of 300 notes were manually annotated by a trained pharmacy student and subsequently used by another researcher with no formal training in NLP or software development to create a Canary model to distinctly detect instances of prior surgery and surgery discussion. Canary vocabulary and grammar rules were created to match the text fragments with maximum accuracy and generalizability. This resulted in a model with 17 semantic word classes and 160 structures.

7.3 Evaluation and Results
A held-out test set of 300 notes were manually annotated by a trained pharmacy student and used for evaluation, which follows the same procedure as described in section 4.3. The results for both categories at the note- and sentence-level are listed in Table 3. This preliminary evaluation is very promising, showing that our models are able to achieve high accuracy and demonstrate the generalizability to new data.

| Table 3: Evaluation results for both categories on the test set of 300 manually annotated gold-standard notes. |
|---------------------------------|--------|--------|
| **Bariatric Surgery Discussion (Note-level)** | 90%    | 90%    |
| **Bariatric Surgery Discussion (Sentence-level)** | 85%    | 69%    |
| **Prior Bariatric Surgery (Note-level)** | 83%    | 90%    |
| **Prior Bariatric Surgery (Sentence-level)** | 44%    | 96%    |
8 Discussion and Conclusion

We described and demonstrated the application of information extraction for identifying actionable insights by mining clinical documents. We showed that this approach can assist with answering quantitative healthcare quality research questions whose answers are not easily derived from structured data sources.

As part of this approach we also presented Canary, an information extraction tool based on user-defined parameters and ontologies. The principal advantage of our tool is that it is a GUI-based software which does not require any technical background. In this study, this was underlined by the fact that it was used by several researchers without any such technical background to successfully create language models of important clinical phenomena. The feedback from these users was positive and the models were then tested on a large-scale set of provider notes.

This approach is also useful for outcomes research, which involves quantifying treatment patterns and measuring the associated patient outcomes, and is an important area of study for identifying potential areas for improving healthcare quality. To this end, the methodology described here could be used to quantitatively measure complex social and demographic issues in health services research.

References

A novel application of point-of-sales grocery transaction data to enhance community nutrition monitoring

Hiroshi Mamiya¹,², Erica E.M. Moodie², David L. Buckeridge¹,²

¹Surveillance Lab, McGill Clinical and Health Informatics, McGill University
1140 Avenue Pine, Montreal, Quebec, Canada, H1A1A3
²Department of Epidemiology, Biostatistics, and Occupational Health, McGill University
1020 Avenue Pine, Montreal, Quebec, Canada, H3A1A2

Abstract

Unhealthy eating is the most important preventable cause of global death and disability. Effective development and evaluation of preventive initiatives and the identification of disparities in dietary patterns require surveillance of nutrition at a community level. However, nutrition monitoring currently relies on dietary surveys, which cannot efficiently assess food selection at high spatial resolution. However, marketing companies continuously collect and centralize digital grocery transaction data from a geographically representative sample of chain retail food outlets through scanner technologies. We used these data to develop a model to predict store-level sales of carbonated soft drinks, which was applied to all chain food outlets in Montreal, Canada. The resulting map of purchase patterns provides a foundation for developing novel, high-resolution nutrition indicators that reflect dietary preferences at a community level. These detailed nutrition portraits will allow health agencies to tailor healthy eating interventions and promotion programs precisely to meet specific community needs.

Introduction

Unhealthy eating is the leading preventable cause of global death and disability, responsible for 11.3 million premature deaths and the loss of 241.4 million disability adjusted life years¹. Obesity and overweight are recognized as a global public health crisis due to the sharp increase in prevalence and their recognized role as risk factors for debilitating chronic diseases including various cancers, cardiovascular diseases, and type II diabetes²,³. Excess sugar intake is one of the most important contributors of weight gain, and carbonated soft drinks (soda) are the primary source of artificially added sugar in the United States ⁴. Soda intake is especially prevalent among individuals with low Socio-Economic Status (SES), thereby contributing to socio-economic inequalities in nutrition-related chronic illness⁵,⁶.

Effective preventive initiatives aimed at reducing nutritional inequalities and improving dietary patterns at population-level are urgently needed. Examples of such approaches include subsidization of healthy food, taxation of unhealthy food, nutrition re-formulation of food product, and increased regulation of unhealthy food marketing activities⁷–¹². At the community level, nutrition education and health promotion programs using existing community ties play a significant role in empowering individuals and communities to adopt healthy dietary behavior ¹³,¹⁴.

The planning and management of these interventions and providing direction to population nutrition research require a capacity to measure dietary patterns at a population scale ¹⁵,¹⁶. Dietary surveillance programs should provide up-to-date information about trends in dietary patterns that will support evaluation of the effectiveness of policies and the influence of socio-economically significant events¹⁶. Such information should be available at high spatial resolution to allow the identification of community-level nutrition disparities, to evaluate neighborhood-specific responses to policy interventions, and to capture the influence of physical environment (e.g. residential availability of food stores and walkability)¹⁶,¹⁷. In addition, the information needs to provide product-specific information to track the change of nutrition formulation over time.

To date, public health surveillance of nutrition has largely relied on (often ad-hoc, non-repeated) dietary surveys, which suffer from the underreporting of food intake and inaccurate recall¹⁸,¹⁹, a lack of detailed information of consumed products (i.e., nutrition contents, packaging, and promotion such as price discounting), and which prohibit
dietary assessment for small areas, especially for surveys with a small sample size. In Canada, a national nutrition survey is conducted only every 10 years, and it only allows dietary habits to be estimated at a low spatial resolution (e.g., provincial level). Consequently, many public health and community initiatives must be taken with little if any relevant and timely information.

Market researchers, food manufacturers, and retail industries use scanned grocery transaction data generated by retail food outlets to guide food product development and promotional activities. The stream of retail transaction records from a sample of grocery stores and non-conventional food outlets, such as pharmacies, are routinely collected and centralized in an automated manner by marketing firms such as the Nielsen Corporation. The data contain product details including product name, purchased quantity, price and promotion status at weekly level along with time-fixed store attributes including location, chain (banner) name and unique store code. Additionally, product-specific Universal Product Code (UPC) can be linked to existing nutrition composition marketing databases to enable automated classification of healthy/unhealthy food and product marketing activities, including packaging design.

Although infrequently used in public health research and practice, these point-of-purchase (i.e., store-level) data could provide a unique input to an automated nutrition surveillance system. Using these data, such a system could produce information on food purchasing patterns, neighborhood product affordability, and the availability and marketing of foods at a high spatio-temporal resolution. The effective application of this information would allow public health agencies and community health workers to access up-to-date community nutrition status and formulate health promotion planning and intervention required at jurisdictional level.

However, generating useful information from these point-of-purchase data is a non-trivial problem. Because the transaction data are available for only a sample of food outlets, assessment of food purchase patterns at small area requires estimation of sales data for out-of-sample stores. The unobserved sales from these out-of-sample stores can be predicted using data from observed stores and the neighborhood and store-level features available from comprehensive government and commercial business registry data. The objective of this study is to develop a sales prediction model using the data from sampled stores, and apply this model to predict sales for out-of-sample stores using store and neighborhood attributes available in the transaction data and business registry data. As an initial example, we develop and apply such a model for the prediction of soda products due to their recognized interest as the major source of artificially added sugar and highly debated product as a target for taxation to reduce population-level consumption.

Methods and Data

The target geographic region was the Census Metropolitan Area (CMA) of Montreal, Canada, which had a population of 3,824,211 inhabitants in 2011. The Nielsen Corporation selects chain retail stores from the Montreal CMA by stratified random sampling, where strata are defined by urban/suburban status, store size, and store type (e.g., supermarkets, pharmacies, mass supercenters). Inclusion criteria for this study are supermarkets (chain grocery stores), chain pharmacies, and supercenters (e.g., Wal-Mart). Excluded store types are independent stores, Warehouse (e.g., Costco), and dollar stores. Our preliminary work indicates that the grocery market share of these target stores among all retail food outlets is 65.2 percent in the Montreal CMA. To maintain the representativeness of the sample in the face of store closures and openings, periodic partial resampling is performed. We extracted the transaction data covering the 2012 calendar year as the study period of interest.

The scanner data consist of weekly aggregated store-level sales volume and information on each food item as defined by Universal Product Code (UPC). Transactions for soda are extracted from a food category labelled as ‘Carbonated Soft Drinks’, from which diet soda is identified and excluded by terms in the product description suggestive of diet beverages. For each store, we generated the average weekly soda sales in 2012 standardized to a single serving size (240ml), resulting in 128 data points (i.e. 128 sampled stores in 2012). Relevant predictive features of soda sales are chain identification code, store type (e.g., supermarkets, pharmacies, supercenters), store size (number of employee), and neighborhood socio-demographic attributes, which are median family income, proportion of individuals who received post-secondary diplomas, proportion of immigrants, and population density as measured by the 2011 Canadian Household Survey. Using linear regression, natural log-transformed soda sales...
were modelled as a function of these predictive features of soda sales. Selection of the predictors and first-order interaction terms was guided by the minimization of the mean squared error (MSE) using 10-fold cross-validation.

The resulting soda sales prediction model was applied to the out-of-sample food outlets in the Montreal CMA. Predictive features of the sales were supplied by the Canadian Business Point of Interest data, which contain annually updated data for all business establishments in Canada. The predicted store-specific weekly average soda purchases in 2012 were spatially interpolated to provide a graphical representation of the soda sales (indicator of unhealthy purchasing) across the Montreal CMA.

Results

The comparison of beverage category-specific sales indicates that soda, along with milk, were the most frequently purchased beverages in the Montreal CMA. The resulting prediction model as selected by cross-validation demonstrated a good model fit (adjusted R square; 0.95) and prediction error (MSE: 0.20). The highly variable nature of soda sales by chain (Figure 3), and their regression coefficients in the selected prediction model (Table 1) indicate that sales are strongly associated with attributes that are differentiated across the chain, such as store size, product assortment, pricing, and promotional patterns. Interestingly, neighborhood socio-economic status (as represented by income and education in the final model) were substantially less predictive of sales as compared to store chain. The spatial distribution of predicted soda sales is presented in Figure 4.
Figure 2. Trend of beverage category-specific transactions (servings in log base 10) observed between 2008 and 2013 in Montreal CMA.

Figure 3. Log (base10) soda sales among sampled stores by store chain in Montreal CMA, 2012.
Table 1. Selected predictive features of store-level natural log soda sales.

<table>
<thead>
<tr>
<th>Selected variable</th>
<th>Predicted change*</th>
<th>Lower 2.5% CI</th>
<th>Upper 2.5% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chain B</td>
<td>-0.27</td>
<td>-0.72</td>
<td>0.19</td>
</tr>
<tr>
<td>Chain C</td>
<td>0.28</td>
<td>-0.22</td>
<td>0.77</td>
</tr>
<tr>
<td>Chain D</td>
<td>-3.60</td>
<td>-4.03</td>
<td>-3.17</td>
</tr>
<tr>
<td>Chain E</td>
<td>-0.13</td>
<td>-0.58</td>
<td>0.32</td>
</tr>
<tr>
<td>Chain F</td>
<td>0.50</td>
<td>0.02</td>
<td>0.98</td>
</tr>
<tr>
<td>Chain G</td>
<td>0.91</td>
<td>0.45</td>
<td>1.37</td>
</tr>
<tr>
<td>Chain H</td>
<td>-0.65</td>
<td>-1.11</td>
<td>-0.18</td>
</tr>
<tr>
<td>Chain I</td>
<td>-1.86</td>
<td>-2.28</td>
<td>-1.43</td>
</tr>
<tr>
<td>Chain J</td>
<td>-0.68</td>
<td>-1.15</td>
<td>-0.21</td>
</tr>
<tr>
<td>Chain K</td>
<td>-7.22</td>
<td>-7.82</td>
<td>-6.62</td>
</tr>
<tr>
<td>Chain L</td>
<td>0.52</td>
<td>0.06</td>
<td>0.98</td>
</tr>
<tr>
<td>Chain M</td>
<td>-3.58</td>
<td>-4.04</td>
<td>-3.12</td>
</tr>
<tr>
<td>Education±</td>
<td>-0.87</td>
<td>-1.85</td>
<td>0.12</td>
</tr>
<tr>
<td>Income‡</td>
<td>0.00002</td>
<td>-0.00004</td>
<td>0.00006</td>
</tr>
</tbody>
</table>

CI: Confidence Interval.
Chain: indicator variable for retail chain, where chain A was used as a reference.
*Predicted change in natural log soda sales in standardized serving (240ml) by each predictor.
±Area-level proportion of post-secondary diploma or certificate. Thus, store-level log soda sales in an area with all residents having post-secondary education is -0.87 (95% CI: -1.85 to 0.12) lower than the store-level sales in an area with no residents attaining post-secondary education.
‡Area-level median family income in 10,000 Canadian dollars.
Discussions and conclusion

Our study harnessed the existing digital grocery transaction data to provide a novel indicator of neighborhood dietary patterns. This is one of the first applications of point-of-purchase transaction data to generate high-resolution surveillance information on population nutrition for public health. The method we propose creates a foundation for further exploration of how scanner data can be used to improve the assessment of population nutrition. The strong predictive power of store chains indicates that the mix of chains in a neighborhood at least partially explains spatial patterns in soda purchasing. Therefore, neighborhood mix of retail chain may serve as an environmental indicator useful in characterizing communities that are prone to excess soda consumption. Because pricing, product assortment, and various food marketing activities are available in the transaction data, further investigation could identify factors across chains that drive differential sales and which may be important in-store risk factors of unhealthy food selection.

The proposed approach provides an effective and novel solution for the automated and accurate estimation of dietary patterns for small area, which is not possible with the current nutrition surveillance relying on dietary surveys. Effective analysis of grocery transaction data will therefore provide important and novel information to enable evidence-based planning and evaluation of preventive strategies aimed at dietary risk factors. Because objective data collection was achieved by scanner technology at the store-level, the purchase and promotional status are free of measurement error arising from consumption reports from participants. In addition, these data are collected on a global scale by international marketing companies. Therefore, standardization of measurements and comparison of population nutrition status across regional, provincial, international national levels can be readily achieved.

The development and evaluation of a valid sales indicator requires several ongoing and remaining steps (planned as part of primary author’s doctoral thesis);
Spatial dependency of sales is being investigated using Conditional Autoregressive Model.

Using large travel surveys conducted in the Greater Montreal region, we are characterizing the shopping-related mobility of individuals, conditional on the area of residence and socio-demographic attributes. The travel distance and shape will be used to re-parametrize the smoothing method used in this study.

Sales at point locations (stores) are being converted into an area-level measure defined by socio-economically meaningful neighborhood spatial unit, and we are investigating their correlation with the area-level health outcome, including diabetes and obesity/overweight.

Finally, we intend to estimate the predictive performance of our area-level purchase indicator for person-level soda consumption records using dietary questionnaires obtained from residents through nutrition survey.

In addition to the sales measures, it is also possible to predict other store attributes, including the availability of (un)healthy food products and their affordability (food price relative to area-level income). Furthermore, these data also offer a highly time-varying view on marketing activities, such as temporary price discounting and flyer promotion, allowing measurement of neighborhood-level susceptibility to food marketing as demonstrated by our exploratory study. Although these promotional activities are recognized as strong drivers of food selection among market researchers, they are prohibitively expensive to capture by manual field (in-store) investigation. Our modeling can be extended to predict store and category-level promotional activities for healthy and unhealthy food, and it allows profiling food outlets (and thus neighborhoods) based on the exposure to food marketing to identify communities at risk of unhealthy food purchasing.

Three major limitations in this initial study should be noted. Although the study focused on non-diet soda items as an initial example, purchase patterns of a single food category provide limited information for the dietary assessment. Because the transaction data provide a full range of healthy/unhealthy food categories (typically greater than 100 categories in a supermarket), our model could be extended to a multivariate approach (joint modelling of multiple response categories), which will exploit the correlation of food sales across categories. Since the Nielsen corporation does not sample independent (non-chain) stores, our transaction data and thus prediction model is only applicable to chain stores, whose market share of all grocery products is approximately 65 percent. However, because chain supermarkets are the primary location of unhealthy food purchase, we included the most important stores determining the spatial trend of soda sales. Finally, our study estimated the volume of soda, rather than the actual quantity of sugar purchased, which is a more directly relevant public health indicator. Therefore, linkage of transaction data with existing nutrition composition databases should be performed in future.

In conclusion, the current lack of neighborhood-level dietary surveillance impedes effective public health and community actions aimed at encouraging healthy food selection and subsequent reductions of chronic illness. The rapidly increasing digitalization of consumer retail activities creates opportunities for creative public health applications of these data. Our method leverages existing grocery transaction data to address an important gap in population monitoring of nutrition status and food preferences.
References


Point of Care Research: Integrating patient-generated data into electronic health records for clinical trials

Sara Marceglia, PhD1,2, Pierluigi D’Antrassi, PhD2, Marco Prenassi, MSc1, Lorenzo Rossi, PhD3, Sergio Barbieri, MD2

1University of Trieste, Trieste, Italy; 2Fondazione IRCCS Ca’ Granda Ospedale Maggiore Policlinico, Milano, Italy; 3Newronika S.r.l., Milano, Italy;

Abstract

Integrating patient-generated data into clinical research would improve the reliability of results, especially when longitudinal, chronic, home-based monitoring is needed. To this end, we designed, implemented, and tested a system that allows integrating patient-generated data into electronic case report form (eCRF), using a standards based architecture that ensured the fulfillment of the major requirements for digital data in clinical studies. The system was tested in a clinical investigation for the optimization of deep brain stimulation therapy in patients with Parkinson’s disease that required both the collection of patient-generated data and of clinical and neurophysiological data. The validation showed that the implemented system was able to provide a reliable solution for including the patient as direct digital data source, ensuring reliability, integrity, security, attributability, and auditability of data.

Introduction

Point of Care Research (POC-R) is now gaining increasing attention as a way to enhance the integration between clinical research and clinical practice1 so that the FDA recently drafted a guidance on the “Use of Electronic Health Record Data in Clinical Investigations”2. Including electronic health records (EHRs) as data source for clinical trials would take advantage of the potentials of combining, aggregating, and analyzing longitudinal patients’ information, thus facilitating the collection of valuable information. It is expected that “EHRs may have the potential to provide clinical investigators and study personnel access to real-time and longitudinal health care data for review and can facilitate post-trial follow-up on patients to assess long-term safety and efficacy of medical products”2.

However, longitudinal clinical trials that require patient’s monitoring in a chronic, home-based condition may benefit of systems able to integrate patient-generated data into EHRs and, in turn, to electronic Case Report Forms (eCRFs). Personal mHealth Apps and the Internet of Things (IoT), at a time when there are more mobile connections than the entire human population, can support patients managing medical conditions, monitor lifestyles, and may even provide medical advice. Such technologies may also be adapted to become a new source of data, directly collected from patients in their ecologic environment, that can be used in clinical trials. To do so, patient-generated data in the chronic, home-based environment needs to be integrated with clinical hospital information systems (HISs) and EHRs.

As previously proposed3,4, a standards-based architecture could be used, independently from the specific EHR system considered, to allow the direct exchange between mHealth Apps and EHR systems. The standards-based architecture require specific document templates taking into consideration patient’s identification issues, privacy, reliability, and interoperability, and it was successfully proven in a prototype environment4.

However, integrating patient-generated data collected while the patient is at home into the streaming of clinical trial data poses some challenges, that reflect the “ALCOA” requirements (A – Attributable, L – Legible, C – Contemporaneous, O – Original, A – Accurate) stated by FDA guidance on electronic data sources in clinical trials:

1. Attributable: each document (and modification or review) has to be attributed to an author;
2. Legible: data should be clearly presented in a human-readable format;
3. Contemporaneous: in addition to the synchronous data recording, any modification should be time-stamped, tracked, motivated, and signed and then synchronized;
4. Original: data should be captured directly and kept safe and de-identified, because the mobile environment is supposed unsafe;
5. Accurate: patient-generated data may pose the problem of reliability, being the patient’s health literacy generally low. Also, data collected from personal medical devices should be considered. Finally, data have to be collected strictly following the clinical protocol, otherwise they cannot be used in the clinical investigation;

Considering these requirements, we designed and implemented a standards-based system that uses
(a) a specifically designed web-based platform that integrates EHR features with signal analysis and management;
(b) a workflow engine to manage the clinical trial process;
(c) a set of mobile apps and wearable devices for patient’s data collection that exchange information using a standards-based document set.

We tested the system in a clinical investigation for the optimization of deep brain stimulation (DBS) therapy in patients with Parkinson’s disease that required both the collection of patient-generated data and of clinical and neurophysiological data.

**Methods**

**System Architecture**

The system integrates information from different actors (Figure 1), namely the Patient, the Doctor, and the Researcher. The Patient may be supported by a Caregiver. The architecture has three main components: the Patient System, the Workflow Manager (WFM), and the WebBioBank (WBB).

![Figure 1. System Architecture](image)

**Patient System**

The Patient System consists of both hardware and software to ensure the provision and use at home of the information necessary for the proper implementation of the clinical trial. The hardware is composed of a Wearable Device (WD) and a Mobile Device (MD), while software refers to the mHealth app for the Patient and the Caregiver. The MD collects all the information attained by the WD (in our specific application, it consists of accelerometer acquisition) and the self-reported data (Patient Diary), and communicates them to the WFM via an external gateway. The main components of this device are a large touchscreen (>3.7”) for better filling out the Patient Diary, a storage system (larger than that made available by the WD) to temporarily store collected data before sending it to the WFM, a Bluetooth transceiver to communicate with and receive acquired data from the WD, and a long-life battery (at least 8h).
The connection from the Patient System to the WFM is implemented by the exchange of XML-based CDA-2 structured documents optimized for the anonymous communication with mobile applications\(^4\), so that no identification data is sent through unsecure connections, or retained into unsafe environments.

The WFM communicates with the Patient System by exposing services that manage reminders and notifications to the mobile application.

**Workflow Manager**

The WFM is a system that provides the interoperability and electronic exchange of information between the Patient System and the WBB. This system offers to the Researcher the ability to model the patient protocol as a graphical flowchart in a workflow editor and to the Doctor the ability to associate each Patient to the specific protocol and execute such process in a workflow engine.

A proprietary workflow composer is used for viewing and creating patient protocols. Key components of each protocol definition are participants (i.e., patients, caregivers, doctors, researchers and inspectors), activities (e.g., data acquisition, filling patient diary, drug administration, …), transitions (e.g., end data acquisition and send the CDA-2 document), schedules (e.g., filling the patient diary each 30 minutes), and variables.

For patient protocol execution, the WFM extends the Mirth Connect health care integration engine\(^6\). Mirth Connect is an open source HL7 interface engine is rich in features as well, including filtering, validation, transformations and routing of documents through different kinds of information system.

Users with different privileges and roles will gain access to the WFM according to their job function and to the clinical center they belong to, according to each specific research protocol. User access is managed and tracked in the WBB and synchronized with the WFM access control.

**WebBioBank**

WBB (www.webbiobank.com) is a web-based platform that integrates clinical data collection with signal management and processing\(^5\). The system is based on the wHospital framework\(^7\) that is a commercial EHR system used in several healthcare institutions and Regional implementations. WBB includes dedicated functionalities to support multicenter clinical studies and an adjunctive module for biosignal storage, management, and analysis\(^5\). The system is accessed through a standard web browser, allowing users to perform various tasks for data management in an anonymous mode according to shared protocols, thus guaranteeing real time interaction between researchers and clinicians in different research centers. Clinicians can add patient’s clinical information using shared clinical forms specific for the clinical study, whereas researchers develop and use shared advanced algorithms for signal processing that can be combined in analysis chains, ensuring that the data processing and analysis are the same in all of the centers involved in the research study. To extend its use to clinical trials, WBB provides data de-identification, ensuring patient’s privacy also when involved in multi-center studies. To enable anonymous data collection, on the web-based platform, only unique patients’ IDs (IDBAC) are saved and records do not include identifying demographic information. The unique IDBAC ensures attributability of records to patients. Patients demographics are visible only to the users belonging to the clinical center (operative unit, OU) in charge of the patient. Each OU can therefore use WBB as an EHR. WBB implements a Role-Based Access Control ensuring that only authorized users, with pre-defined roles, will gain access according to their job function in the clinical center during each specific research protocol. Also, the system keeps track of logins and failed login attempts. Since WBB integrates the functionalities of traditional EHRs with those of research support systems, it implements a “research” EHR (rEHR)\(^5\).

**Structured documents templates**

Data reliability, integrity, and attributability was achieved by exchanging standard CDA-2 documents. The CDA-2 template used is a modified version of the Personal Healthcare Monitoring Report (PHMR) CDA2 template (CDAR2_IG_PHMRPTS_R1.1_DSTU_2010OCT) adapted to fulfill the requirements of the mHealth App/EHR data exchange as already described in \(^4\). We will refer to this template as mobile PHMR (mPHMR) The major changes introduced in mPHMR regard the possibility to recognize the patient as data author (and not only as <recording target>), and the use of de-identified patient’s information, to ensure security. In this implementation, the experimental session ID is also reported in the header, to identify the precise location of the data collected within the clinical study protocol. The <Custodian> is the OU in charge of the patient. The mobile app and device are described as equipment and participate to the data collection as data source. The body of the document is composed by four mandatory sections: Results (LOINC 30954-2), Medical Equipment (LOINC 46264-8), Medications (LOINC 10160-0), and Purpose (LOINC 48764-5)\(^4\).
Validation case study

Our first target for system validation is the telemonitoring of patients with DBS implant in order to obtain information on therapy optimization. Since our patients face a fragile stabilization period immediately after electrode placement, we decided to make the first system testing in a controlled environment, to keep the risks for patients as low as possible.

We hence used the implemented system to support a clinical trial ongoing at the Fondazione IRCCS Ca’ Granda Ospedale Maggiore Policlinico of Milan (Italy) on patients with Parkinson’s Disease implanted with DBS electrodes. The specific research aimed to test a newly developed CE-marked external prototype for closed-loop adaptive DBS (aDBS). The aDBS device automatically adapts stimulation parameters moment-by-moment to the patient’s clinical state, thus providing optimal stimulation parameters\(^8\). This external prototype we tested implements an aDBS approach based on the analysis of local neuronal activity (local field potentials, LFPs) captured from the DBS electrodes while stimulation is switched on\(^8\). In particular, the prototype adapts DBS parameters according using modulations of the LFP beta band (13-35 Hz). In this research study, aDBS was tested in freely moving PD patients for 8 consecutive hours, by continuously monitoring beta band modulations and by assessing aDBS ability to respond to changes in patient’s state (Figure 2). The data collection system we implemented was used to support the research in two ways: (1) collecting clinical data generated by the expert neurologist who assessed the patient during the 8-hours period; (2) collecting patient-generated data in an ecologic but controlled environment, while the patient was left alone doing his/her normal activities in his/her room with the aDBS prototype on.

Figure 2. 8-hours experimental protocol

The experimental protocol is depicted in Figure 2. The study was approved by the institutional review board and by the Italian Ministry of Health, and conformed with the Declaration of Helsinki. We enrolled 4 PD patients who underwent surgery for DBS electrode implant having externalized leads for a week to connect a wearable device for aDBS testing. Two perioperative sessions lasting from 7 to 8 hours were conducted the day 5 and 6 after surgery while the patient was doing his/her normal activities. In the first session, the beta band power was continuously recorded while the patient took its post-operative daily medication dose. In the second session, we added aDBS treatment to beta band power monitoring and daily medication. A neurologist assessed the clinical state and fluctuations at 5 time points (Figure 2) through the motor Unified Parkinson’s Disease Rating Scale part III (UPDRSIII) and the Unified Dyskinesia Rating Scale (UDysRS). During the entire duration of the experimental session, the patient filled in a diary (every 30 minutes) and wore a bracelet to assess his/her bradykinesia state.
Patient’s wearable device and mobile application

Patient’s generated data in the implemented system were of two kinds, the first one being a personal diary, collecting the patient’s state every 30 minutes, with scheduled alerts; the second one being accelerometer data acquired by a commercially available wearable device.

The device used was a Pebble Time smart-watch. It includes a three-axis accelerometer and a Bluetooth connection with a mobile device (Android Phone). A dedicated app was developed to both acquire and store data, and to provide a clinical diary to be filled in by the patient at predefined times. The patient has a personal ID and password to access to the app (enforced by an initial login screen). The mobile phone app implements an algorithm that finds the Bradykinesia Acceleration Scores (BAS) using the accelerometric values of the smart bracelet. The BAS algorithm was adapted from the patent of Griffiths & Horne9.

System testing

The system Validation test consisted into two experiments: the first one checked the functionalities of the WBB system and its integration with the WFM, and the second one verified the ability of the app to collect patient-generated data in an accurate and secure way. Tables 1 and 2 (columns 1 and 2) detail the test protocols of the two experiments and their expected results. Errors are detected when the test result differs from the expected result.

Results

System implementation

The system was fully implemented to be used in the validation case study. WBB was configured in terms of users, roles, and forms to support the 8-hours research study in the hospital-based ecologic but controlled environment. WBB fulfills the requirements for clinical study data collection: patients are de-identified to ensure security; the clinical forms are developed by the researcher (usually the principal investigator), filled in and signed by the author, reviewed by the principal investigator, and changes/modifications are tracked, time-stamped, and signed whenever they occur, thus guaranteeing integrity, attributability, and reliability. The definition of the “inspector” user role allowed auditability. Figure 3A shows the snapshots of the main rEHR forms created to collect data from PD patients undergoing the experiment.

Table 1. Validation tests for the WBB and WFM systems. The ALCOA cell line (column “Results”) represents the verification of requirements fulfillment with the results obtained by the tested action (green = achieved).

<table>
<thead>
<tr>
<th>Experiment 1</th>
<th>Expected result</th>
<th>Result and requirements fulfillment</th>
</tr>
</thead>
<tbody>
<tr>
<td>The doctor creates a new de-identified EHR in WBB and fills in the clinical form for the research study (UPDRSIII, UDysRS).</td>
<td>The clinical form is filled in, connected to the IDBAC, and signed</td>
<td>No errors in clinical forms and signatures.</td>
</tr>
<tr>
<td>The doctor changes the content of an existing clinical form</td>
<td>Modifications are tracked, time-stamped, and signed</td>
<td>All clinical forms show the modifications with the time stamp and the signature</td>
</tr>
<tr>
<td>The researcher logs in the WFM to create the patient protocol and the doctor logs in the WFM to assign the patient to the protocol</td>
<td>The doctor is not allowed to create a new protocol and the researcher is not allowed to assign the patient to a protocol</td>
<td>Each user can access to the WFM and can manage protocol actions according to the user’s role</td>
</tr>
</tbody>
</table>
Table 2. Validation tests for the Patient System. The ALCOA cell line (column “Results”) represents the verification of requirements fulfillment with the results obtained by the tested action (green = achieved).

<table>
<thead>
<tr>
<th>Experiment 2</th>
<th>Expected results</th>
<th>Result and requirements fulfillment</th>
</tr>
</thead>
<tbody>
<tr>
<td>1) A caregiver logs into the app with the patient ID and password, configures the timer of the app to acquire only accelerometric data from the Pebble Watch in an 8h session and press play to start the acquisition from the mobile phone.</td>
<td>At the end of the session, there will be a rEHR inside the WBB with the BAS data of the patient and another rEHR with the UPDRSIII scores compiled by the clinician. The mPHMR data on the mobile device must be absent at this time and the tables on the SQLite Database must be empty. The IDBAC of the two rEHR and the IDBAC into the mobile device must be the same.</td>
<td>There are two correct rEHR on the WBB. Some BAS data in particular timeframes are absent due to the distance of the mobile device from the patient caused by forgetfulness or special conditions (e.g. MRI or other exams). The correlation between UPDRSIII scores and the BAS is -0.582 (p-value&lt;0.009, Pearson) for 5 sessions regarding 4 different patients.</td>
</tr>
<tr>
<td>2) The patient wears the bracelet helped by the caregiver and starts the acquisition pressing the start button on the watch, at the end of the session the caregiver synchronizes the mobile device pressing “Send data” with the Workflow Manager.</td>
<td>At the end of the session, there will be a rEHR inside the WBB with the diary data of the patient. The mPHMR data on the mobile device must be absent at this time and the tables on the SQLite Database must be empty. The IDBAC of the rEHR and the IDBAC of the patient must be the same.</td>
<td>There is a correct rEHR on the WBB; the BAS score was not evaluated. One patient did not fill the diary; the rEHR data contains null values for every missing timeframe.</td>
</tr>
<tr>
<td>3) A doctor logs into the WBB and fills the clinical form of the patient with the UPDRSIII scores assessed during the session.</td>
<td>A caregiver logs into the app with the patient ID and secret password, configures the diary notifications interval of the app to 30 minutes and sets the session time to 8 hours to acquire the personal diary data. The patient, after the setup, fills the diary every half an hour, at the end of the session the caregiver synchronizes the mobile device with the Workflow Manager pressing “Send Data” on the mobile app.</td>
<td>Testing the range of the Bluetooth connection between mobile device and smartwatch. It must cover all the patient room including the bathroom and the nearby corridor (at least 8 meters with a wall in between, not RF shielded). Bluetooth Low Energy has &gt;100 m declared line-on-sight range with the on-chip antenna, in normal patient rooms the RF noise on the 2.4 GHz band is low and the walls are not RF shielded.</td>
</tr>
<tr>
<td>A caregiver logs into the app with the patient ID and password, configures the diary notifications interval of the app to 30 minutes and sets the session time to 8 hours to acquire the personal diary data. The patient, after the setup, fills the diary every half an hour, at the end of the session the caregiver synchronizes the mobile device with the Workflow Manager pressing “Send Data” on the mobile app.</td>
<td>Testing the battery life of both the pebble smartwatch and the mobile devices. The prerequisite states that the WiFi and Bluetooth connection must be always on, also the accelerometric acquisition must be always on in a daily living situation, for at least 8 hours.</td>
<td>The Pebble Smartwatch app acquires data with a sample rate of 100 Hz and sends packets every 5 seconds to save battery. The background service on the mobile device must be always on. It is a power consuming activity but it is expected at least 8 hour of operation.</td>
</tr>
</tbody>
</table>
Two web services were developed to support the communication and integration between the WBB and the WFM. The first one (wHvpc) is devoted to the integration of user roles and patients’ IDs: it allows the verification of the privacy criteria for doctors/researchers who access the WFM to create or assign the patient’s protocol and, once the doctor accesses to assign the study protocol, allows the exchange of patient’s ID and contact information from WBB to WFM. Then, when the protocol has been assigned and the patient is registered in WFM, the WFM deletes contact information and keeps only the patient’s ID. In this way, the synchronization between the two systems is guaranteed thus allowing attributability and integrity, and patient’s contact information do not reside on WFM, thus allowing security and privacy. The second one (wHcda) is devoted to the exchange of CDA-2 standard documents between the WFM and the WBB. Once the WFM receives patient-generated data from the mobile application, it creates and encodes the CDA-2 document according to the mPHMR template, and sends it to the WBB using the wHcda web service. Data reliability is therefore ensured by the use of standard documents that are accepted by the WBB platform and processed as clinical documents.

On the patient side, the WFM provides the functionalities for patient’s registration and protocol fulfillment. The WFM stores the patient data into the correct rEHR on the WBB by calling the wHcda service passing the identification numbers of the user (uID), the patient (IDBAC), and the custodian (ID OU). The patient data inside the database are anonymous for all users, and only the patient’s doctor can re-identify them by means of a local registry. De-identification is guaranteed also by the WFM registration process that does not require patient’s demographic information, but uses the contact information retrieved at the time of assignment that are then deleted when the patient is successfully registered. Thanks to the definition of the patient’s protocol, the WFM is able to send the activity program to the patient’s mobile app, and to provide remainders and alerts when a task is due (e.g., when the patient has to fill-in the diary). Also, in the case WFM does not receive the patient-generated data on time, according to the protocol, it sends new requests, and then notifies the WBB of the deviation from the protocol, sending a specific CDA-2 with the indication of the deviation using the wHcda web service.

Patient-generated data collection for the case study, including a patient’s diary to be filled in and a wearable bracelet for bradykinesia evaluation (Figure 3B), is shown in the sequence diagrams in Figure 4. The algorithm in the mobile app generates a BAS value every 4 minutes of data. BAS is lower when the patient is bradykinetic (Figure 3C). Bracelet data are preprocessed in the mobile app to retrieve the bradykinesia score (Figure 3C). The forms for the patient’s diary consists of a multiple-choice mutually exclusive list that asks the perceived motor status through 4 different answers: “OFF: Bradykinesia and rigidity”, “ST: Transition”, “ON: normal mobility”, and “ON: disabling dyskinesia”.

The diary and the accelerometer data is stored internally in a SQLite DBMS. When the device is synchronized with the WFM, a mPHMR document is generated with the all the BAS data compressed and encoded in MIME format in an observation of the CDA-2 document (content-type: application/x-compressed, Content-Transfer-Encoding: BASE64) and all the diary data on another plain text observation of the same document. If the mPHMR document is correctly stored and approved by the WFM, a positive feedback is sent back to the mobile device and the internal database is erased for security reason. This feedback and the other remainders are sent through a web-service exposed by the WFM. The use of standard CDA-2 documents between the App and the WFM ensures data integrity, attributability, and safety (in the CDA-2 the author is the patient, identified only by his/her ID). Also, the mobile app does not retrieve any clinical data, but deletes them when the WFM correctly receives the CDA-2 and validates the data.

System validation

In our experiment, we enrolled 4 patients, 2 doctors, and 1 researcher (the principal investigator). Also, a user with role “inspector” was created to test the audit functionalities. On WBB, the researcher created 7 forms, one for patient’s disease history, one for DBS surgery details (target position, electrodes implanted, intraoperative monitoring results), one for the details of the experimental setting (levodopa equivalent dose administered to the patient, neurophysiologic parameters retrieved to set the aDBS device), one for each clinical assessment including the UPDRSIII and the UDysRS scale, and one for the visualization of patient-generated data. The results of the planned validation testing are reported in Tables 1 and 2 (column 3).
Figure 3. System implementation for the validation case study. A) Snapshot of the WBB form for UPDRS III. B) Patient assessment and patient diary. C) Accelerometer exemplary data. D) Correlation between UPDRS III score and bradykinesia accelerometer score (BAS). Note that the system was implemented in Italian.

All the expected 128 diary recordings were received by the system. Of them, 105 were filled-in whereas 23 arrived with null values. The major reason was one poor compliant patient who lacked compiling the diary several times. There were no errors in the data transmission. A total of 40 hours of accelerometer data were recorded. The accelerometer data loss was due to a poor connection between the wearable device and the mobile app. Despite this, as shown in Figure 3D, there was a significant correlation (-0.582, p<0.009) between clinical assessments and patient-generated data, thus suggesting that the measures obtained by the wearable device are reliable for assessment purposes, even though data are incomplete.

Conclusion

We designed, implemented, and tested in a real clinical research study a system that allows integrating patient-generated data regarding clinical trial into rEHRs, using a standards based architecture that ensured the fulfillment of the major requirements.

Health information exchange between mHealth Apps (Patient Reported Outcomes, PRO), clinical research studies, and EHR systems is a recent challenge from the technological and regulatory point of view. Many example bi-directional health information exchange between mHealth Apps and EHR systems are available but they all differ from the architecture described in the present work, because they were all developed for the hospital environment and for the use by healthcare professionals. Other solutions provide the integration between clinical research studies and PRO without EHR support. Our web based solution provides the integration of all the three elements: PRO, EHR and clinical trials.
Moreover, the presented system architecture could be used in different healthcare domains. For instance, a similar architecture, still in a prototype stage, was developed to support longitudinal studies in nutrigenomics\textsuperscript{14}, including the management of behavioral and nutritional habits of patients with cardiovascular risk that was guaranteed by the integration of the patient reported meals, his/her EHR, and food dictionary.

It has been proposed that interoperability issues in health IT can be addressed by using web services\textsuperscript{15,16}. Our architecture, in agreement with this hypothesis, implements specific web services to ensure interoperability among the different systems (WebBioBank, Workflow manager, and mHealth App) and to enable the communication among patients/caregiver, researchers and doctors by using standards such as PHMR template compliant with RIM (CDA2) and dictionaries.

Considering that DBS patients after surgery for electrode implant face a fragile stabilization period, we expected poor compliance and increased risk. For this reason, we decided to make the first assessment of the mHealth App in a controlled hospital environment. However, the system requires, in the next future, a more focused testing procedure with the patients in their home environment to further verify usability and robustness in longitudinal studies.

The loss of data during the evaluation stage was mainly caused by the low available data storage memory in the bracelet, thus requiring a continuous connection with the mobile phone to transfer data and free the memory. This study shows how the capability of a device to collect data autonomously is critical during the daily living of these fragile patients who are not used to keep the mobile phone close to them but tend to wander off outside its range. This capability will be a major requirement for future implementations of the proposed architecture.

The validation showed that the implemented system and architecture were able to provide a reliable solution for including the patient as direct digital data source, ensuring ALCOA requirements for data generated either by the patient him/herself or by personal wearable device.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{sequence_diagram.png}
\caption{Sequence diagram of Patient-generated data collection. A) Accelerometric data acquisition. B) Diary generation}
\end{figure}
References

Fast and Accurate Metadata Authoring Using Ontology-Based Recommendations

Marcos Martínez-Romero, PhD, Martin J. O’Connor, MSc, Ravi D. Shankar, MS, Maryam Panahiazar, PhD, Debra Willrett, MS, Attila L. Egyedi, Olivier Gevaert, PhD, John Graybeal, and Mark A. Musen, MD, PhD
Center for Biomedical Informatics Research, Stanford University, Stanford, CA, USA

Abstract

In biomedicine, high-quality metadata are crucial for finding experimental datasets, for understanding how experiments were performed, and for reproducing those experiments. Despite the recent focus on metadata, the quality of metadata available in public repositories continues to be extremely poor. A key difficulty is that the typical metadata acquisition process is time-consuming and error prone, with weak or nonexistent support for linking metadata to ontologies. There is a pressing need for methods and tools to speed up the metadata acquisition process and to increase the quality of metadata that are entered. In this paper, we describe a methodology and set of associated tools that we developed to address this challenge. A core component of this approach is a value recommendation framework that uses analysis of previously entered metadata and ontology-based metadata specifications to help users rapidly and accurately enter their metadata. We performed an initial evaluation of this approach using metadata from a public metadata repository.

Introduction

Reproducibility of biomedical discoveries has become a major challenge in science. Investigations in a variety of fields have shown alarmingly high levels of failure when attempting to reproduce published studies.¹,² To help address this issue, many funding agencies and journals are now demanding that experimental data be made publicly available—and that those data have associated descriptive metadata.³ In the last few years, the biomedical community has met this challenge by driving the development of metadata standards, which scientists use to inform their annotation of experimental results. For example, the MIAME standard⁴ describes metadata about microarray experiments. The overarching goal when defining these standards is to provide sufficient metadata about experimental data to allow the described experiment to be reproduced. Community-based groups have defined an array of standards describing metadata for a variety of scientific experiment types. A large number of standards-conforming repositories have been built, greatly enhancing the ability of scientists to discover scientific knowledge.⁵

Despite the increasing use of these standards, the quality of metadata deposited in public metadata repositories is often very low.⁶ A central problem is that metadata authoring process itself can be extremely onerous for scientists.⁷ A typical submission requires spreadsheet-based entry of metadata—with metadata frequently spread over multiple spreadsheets—followed by manual assembly of multiple spreadsheets and raw data files into an overall submission package. Validation is often post-submission and weak. A secondary problem is that metadata standards are typically written at a high level of abstraction. For example, while a standard may require capturing the organism associated to a biological sample, it typically will not specify how the value of the organism must be supplied. Little use is made of the large number of controlled terminologies currently available in biomedicine. Submission repositories reflect this lack of precision and usually have weak or nonexistent mechanisms for linking terms from controlled terminologies to submissions. Faced with this lack of standardization, users often provide ad hoc values or simply omit many values. These difficulties combine to ensure that typical metadata submissions are sparsely populated and poorly described, and thus require significant post-processing to extract semantically useful content.

In this paper, we describe the development of a methodology and associated tools that aim to improve the metadata acquisition process. We outline a recommender framework that provides an intuitive and principled approach to metadata entry. The framework uses analyses of previously entered metadata combined with ontology-based metadata specifications to help guide users to rapidly and accurately enter their metadata. The recommender framework is part of the CEDAR Workbench (https://cedar.metadatcenter.net), an end-to-end metadata acquisition and management system under development by the Center for Expanded Data Annotation and Retrieval (CEDAR).⁸ The ultimate goal is to speed up the creation of metadata submitted to public repositories and to increase the quality of that metadata.
Related Work

Browser-based auto-fill and auto-complete functionality has a long history on the Web. Common auto-fill examples include the automatic population of address and payment fields by Web browsers in standard HTML forms. Auto-complete suggestions are commonly made for page URLs, where browsers typically maintain a history of visited pages and suggest likely pages based on a simple frequency analysis of previously visited pages. More advanced auto-complete functionality can be seen in search engines from major Web search vendors, where suggestions are based on analyses of both Web content and searches made by users.

A variety of auto-fill and auto-complete recommendation systems have been developed that perform more substantial analyses of previously entered content. A common approach is to process raw form content to extract high-level semantic concepts that drive the recommendation process. A system called Carbon⁹ presents auto-complete suggestions based on an analysis of Web forms previously filled in, combined with semantic information from those forms. The system uses this information to help users fill in structurally different forms. A related system called iForm¹⁰ was developed to assist form completion by analyzing both previously filled versions of a form and free text to extract likely values for fields. The approach focused on performing a semantic analysis of data-rich input text to automatically select text segments and then associating the text segments with fields in a form.

Several recommender systems that support auto-fill and auto-complete with ontology terms have been described in the literature. RightField,¹¹ which is distributed as an Excel plugin, provides mechanisms for embedding ontology-based value fields in spreadsheets. Users populating the resulting spreadsheets are presented in real time with suggestions restricted to terms from subsets of specified controlled vocabularies. Ontology-based systems that specifically address the metadata acquisition challenge include Annotare¹², which is used to submit experimental data to the ArrayExpress metadata repository,¹³ and ISA-Tools,¹⁴ which provides a generic spreadsheet-based tool chain for metadata authoring. Both systems provide strong support for using controlled terms and allow users to link metadata to controlled terminologies. None of them provides value-recommendation functionality, however.

By combining the analysis-driven and ontology-based recommendation strategies used by these systems, we can generate more powerful suggestions than is possible with each approach alone. We believe the combination of the two techniques can provide the speed of analysis-driven recommendations coupled with the added precision of ontology-based suggestions. This paper advances our preliminary work on metadata prediction¹⁵,¹⁶ by outlining the development of a methodology and associated tools that demonstrate this combination.

Methods

We designed an approach for metadata recommendation that simplifies the metadata authoring process in the CEDAR Workbench. The CEDAR Workbench is a suite of Web-based tools and REST APIs for metadata authoring and management, centered on the use of metadata-acquisition forms called metadata templates (or simply templates). In the CEDAR Workbench, templates are used to formally encode metadata standards and to create highly-interactive interfaces for acquiring metadata conforming to those standards. Templates define the data attributes (called template fields or fields) needed to describe experimental data. For example, an experiment template may have a disease field containing the name of the disease studied by a particular experiment. Our approach simplifies metadata authoring by suggesting the most appropriate values for template fields when acquiring metadata. We outline our approach and then explain how we implemented it in the CEDAR Workbench.

Description of the approach

Let \( t \) be a metadata template, which contains a set of template fields \( f_1...f_n \). Now suppose that a user is filling out the template \( t \) with metadata. Our approach generates a ranked list of suggested values for fields \( f_1...f_n \) based on: (1) the template instances previously authored for the template; and (2) the field values already entered by the user for the current template, which we call the recommendation context.

For a template field being filled out, our approach retrieves all values previously entered into that field and calculates a relevancy score in the interval \([0,1]\). This score represents the likelihood of the value occurring again based on previously created template instances and on the recommendation context. The relevancy score for a field-value pair \( p \) in a template instance \( s \), derived from a template \( t \), is calculated as:

\[
\text{score}(p, s) = \frac{|\text{matchingInstances}(W)|}{|\text{matchingInstances}(fv(s))|}, \text{with } W = \{p\} \cup \text{fv}(s)
\]
where instances(t) are all previously authored instances of the template t, and matchingInstances(A) returns the instances that contain all the field-value pairs in A. The matchingInstances(A) function is defined as:

\[ \text{matchingInstances}(A) = \{ x \in \text{instances}(t) \mid \forall a \in A, a \in f(v(x)) \} \]

Here, \( f(v(s) \) represents all the field-value pairs for an instance s:

\[ f(v) = \{ (f, v) \mid f \in \text{fields}(s) \land v \in \text{values}(s) \land \text{value}(f) = v \} \]

where fields(s) are the fields in the template t, from which s is derived, values(s) are all the values in the instance s, and value(f) is the value assigned to a field f in the instance s. While the instance is being created \( f(v(s) \) represents the recommendation context.

Example: Suppose we have a template t with disease and tissue fields and have four instances of t with values as shown in Table 1.

Table 1. Field names and values for four sample instances of a template with disease and tissue fields.

<table>
<thead>
<tr>
<th>Field</th>
<th>instance 1</th>
<th>instance 2</th>
<th>instance 3</th>
<th>instance 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disease</td>
<td>liver cirrhosis</td>
<td>liver cirrhosis</td>
<td>liver cirrhosis</td>
<td>breast cancer</td>
</tr>
<tr>
<td>Tissue</td>
<td>liver</td>
<td>liver</td>
<td>blood</td>
<td>Breast</td>
</tr>
</tbody>
</table>

Suppose that the user is populating a new instance of t and has entered liver cirrhosis as a value for the disease field. The relevancy scores for the values liver, blood, and breast of the tissue field can be calculated as follows:

\[
\begin{align*}
\text{score}((\text{tissue}, \text{liver}), s) &= \frac{|\text{matchingInstances}(((\text{tissue}, \text{liver}), (\text{disease}, \text{liver cirrhosis})))|}{|\text{matchingInstances}(((\text{disease}, \text{liver cirrhosis})))|} = \frac{2}{3} = 0.67 \\
\text{score}((\text{tissue}, \text{blood}), s) &= \frac{|\text{matchingInstances}(((\text{tissue}, \text{blood}), (\text{disease}, \text{liver cirrhosis})))|}{|\text{matchingInstances}(((\text{disease}, \text{liver cirrhosis})))|} = \frac{1}{3} = 0.33 \\
\text{score}((\text{tissue}, \text{breast}), s) &= \frac{|\text{matchingInstances}(((\text{tissue}, \text{breast}), (\text{disease}, \text{liver cirrhosis})))|}{|\text{matchingInstances}(((\text{disease}, \text{liver cirrhosis})))|} = \frac{0}{3} = 0
\end{align*}
\]

Our method takes advantage of previously populated fields to generate a context-sensitive estimate of the values for an unpopulated field. When there is no context (i.e., when no other fields in a template have been filled out), it simply computes the frequencies of all the values found for the unpopulated field and ranks the values accordingly. The method does not impose any restriction on the order in which the fields must be filled out. The analysis for a field can be performed on all instances in the repository and using all the values previously entered for other fields, independently of the order they were filled out. This ability of the method to consider contextual information enables it to go beyond simple one-cause, one-effect relationships and to consider the combined effects that previously entered values may have on the target field.

The basic approach can be extended to deal with fields whose values have been constrained to particular ontologies, ontology branches, or lists of ontology terms (e.g., a disease field could be constrained to contain diseases from the Disease Ontology). When dealing with these ontology-based field values, our method calculates the frequencies of the underlying term identifiers independently of the display value used. For example, suppose that the repository contains template instances that refer to hypertension in different ways, such as HTN, increased blood pressure, and high blood pressure, and that those instances have been linked to the identifier of the term hypertension in the Disease Ontology (http://purl.obolibrary.org/obo/DOID_10763). In this case, our analysis approach would use the term identifier to calculate the frequency of the Disease Ontology term, effectively aggregating the frequencies of all synonyms of hypertension.

The generated recommendation scores can be used to produce a ranked list of suggested values for a target template field. Each recommendation consists of the suggested value and a number in the interval [0,1] that represents the frequency of the value in previously populated instances. For plain text metadata, the system suggests textual values. For ontology-based metadata, the system suggests ontology term identifiers. These recommendations can be
presented to the user using a user-friendly preferred label for the ontology term defined in its source ontology (e.g., hypertension is the preferred label for http://purl.obolibrary.org/obo/DOID_10763 in the Disease Ontology). The recommendations for a particular field can be calculated in real time as a template is being filled in. The metadata repository’s recommendation index can be updated whenever a template instance is saved, allowing other metadata instance creators to immediately use the updated recommendation values.

Implementation

We implemented our approach for metadata recommendation as a Web service called the Value Recommender, and integrated it into the CEDAR Workbench. The CEDAR Workbench provides two core tools that form a metadata authoring pipeline: the Template Designer and the Metadata Editor. The Template Designer allows users to interactively create metadata templates in much the same way as they would create survey forms. Using live lookup to BioPortal, the Template Designer allows template authors to find terms in ontologies to annotate their templates, and to constrain the values of template fields to specific ontology terms18. The Metadata Editor uses a template specification generated by the Template Designer to automatically generate a forms-based metadata acquisition interface for that template. The generated interfaces allow users to populate metadata templates with metadata.

We modified both tools and several other CEDAR components to work with the Value Recommender service (see Figure 1). We extended the Template Designer to allow users to specify the fields for which value recommendations are enabled. We enhanced CEDAR’s template specification model to store this preference. This preference is used to signal to CEDAR’s metadata indexing engine that field-level metadata in its metadata repository should have additional analysis steps applied to it. Fields marked for value recommendation are indexed by CEDAR’s Elasticsearch-based engine (https://www.elastic.co) such that their values can be compared with other value recommended fields. These statistics are used in real time by the newly developed Value Recommender component. Note that users can also use standard CEDAR functionality to constrain fields to contain values from controlled terminologies held in the BioPortal server. Both constraint types can be specified simultaneously for a field.

We extended the Metadata Editor to use the Value Recommender service to suggest appropriate values for metadata fields during field entry (see Figure 2). Users entering metadata using the Metadata Editor are prompted in real time.
with drop-down lists, auto-completion suggestions, and verification hints supplied by the Value Recommender service. Recommendations for unfilled fields are updated in real time as users incrementally complete metadata acquisition forms. The editor presents a drop-down list for value-recommended fields containing suggested values ranked in order of likelihood. The editor can also be configured to indicate whether suggested values are ontology terms, in which case it also shows the BioPortal acronym for that ontology.

Figure 2. Screen shot of the CEDAR Metadata Editor showing recommended values for a particular field. In this case the editor shows suggestions for disease values. It presents a drop-down list containing suggested values ranked in order of likelihood. Ontology-based terms are indicated with an ontology icon. The relevancy score for each suggested value is presented as a percentage.

Evaluation

We analyzed the performance of our framework when suggesting appropriate metadata values using both plain text metadata and metadata represented using ontology terms. We constructed an evaluation pipeline to drive the analysis workflow (see Figure 3). The main steps of our evaluation workflow are as follows.

1. Preprocessing and ingestion

We used the CEDAR Workbench to design a metadata template targeted to the BioSample metadata repository. This repository, which is provided by the National Center for Biotechnology Information (NCBI), captures descriptive information about biological materials used in scientific experiments. BioSample defines several packages that represent specific types of biological samples, and specifies the list of attributes by which each sample should be described. The BioSample Human package, for example, is designed to capture metadata from studies involving human subjects, and includes attributes such as tissue, disease, age, and treatment. We used this package specification to develop a BioSample template in CEDAR to describe human samples.

For the purpose of our evaluation, we populated BioSample template instances using metadata from the Gene Expression Omnibus (GEO), a database of gene expression data which contains experimental metadata largely authored by original data submitters. The GEO database currently contains over 2 million records, and includes over 80,000 studies, each of which contains metadata for related biological samples.

We downloaded metadata from the GEO repository using GEOmetadb, and extracted all corresponding metadata elements for all human samples. We chose the fields title, sample_id, series_id, status, submission_date, last_update, type, sources_name, organism, and characteristics (including disease and tissue). Then, we picked the human samples that contained both disease and tissue metadata (35,157 samples), and transformed them into BioSample template instances conforming to CEDAR's JSON-based model.
Figure 3. Evaluation workflow. (1) Design a template for the BioSample repository, and populate it with metadata from the Gene Expression Omnibus (GEO); (2) Annotate the template instances obtained with terms from biomedical ontologies; (3) Upload the training set to the CEDAR Workbench; (4) For each of the test instances, generate suggestions for the disease, sex, and tissue fields; (5) Compare the suggestions obtained using the Value Recommender with the suggestions obtained using the baseline method.

2. Semantic annotation

We define semantic annotation (or simply annotation) as the process of finding a correspondence or relationship between a term in plain text and an ontology term that specifies the semantics of the term in plain text. We used a component of the CEDAR system, called the Semantic Annotation Pipeline (SAP), to automatically annotate all the fields values in 35,157 BioSample instances using biomedical ontologies. Table 2 shows the BioSample fields used in our evaluation. It presents both the number of plain text values and the number of ontology terms resulting from the semantic annotation process. The annotation ratio represents the mean number of plain text values per ontology term. For instance, we observed that the concept female was represented in plain text using values such as female, Female, f, F, and FEMALE.

Table 2. Comparison between the number of plain text values for the fields disease, sex, and tissue, and the number of ontology terms resulting from applying our Semantic Annotation Pipeline (SAP) to the plain text values.

<table>
<thead>
<tr>
<th>Field</th>
<th>Description</th>
<th>Plain text</th>
<th>Ontology terms</th>
<th>Annotation ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Values</td>
<td>Examples</td>
<td>Values</td>
</tr>
<tr>
<td>disease</td>
<td>Disease diagnosed</td>
<td>1,064</td>
<td>Lung carcinoma, carcinoma of lung</td>
<td>261</td>
</tr>
<tr>
<td>sex</td>
<td>Sex of sampled organism</td>
<td>16</td>
<td>female, Female, f, F, FEMALE</td>
<td>2</td>
</tr>
<tr>
<td>tissue</td>
<td>Type of tissue the sample was taken from</td>
<td>604</td>
<td>liver, Liver, liver tissue, liver biopsy, Liver biopsy tissue</td>
<td>171</td>
</tr>
</tbody>
</table>
3. Training

We partitioned the sample data—both for plain text values and for values annotated with ontology terms—into two sets. We used 80% (28,126 instances) of the sample data for training and 20% (7,031 instances) of the data for testing. We uploaded the training set to the CEDAR Workbench using the CEDAR API. We then indexed the training set with Elasticsearch.

4. Testing

For each of the test instances, we used the Value Recommender to generate value recommendations for the disease, sex, and tissue fields. The Value Recommender suggested values for each field based on the values of the other fields (e.g., suggestions for tissue were generated using the values for disease and sex). We used the majority vote as our baseline, which means picking the value with more occurrences in the training data. This process differs from the Value Recommender in that it ignores co-occurring values. For each field, we compared the suggestions provided by both the Value Recommender and the baseline with the expected value. We considered that the expected value for a field was the value for the field contained in the instance.

The Value Recommender produces a list of suggested values ranked by relevance. We assessed the performance of our method using the mean reciprocal rank (MRR) statistic. This statistic is commonly used for evaluating processes that produce a list of possible responses to a query, ordered by probability of correctness. We limited the output to the top three recommendations, and calculated the reciprocal rank (RR) as the multiplicative inverse of the rank of the first correct recommendation.

For example, if the correct value were ranked in the 3rd position, the reciprocal rank would be calculated as 1/3. Then, MRR was calculated as the mean of all RRs obtained. Table 3 shows the RR statistic for some example recommendation results. The MRR of the three values suggested in the table can be calculated as the mean of the RR value for all three rows: $(1 + 1/2 + 1/3) / 3 = 0.61$.

Table 3. Example of recommended values and reciprocal rank (RR) for the disease field.

<table>
<thead>
<tr>
<th>Expected value</th>
<th>Recommended values</th>
<th>RR</th>
</tr>
</thead>
<tbody>
<tr>
<td>asthma</td>
<td>1) asthma</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2) lung cancer</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3) lipid metabolism disorder</td>
<td>1</td>
</tr>
<tr>
<td>lymphoma</td>
<td>1) rheumatoid arthritis</td>
<td>1/2</td>
</tr>
<tr>
<td></td>
<td>2) lymphoma</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3) acute myeloid leukemia</td>
<td></td>
</tr>
<tr>
<td>lung cancer</td>
<td>1) rheumatoid arthritis</td>
<td>1/3</td>
</tr>
<tr>
<td></td>
<td>2) asthma</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3) lung cancer</td>
<td></td>
</tr>
</tbody>
</table>

5. Analysis of results

Figure 4 compares the mean reciprocal ranks obtained using our recommendation framework with the majority value baseline. It shows results both for plain text values (left) and values annotated with ontology terms (right).

The results indicate that our framework performs considerably better than the baseline for the three fields. Our context-sensitive recommendation method obtained an average MRR of 0.78 for plain text values, and 0.77 for ontology terms, compared to the baseline method’s average MRR of 0.21 for plain text values and 0.41 for ontology terms. By examining the results we see that, for example, our method correctly suggested asthma as a value for the disease field when the tissue was epithelium of bronchus. However, the baseline method suggested hepatocellular carcinoma, a disease that is not related to that kind of tissue. Our approach performs consistently well both for plain text values (0.78) and for ontology-based metadata (0.77). The importance of contextual information is particularly evident when analyzing the results obtained for plain text values, where there are substantially more values for each field than for ontology terms (for example, as shown in Table 2, there are 1,064 plain values versus 261 ontology
terms for the disease field). The average MRRs of the baseline are considerably lower for plain text values (0.21) than for ontology terms (0.41).

![Figure 4](image)

**Figure 4.** Mean reciprocal rank for BioSample instances with plain text values (left) and with ontology-terms (right), both for the baseline and the Value Recommender (VR).

Finally, we investigated more closely the effect of the context on the recommendations. The best results were obtained for the tissue field, with MRRs of 0.88 for plain text and 0.86 for ontology terms, illustrating the strong influence of the context on that field. Once disease and sex field values are provided, our approach is able to identify the appropriate value for the tissue field in most cases. Figure 5 shows the top suggestions provided by the Value Recommender for the disease field, with increasing levels of context. The figure shows how lung cancer can be much more clearly suggested as a likely choice when sex and tissue have been specified. This example reflects the increase in discrimination that is possible when more contextual information is available.

![Figure 5](image)

**Figure 5.** Top 10 suggestions provided by the Value Recommender for the disease field of the BioSample template, with increasing levels of context. The last plot shows 6 different values because only 6 suggestions were returned.
Discussion

We developed and deployed a metadata recommender service as part of an end-to-end metadata management system called the CEDAR Workbench. We found significant improvements could be obtained by considering contextual information when making recommendations. Our evaluation suggests that, by adding our recommendation capabilities on top of already-offered user interface optimizations, the CEDAR Workbench can provide major enhancements in both speed of metadata creation, and accuracy of those metadata.

A limitation of our evaluation is that we did not use all values in GEO, restricting our analysis to human samples only. We plan to carry out further analyses using all samples in GEO to determine how our method generalizes to additional sample types. We also plan to perform similar analyses on data from the BioSample repository, which contains metadata on 2,787,750 public biological samples used in scientific experiments. Additionally, we plan to study how the metadata recommender service performs with templates that contain a greater number, and more diverse fields. While disease, sex, and tissue are relevant metadata for biological samples, they are a small and relatively simple set of fields and generalizability may be a challenge.

The Value Recommender system is the first of a planned set of intelligent authoring components in the CEDAR system. Future efforts will concentrate on deeper analyses of metadata to discover more complex relationships among metadata fields, which will then drive tools to assist users when entering metadata. As a first step, we plan to extend the recommender to derive and use more in-depth knowledge of correlations among values in the dataset. Specifically, we will apply our previous research on association rule mining to identify degrees of correlation among metadata items. This approach will strengthen the positive associations that our current recommendation engine provides, and will allow us to point out possible errors by identifying unlikely values. With sufficient levels of accuracy, our system may be used to automatically fill in missing values for a significant number of metadata fields.

Our work also has implications for scientists focused on retrospective augmentation of metadata. For example, the system could be used to interactively help curate previously submitted data by suggesting more specific values for populated fields, in addition to suggesting values for empty fields. It could also assist curators with suggestions for correcting incorrectly entered element values. In particular, the system could be targeted to both retrospective and real-time quality assurance by detecting unlikely field combinations. Strong discrepancies detected between the element value entered by a user and the predicted value could be highlighted to human curators for review. By rapidly providing highly interactive recommendation, the system could also help curators quickly deal with greater volumes of metadata submissions, and help address the problem of curation scalability faced by many repositories.

Conclusion

We have described the development of a recommendation framework that focuses on helping biomedical investigators annotate their experimental data with high quality metadata. The framework takes advantage of associations among the values of multiple fields in existing metadata to recommend context-sensitive metadata values. A key focus is on interoperation with ontologies. Using formal ontology-based specifications and interactive look-up services linked to the BioPortal ontology repository, the system tunes its recommendations to target controlled terminologies. We outlined an initial evaluation of the framework using metadata from the GEO repository, and provided an implementation of the system in a metadata management system called the CEDAR Workbench.

These tools aim to provide a series of intelligent authoring functions that lower the barrier to the creation and population of metadata templates, and help ensure that the resulting metadata acquired using these templates is of high quality. The ultimate goal is to provide the ability for investigators to easily create metadata that are comprehensive, standardized, and make the corresponding data sets conform to FAIR principles.

Acknowledgements

CEDAR is supported by the National Institutes of Health through an NIH Big Data to Knowledge program under grant 1U54AI117925. NCBO is supported by the NIH Common Fund under grant U54HG004028. The CEDAR Workbench is available at https://cedar.metadatacenter.net and on GitHub (https://github.com/metadatacenter).
References

Developing a concussion assessment mHealth app for certified Athletic Trainers.

Scott P. McGrath, MS¹, Melanie L. McGrath, Ph.D. ATC², Dhundy Bastola, Ph.D.¹
¹ School of Interdisciplinary Informatics, University of Nebraska at Omaha, Omaha, NE; ² Dept. of Health and Human Performance, University of Montana, Missoula, MT

Abstract
Annually, 1.6-3.8 million concussions occur from sports in the United States, which account for 5-9% of all sports injuries. The dangers of concussions include prolonged post-concussive symptoms, increased risk of subsequent concussions, seizures, mental health issues, and in cases of second-impact syndrome (SIS), possible death. Certified Athletic Trainers (ATC) continue to serve an important role in providing assessment and treatments for athletes with sports-related injuries. They provide a critical safety net due to limited knowledge and misconceptions of concussion held by some youth sports coaches and athletes. However, availability of services from ATCs in rural areas is a challenge. In order to help extend coverage to more rural student athletes, we propose designing a telemedicine app following the mHealth development roadmap from the Center for eHealth Research (CeHRes). In this paper we will document contextual inquiry, user requirements capture, design phases, and app evaluation from the targeted user base.

Introduction
There are an estimated 1.6-3.8 million concussions that occur from sports and recreational activity in the United States annually¹. Board Certified Athletic Trainers (ATC) provide a critical safety net due to limited knowledge and misconceptions of concussions held by some coaches and athletes. Studies have shown schools with ATCs have lower overall injury rates, and higher rates of concussion diagnoses due to increased reporting and accurate assessment. However, only 55% of athletes in United States public secondary schools have access to a full-time ATC, and only 37% of schools employ a single ATC. Currently 30% of public secondary schools in the US lack an ATC, leading to 24% of sporting events that have no medical coverage². Due to budget restrictions, many schools resort to contracting an ATC from local physical therapy clinics or hospitals. This is an alternative to hiring a full time ATC, and often the ATC under contract will cover two or more schools. These ATCs have to select which school they will cover when events are held concurrently. This contributes to the overall medical coverage gap. Given the budgetary concerns in many schools, tools that facilitate remote diagnosis and/or treatment of patients by means of telecommunications technology may be one solution; specifically, an mHealth app may be ideal for many schools.

Concussions account for 5-9% of all sports injuries, which often translate into visits to the emergency room³. For most athletes (80-90%) concussions resolve in 7-10 days, however in some cases, symptoms can persist for months or even years⁴. The dangers of these injuries include prolonged post-concussive symptoms, increased risk of subsequent concussions, seizures, mental health issues including depression, and in rare cases, death. Of particular concern is second-impact syndrome (SIS), which may occur if an athlete returns to activity while still exhibiting concussion symptoms. SIS can cause a rapid increase in intracranial pressure, brain stem herniation, and often death. The most worrying aspect is while still recovering from the initial concussion, an additional impact of lower severity can lead to the onset of SIS. Chronic traumatic encephalopathy (CTE) is emerging as a long-term health concern that may be linked to concussions. The key to minimizing these dangers involves timely diagnosis, comprehensive treatment and rehabilitation plans, and return-to-play guidance from trained medical professionals like ATCs. In the absence of medical professionals on the field, the ER is often the only outlet for athletes with suspected concussions. When reviewing reports from ER visits, a 62% increase in non-fatal traumatic brain injuries occurred between 2001-2009⁵, indicating that ER visits for concussions have been on the rise. It is estimated that 12% of ER visits are non-urgent, and studies have shown that overuse of ER detrimentally impacts the entire hospital⁶. Unnecessary visits to the ER for non-urgent conditions also carry a risk of nosocomial infections, overtreatment, and lack of continuity of care. Hence, there is an incentive to provide these injured athletes access to medical care without having to escalate that care into a trip to the ER.

For concussions, the standard of care when evaluating concussions is by a physician or licensed healthcare provider who are specifically trained. Diagnosis of a concussion involves identifying one or more of the following clinical
domains: somatic symptoms, cognitive impairment (e.g., slow reaction times), behavioral changes (e.g., irritability), sleep disturbance (e.g., insomnia), and physical symptoms (e.g., balance, amnesia, loss of consciousness)\(^5\). To assist with the diagnosis of sports concussions, there are several technological approaches available. Functional neuroimaging is the most popular approach; there is also qualitative EEG, and impact sensors. A computer-based neuropsychological testing tools has been developed called the Immediate Post-Concussion Assessment and Cognitive Testing (ImPACT)\(^6\) which is also very popular. Telemedicine and mobile devices are options that have been relatively unexplored as potential tools\(^7\). Work by Vargas et al. has proposed a wheel and spoke telehealth approach, centralizing specialists who are available via a telemedicine robot\(^8\), but the utilization of an mHealth app solution is largely untapped. There are over 20 mobile concussion assessment apps available on the market, but none of these are part of a telehealth system. Most of these existing apps only assist the user in detecting if a trained healthcare provider should evaluate the individual. A recent systematic review of smartphone and tablet concussion apps raises concerns about the appropriateness of these tools for different groups of end users, and they also found that the apps were not always in compliance with testing guidelines\(^9\). There are some validated tools used by medical professionals for concussion evaluations, like the Sport Concussion Assessment Tool 3 (SCAT3)\(^10\). When used by a non-medical layperson, the SCAT3 test can be used to help identify signs of concussions, but diagnosis should remain a task left to medical professionals. A less technical sport concussion recognition method, like the Pocket Concussion Recognition\(^11\) tool is recommended for anyone without a medical background. This is where we would distinguish our proposed telehealth app. With a user-centric development path, and using ATCs as members of the focus group, the intent is to ensure ATCs remain the nexus of the medical evaluation process. Many of the available apps aim to place the coaches, parents, or athletes into the drivers seat for these decisions. This can be problematic due to biases (an athlete trying to “tough it out” on an injury) or possible conflicts of interest (placing winning over the health of the athlete).

Visual examination during concussion testing is a key component, which is what led us to pursue a telehealth model. There are sections in the SCAT3 test that rely on observing the athlete, like the balance test. It is also important to watch facial expressions to ensure the subject comprehends the question. For example, the ATC will want to make sure they understand all the questions being asked, such as if they are experiencing nausea. Does the athlete know what nausea is, or is there any attempt at evasion during the questions? This begs the question, why not just have the ATC complete the SCAT3 test via Skype, FaceTime, or other similar video conferencing tools? It is possible to do, but not practical. Usability and secure storage of health information are two major obstacles encountered with this approach. An ad hoc solution of conducting the SCAT3 over a conferencing app still would still rely on using paper records. If the video conferencing was integrated into a single app, there are inherent advantages. Currently, the SCAT3 is principally designed to be paper-based tool, where sections of the test are completed via print outs of the test. Any paper based data storage approach is faced with major concerns about issues with readability, data entry errors, retrieval, and backup\(^12, 13\). Many of these problems can be avoided by performing data collection into a secure HIPAA compliant database, following guided data entry steps with error checking, all into a file attached to a particular individual. These paper files also would have a high likelihood of being stored under lock and key in the ATC’s office in addition with any other medical files. Retrieval is not a problem if both the athlete and the ATC are in that office, but in the rural environment, the ATC is often working remotely. Secure logins and file encryption would ensure the data would be accessible to the ATC as long as they were able to connect online. Using alternate video conferencing tools is a shortcut that would solve the access to medical professionals temporarily, but we feel the complications associated with it would discourage repeat use. Initial conceptualization of the

The primary aim of this paper is to detail the development of this proposed mHealth app following the protocol recommended by the Center for eHealth Research (CeHRes)\(^14\), presenting a prototype version of the app, and reviewing its reception by ATCs. Specifically, four primary components of the five stages of the CeHRes roadmap will be conducted: contextual inquiry, value specification, design, and operationalization.

**Importance of Research**

Telemedicine is an established tool for extending medical coverage to underserved and rural areas\(^15, 16\). Telestroke\(^17\) is one example of a successful telehealth tool that has been able to provide cost-effective\(^18\) coverage to rural areas where stroke specialists are lacking\(^19\). However, there is a lack of work in developing a telehealth solution for assessing concussions and helping to extend care and treatment options for rural high school athletes\(^7\). When looking into application options for this technology, there is a distinct opportunity for rural Nebraska. The distance between rural high schools is a factor impacting the availability of full time ATCs to provide medical coverage. The

\(^{1276}\)
Nebraska School Activities Association (NSAA) is the organizing body that oversees all interscholastic activity for the state. Table 1 highlights sports in NSAA’s seasonal schedule with increased risk and the number of rural participants in those highlighted sports for that season. The tally was taken from NSAA’s 2014-2015 statewide report of participants per school. Each school was identified as rural or urban by using the Economic Research Service Rural-Urban Commuting Areas (RUCA) for counties with codes 4-10, codes 1-3 were designated as urban. Detailed statistics on the exact number of ATCs covering these athletes in Nebraska is difficult to find. There is a lack of standard reporting to indicate the number of staff ATCs employed by high schools, as discovered by Schwaderer and Unruh. Some schools administrators recorded full time ATCs in the same fashion as ATCs that were contracted out to multiple schools, making it challenging to identify exact coverage gaps. They were able to build a general ATC coverage map that depicts the disparity of ATC availability between rural and urban areas in Nebraska (Figure 1). The numbers of rural participants at risk for concussion and the large areas of Nebraska with coverage gaps lend itself to exploring a solution through mHealth.

<table>
<thead>
<tr>
<th>Sport</th>
<th>Fall</th>
<th>Winter</th>
<th>Spring</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cross Country</td>
<td>Basketball</td>
<td>Baseball</td>
<td></td>
</tr>
<tr>
<td>Football</td>
<td>Swimming</td>
<td>Boys Golf</td>
<td></td>
</tr>
<tr>
<td>Softball</td>
<td>Wrestling</td>
<td>Soccer</td>
<td></td>
</tr>
<tr>
<td>Boys Tennis</td>
<td>Cheerleading</td>
<td>Girls Tennis</td>
<td></td>
</tr>
<tr>
<td>Volleyball</td>
<td>Track and Field</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Girls Golf</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| Total High Risk Rural Participants | 9,173 | 13,260 | 1,992 |

*Sports in bold are sports with the highest risk for concussion*

Figure 1. ATC Coverage in Nebraska (image courtesy of Schwaderer & Unruh @ UNK)

Roadmap

The Center for eHealth Research roadmap was selected as the template for developing this project. One of the primary goals of this app is to make it appealing to the audience we are trying to reach, primarily ATCs and their student athletes. The human-centered design of the CeHRes road map helps to achieve this aim.

There are five phases of the road map, with iterative elements to help ensure that improvement is occurring through each stage of the mHealth app development.

1. **Contextual inquiry** – the app’s target user perspective is evaluated and the current standard of care is considered.
2. **Value specification** – Stakeholders and values are identified and translated into user requirements.
3. **Design** – taking the user requirements, developers, stakeholders, and targeted user collaborate in building the app.
4. **Operationalization** – once the technology has been built, steps are taken to market the product and integrate it into practice.
5. **Summative evaluation** – the final stage is a review of the product, how is it being used and what kind of impact is it having on healthcare.

**Research Approach or Methodology**

In following the CeHRes roadmap, the first step is to identify key stakeholders and explore the relevance of the idea in the contextual inquiry phase. A focus group comprised of these stakeholders was held as a part of the value specification phase. The output from the focus group helped to guide the development of the app in the design phase of the CeHRes roadmap. The target market for this app is aimed at high school level athletics (grades 9-12). High schools often are faced with restrictive budgets that contribute to the previously described ATC shortages, in comparison to college and university athletics, which have more robust budgets. At this point, the contextual inquiry, and value specification, and design phases have been completed. The Operationlization stage has started, being with revisiting the requirements captured at the beginning of the process with the focus group and conducting an End-User Computing Satisfaction (EUCS) survey. For the focus group component of the study and the EUCS survey, IRB 206-16-EX was submitted as an exempt educational, behavioral, and social science research review, and was accepted by the UNMC IRB board. The last two components of the CeHRes roadmap, operationalization and summative evaluation will be presented when those stages are complete and field-testing has been completed. These final two stages focus on marketing and market impact once the app has been released. Completion of the design phase provided a tangible product, which we felt presented a natural break in the process that could be shared. Additional details on these plans will be explained later in this paper.

**Research question**

There are three principal research questions that will be investigated in this paper:

1. **Who has a stake in addressing the healthcare problem of concussions?**
2. **How can technologies contribute to a better quality of care, and in what manifestation can technology contribute?**
3. **Was the design process able to meet user requirements established by the stakeholders and targeted users?**

**Results**

**Contextual inquiry phase: Stakeholder and targeted users**

After a literature review and market examination was completed, the authors of this paper conducted a brainstorming session to identify all stakeholders and targeted users. The results from that session generated the following list:

1. **Patients** – the student athletes participating in the sporting event, and are suspected of suffering from a concussion.
2. **Certified Athletic Trainers** – medical providers assigned to provide coverage for the sporting events. Typically assigned to a particular team or school, for the purpose of this app, these ATCs would be at a different location than the patient.
3. **Emergency Medical Technician (EMT)** – EMTs occasionally have been hired to provide medical coverage for these sporting events.
4. **School nurses** – school provided medical professional.
5. **Parents & Guardians** – due to the targeted patients for this app are high school aged athletes, many would be considered minors, and thus require their involvement for this mHealth app.
6. **Coaches, Staff, and Game Officials** – individuals that are involved with the patient’s sports team, or are involved with the officiating of the match.
7. **Principals, Athletic Directors, High School administration** – many of the matches and practice sessions where a potential concussion could occur happen on school property.
The primary user base was designated as ATCs who work with rural athletes, and the coaches of those rural athletes, and thus were critical to the design stage. Other stakeholders could be approached once the app was developed.

**Value specification phase: focus group**

The focus group was conducted on May 25th, 2016 for 90 minutes. Five members were included in this interview session. Four of the members are certified athletic trainers with various backgrounds and levels of experience, and one member was a developer with medical software applications experience. The first ATC is practicing at a local high school in the Omaha metropolitan area. The second ATC is employed at the University and is involved in an athletic training education program. This program offers both undergraduate and graduate degrees in athletic training. The third ATC previously worked at the collegiate level, and now works at a medical center as director for sports medicine in addition to being member of the national certification board for athletic training. The fourth ATC is a director of operations for a clinic that subcontracts out ATCs and holds an executive role in the Nebraska State Athletic Trainers’ Association (NSATA), part of the National Athletic Trainers’ Association (NATA). This provided a wide range of views and perspectives to draw from.

The focus group was recorded and transcribed into a word document to serve as input for app requirements. The transcripts and artifacts generated by the focus group were used to guide efforts in the design stage. The paper versions of the definitions of the "10 E’s in e-health" and a print out version of the preliminary app mock-up (Fig. 2) were collected after the session was complete.

![App mock-up screens depicting data entry and data retrieval screens](image)

**Figure 2.** App mock-up screens depicting data entry and data retrieval screens

Eysenbach’s “10 E’s” are used to help establish a clear vision for the mHealth app. They are values that encompass what e-Health should aim for:

- Efficiency
- Enhancing quality
- Evidence based
- Empowerment
- Encouragement
- Education
- Enabling
- Extending
- Ethics
- Equity

In Eysenbach’s paper, each of these terms is explained in more depth. However, having the focus group define these values in context of the mHealth app helped to identify what elements are most desired from the final product. Following up and having the group pick the top three after the definitions are complete provided insight to what was desired the most from the mHealth app. These documents were treated as additional research artifacts, which contained notes, and questions that the participants had during the focus group interview.

**Guidelines for designing a telehealth concussion app**

To help elucidate the design guidelines, there needs to be a clear directive captured from the stakeholders in the focus group. These were specified as values, and helps to establish the boundaries and directions our development work should take when building the app. To help identify the key values for this project, the focus group reviewed Eysenbach’s list of “10 E’s”. From the list, three stood apart from the rest: *enhancing quality of care, efficiency, and extending the scope of care.*
The focus group was encouraged to define what these terms mean, ‘in their own words’, when discussing how an mHealth app could aid in remote concussion assessment. For enhancing quality of care, the emphasis was more on follow up care over the initial assessment. As one member stated: “There is value for using this tool on the sidelines, but I think it would be more helpful when we check up on them later.” The consensus of the group was that this would not be a tool that they would expect either athletes or coaches to pull out and use in the middle of a match. After the match or practice was seen as the more likely opportunity for the mHealth app to come into play. This also was reflected in how the group defined efficiency. Eysenbach states that efficiency decreases cost, by either cutting unnecessary or duplicate diagnostic interventions. The focus group felt that efficiency was more along terms of time saved. By allowing ATCs to follow up with athletes across town, or in another town entirely, the time spent traveling to those locations could be reduced. Often, when ATCs are contracted out to multiple locations, they spread their visits out across the week. It is possible that an athlete may have to wait a few days before the ATC is scheduled to visit their location. This exchange was particularly enlightening because our initial vision for the app was a real time assessment tool. However, these stakeholders helped us refine the vision for the app to more of a post match assessment tool. As they explained, it is not uncommon for athletes to shake off signs of concussions as just wear and tear of the sporting event, only to notify ATCs when they haven’t fully recovered over the next 24-48 hours.

The third major value was extending the scope of care. Both the focus group and Eysenbach frame this value in terms of overcoming geographic barriers. Finding times to be available for a teleconference using the mHealth app would be much easier than trying to plan actual trips to each location. Following this path would also help to overcome potential technical hurdles, such as wireless signal strength. If the app helps the health care provider with their scheduling, it would be possible for both the athlete and the ATC to meet at a time where both can ensure strong Internet connections.

The focus group also discussed requirements that they would like to see for how the app operated. The SCAT3 exam is a validated tool, but only when used by properly trained individuals. In order to help keep the evaluation under the control of the medical professionals, it was suggested that the app have different tiers of access. Some of the more technical questions that move beyond the basic capture of the mechanism of injury and who was involved should be handled by the ATC. To accomplish this request, the idea was to have initial screening data that could be entered by a layman (another athlete, a parent, or a coach). That data would be forwarded on to the ATC for them to make the call if the evaluation should continue. These additional conversations helped to outline key values for the end-user of this application. 1) Ease of use. 2) Confidentiality of athlete records. 3) Timely notifications and information retrieval.

The focus group established that the mHealth app should be defined as a success or failure if it could measurably increase the detection of concussions in comparison to events where medical supervision was absent. Although beyond the current scope of the project, the intent is to continue development to reach a point where clinical data can be captured with the app in order to measure this metric.

App concept

The app was built using agile development methods with student developers from the University of Nebraska at Omaha. Agile development cycles were tracked using the web collaboration tool Trello. The initial vision for this app centers on converting the SCAT3 test to a mobile platform. The process flow (Fig. 3) was an output from the focus group session.

![Figure 3. ATC telehealth mHealth app process flow](image-url)
Player information can be captured prior to participation in a sporting event, helping expedite the submission process. The SCAT3 has eight primary sections, with additional data entry for mechanisms of injury, athlete background, and signs of concussions. Accessing the system does require log in for any user. Two primary accounts types will be available, a limited access account restricted to data entry and review of their own submitted reports (coach view), and an account with both submission and retrieval access to complete the report reserved for medical professionals (trainer view). A third account type (administrator), will be used to create new users and maintain the system. Once logged in, the layperson can complete three initial sections of the SCAT3. The first section is titled “Potential signs of concussion”, which ask simple yes/no questions for the observer of the injury (e.g., “Any loss of consciousness?”). The second section is the Maddocks Score test $^{25}$, which helps to identify situational awareness by asking the athlete simply memory tests like “What team did you play last week?”. The final section includes the injured athlete’s background (demographics, any history of concussions, etc.). This data capture is consistent with what information is collected in the Pocket Concussion recognition tool, and this does not require medical expertise.

When complete, a report would then be sent to the database, and notification will be forwarded on to an ATC of its entry. The ATC would then log into the system via the website, retrieve and review the report, and make a decision. At this point, if they decide no further intervention is needed, they can log the report for follow up later and clear the athlete to return to play. Alternatively, they can decide to continue the evaluation, and complete the additional six elements remaining on the SCAT3 test in order to generate a score. Return to play decisions are on a 6-point graduated scale ranging from abstaining from all activity, to allowing light aerobic exercise and incrementally increasing activity levels until culminating at clearance for normal game play$^{3}$.

The database for the app was developed in PostgreSQL and is currently not HIPAA compliant. Postgres can easily be converted into a MySQL or Microsoft SQL database if needed. Before any real patient data can be entered, the app will need to be ported to a more secure platform. Details about this task are addressed in the future work section of the paper. As suggested by the developers, this project was broken in three phases. The first phase was building the core elements of the app for data capture, storage and retrieval between multiple users, which is now complete. The next phase would be porting it over to a secure, HIPAA compliant platform, and the final piece would be the video feed integration, since it would be the most technically challenging, and expensive component.

**Figure 4.** App data capture (via an app on a smart phone) and retrieval screens (via website on a tablet)

**Evaluation phase**

With a useable prototype complete, the app could be shared for feedback from ATCs and athletic training students (individuals who have not yet taken the certification exam). The focus group was invited to review the app and see if it met the three initial key metrics: 1) the app’s ease of use 2) maintaining confidentiality of athlete records 3) timely
notifications and information retrieval. Additional users were encouraged to download either an Android or iOS version of the app, build their own test cases, and provide feedback and complete an End-User Computing Satisfaction (EUCS) survey. EUCS is a validated tool designed to evaluate factors impacting user satisfaction, content, format, accuracy, timeliness, and ease of use\(^6\). The tool has also been used to evaluate other medical systems\(^{26, 27}\).

**Conclusions**

At this current phase of this project, we have a functional Android, iOS, and web based version of the app. We were able to define the requirements with the insight from the focus group. In addition, the process helped to construct answers to our primary research questions. Funds were not sufficient at this time to construct the videoconferencing element as originally proposed. This was a disappointing development, but as this was our first effort in mHealth app development, learning about cost limitations was an important takeaway. Pricing out bids for app development prior to the focus group session would have avoided this setback. Grant funding will be pursued to cover the additional costs associated with the development of the video component at a later date.

- **Who has a stake in addressing the healthcare problem of concussions?**
  
  We were able to provide a detailed list of stakeholders in the results section who have a vested interest in improving detection of concussions. From this list, we identified a primary user base of ATCs working with rural athletes and the coaches of those athletes. Success or failure of this mHealth app would depend on our ability to properly deliver a tool that meets the needs for this group.

- **How can technology contribute to better quality of care and in what manifestation can technology contribute?**
  
  By offering a solution that integrates telehealth and mHealth into an already accepted and validated concussion tool, we don’t have to worry about re-inventing the wheel. The augmentation of the SCAT3 will require testing to ensure that the mHealth app works to the same degree as its paper based iteration, but this change offers many advantages. Data capture and retrieval would be improved over the paper versions of the tests. Follow-up on athletes suspected or verified to have a concussion can be simplified and can occur more frequently.

- **Was the design process able to meet user requirements established by the stakeholders and targeted users?**
  
  The focus group was satisfied with the progress of the project so far. All three key metrics were deemed satisfactory, although there are noted points for improvement on some “ease of use” elements. This feedback will be incorporated into the next version of the app, which will be shared again with the focus group. Additional data was planned based off the results from the EUCS survey, but delays due to launching the app on Google Play and Apple iTunes prevented us from reaching our goal of 20+ completed surveys. Participation in the survey is still being solicited, and additional iterations of the app will use EUCS to measure progress.

One of the big lessons learned from this project so far has been the transition away from a concept of immediate access to a healthcare professional, toward more of a tool that can facilitate increased contact between the athlete and ATC. By following the CeHRes roadmap, there have been structured guidelines that helped us reach these new insights. Without taking the time to work through these stages, the app development would have most assuredly been built with the focus on sideline interventions. While the tool built with these specifications may have been effective in it is aims and deliver on its goals, adoption of the tool may have never taken off. It is not uncommon for an athlete to suspect something is wrong the day after a concussion, when symptoms have still persisted. As highlighted by the focus group, this app would be beneficial to extend care to those athletes too. There are two suggestions we would provide for enhancing the CeHRes roadmap, an emphasis on budget building/cost analysis and guidance for launching an app on either an Android or iOS platform. Overall, it is a very effective tool, and one recommend by us to others, with the caveat for these two areas. We would emphasize on having a clear handle on building costs before entering the value specification phase of CeHRes. Also, launching our app in beta proved to be a significant hurdle, one that should not be overlooked in planning for a successful mHealth app. The certificate, keys, and profile generation steps required for launching the app on the Apple iTunes store and Google’s Play store has a sizeable learning curve if it is the first time a user is publishing their app. Developer accounts also are required by the user (at the time of publication, $25 for a lifetime license on Android, and $99 per year for Apple’s iOS).
Both the Google Play store and Apple iTunes have mechanisms to run closed and open beta tests. Beta testing via the Google Play store is fairly simple, it is managed from the Google Play Console and only requires users to have a Gmail account to participate. Apple beta testing follows a different framework, which should be accounted for in planning for testing phases. Apple uses a program called Testflight for beta tests. Before beginning the beta testing program, the app needs to pass evaluation from Apple, and reviews are targeted to be complete 24-48 hours after submission. Once the app has passed review, beta testers can be added via their Apple ID. Users first need to download the Testflight app. Once their account has been added to the beta testing list, the app will appear inside of the Testflight app. Additional time should be budgeted for this overall process on top of helping beta testers who might not be familiar with the process themselves.

Future work

In April of 2017, the group in charge of SCAT skipped over SCAT4 and released the latest version, SCAT5\textsuperscript{28}. We plan on reviewing this latest version and adjusting our app as needed. Additional EUCS survey results will continue to be collected for further feedback on the app. There are three major goals remaining for this project: converting it over to a HIPAA-compliant environments for secure patient data storage (most likely using REDCap\textsuperscript{29}), building the video conferencing component of the app and eventually conducting a randomized control trial study to test the app in the field. Using FaceTime or Skype APIs are the most likely solution, since they offer encryption and security options needed for medical consultations. The purpose would be to evaluate the app’s effectiveness in detecting concussions in practice. We are excited by the stakeholder’s interest when discussing this project and plan on using our prototype to secure additional funding to support future development.

References


22. Nijland N. Grounding eHealth: towards a holistic framework for sustainable eHealth technologies: University of Twente; 2011.


Does a Community-Engaged Health Informatics Platform Facilitate Resource Connectivity? An Evaluation Framework

Mari Millery, PhD¹, Alejandra N. Aguirre, MPH² & Rita Kukafka, DrPH, MA, FACMI¹
¹Columbia University Department of Biomedical Informatics, New York, NY
²Irving Institute for Clinical and Translational Research, Columbia University Medical Center, New York, NY

Abstract
Community-engaged health informatics (CEHI) integrates informatics with community-based participatory public health. Addressing social determinants and population health requires mobilization of health-related resources in communities. We present a framework for evaluating the process and outcomes of a CEHI platform designed to improve connectivity among community health resources.

The GetHealthyHeights.org CEHI platform was implemented in an urban low-income community. It was designed to facilitate connectivity among health-related community-based organizations (CBOs). To evaluate the process towards and the achievement of connectivity, a conceptual framework, methodology, and operational measures were defined.

A system-level approach, such as social network analysis, is required to capture the community as one dynamic unit. The evaluation framework specifies network connectivity metrics based on a social network survey. A network survey of CBOs (n=35) at baseline demonstrates utility of social network data for characterizing connectivity among community resources. The evaluation framework models how informatics and community resources improve population health.

Introduction
Evidence indicates that zip code rivals genetic code as a determinant of overall health, urging more interventions on the “zip code level”¹. Accordingly, there is increasing demand for an ecological systems paradigm for community interventions, with focus on: community capacity; problem identification through community engagement; empowerment of the community; and the permeating role of culture². As the importance of social determinants of health is increasingly recognized, health systems must seek collaborations with community-based partners that are equipped to meet diverse social service needs of patients ³. From the healthcare financing perspective, population-based accountable care models provide additional incentives for health systems to address social determinants through community-based services⁴.

Community-engaged health informatics (CEHI) is an emerging field with many opportunities for novel discoveries, as it combines concepts and methods from biomedical informatics⁵, community-based public health approaches⁶, as well as other fields such as community informatics⁷. CEHI takes an ecological systems approach to community health². It extends the notion of Learning Health System to the community-level, and investigates the ecology of community health information in a “cyber-social ecosystem”⁸.

We applied community-based participatory research (CBPR)⁶ to design a CEHI platform GetHealthyHeights.org (GHH) for the medically underserved, urban, largely Latino community of Washington Heights-Inwood (WAHI)⁹. GHH features an innovative participatory architecture with functionality for facilitating community stakeholder interactions. In addition to community residents, and a range of other community stakeholders, GHH specifically serves health-oriented CBOs, as they are viewed as essential assets for building community health interventions¹⁰. As an intervention, GHH was designed to facilitate connectivity among stakeholders in the community.

The design of GHH builds on the lessons learned from GetHealthyHarlem.org¹¹, ¹². While the Harlem portal was designed primarily for community member participation, the GHH team decided to primarily focus on community-based organizations and other groups and agencies that care about health of WAHI. GHH is intended to be a CEHI platform that allows exploration of novel functionalities that promote community engagement and community health. The mission of GHH was collectively defined by the GHH Steering Committee as “an online community that engages people and organizations in Washington Heights-Inwood to discover, connect, and share resources to get healthy”¹³.
Selected functions include a community calendar, a local service directory, posting of multiple types of content (e.g., articles, videos, and links), the ability to comment and rate content, integration of social media for content sharing, use of Google Translate (especially for Spanish translation of content), creation of pages for local organizations and the ability to form groups that other users can join. Community organizations have an essential role in creating and disseminating content through GHH. Figure 1 shows an illustrative screen shot of GHH. An initial version of GHH was launched to the public in April 2015. As of February 2016 (23 months of public use), the website had 43,990 page views, 16,860 use sessions (average 733 per month), 39 partner organizations, and 286 registered users.

Figure 1: Screenshot of GetHealthyHeights.org

GHH seeks to include and engage all health-related resources in the community. A comprehensive resource mapping was conducted and is described elsewhere. Resources encompass public, non-profit, private, and informal sectors, and range from large organizations (e.g., universities, hospitals) to small informal entities (e.g., bulletin board, walking group). The GHH team is currently in the process of defining a taxonomy to better classify types of health-related community resources, organizations, and services. CBOs are an important sub-type of resource and play a central role in GHH. The evaluation approach described in this paper specifically focuses on detecting change over time in network connectivity among CBOs, and attributing the change to GHH.

CBOs play an important role in networks of community health resources and services but they have poor capacity for processing, using, and exchanging information. Care coordination is a well-documented problem within the health care system generally, suggesting similar challenges with coordination of community-based services and resources. Emerging literature in the area of knowledge translation is beginning to highlight the potential of CBOs in translating evidence into the community. But many barriers hinder evidence use and evidence-based practices of CBOs. Research has consistently identified that CBOs struggle with: access to evidence; time to process evidence; skills to review, summarize, and synthesize evidence; research terminology; and local applicability and acceptability of evidence. Strategies are needed to address these barriers in order to realize the potential of CBOs as important agents of improving population health. The GHH CEHI platform targets these challenges through a set of online functionalities specifically designed to facilitate connectivity and information access among CBOs in the WAHI community. Enhanced connectivity of a community health resources network may provide positive synergistic effects, and has a large potential pay-off in terms of health-related social capital, and resulting benefits to community wellness.

Network connectivity means that local community-based organizations (CBOs) are sharing information, collaborating, and cross-refering clients. Paarlberg and Varda argue that network exchanges are the most essential driver of capacity in the non-profit sector. Connectivity is an important outcome because it is expected to make delivery of programs and services more efficient, and to facilitate health-related social capital, contributing to improved community wellness and reduction in health disparities. In the context of health informatics, there is a vision of seamless exchange of health data across organizations, and CBOs are increasingly considered part of the...
health data exchange network. For this future vision of sharing data among CBOs to become reality, trust among organizations becomes a critically important intermediate connectivity construct. Broad-based data sharing will require trust as a quality of organizational linkages.

A system-level approach is required to study connectivity on the community level. Social network analysis provides such approach. Some prior research has used social network analysis to demonstrate outcomes of community-level interventions. For example, Ramanadhan and colleagues 24 used network analysis to show that a cancer prevention coalition building intervention increased the connectivity among members of an intersectoral agency network in Massachusetts. The PARTNERtool has been used in several studies to measure social networks of community partnerships26. It provides an established instrument for operationalizing the kinds of linkages and exchanges that constitute network connectivity.

The overall rationale of the GHH CEHI platform as an intervention and its impact on health draws from several theoretical influences, including literature on community engagement, social networks, and social capital27. Based on literature, our conceptual model postulates that network connectivity contributes to social capital22, and social capital contributes to community health23.

The GHH CEHI platform is purposefully designed to facilitate health resource network connectivity. The architecture of participation in GHH9 mediates connections and interactions among community stakeholders. The evaluation framework we present here seeks to specify some of the processes by which functionality in GHH, and participation in GHH, are expected to lead to more resource connectivity.

Prior research has not investigated the impact of a comprehensive CEHI intervention such as GHH. Studies are starting to explore ways of applying informatics methods in the context of community engagement28-32. Compared to other applications of health informatics in the community engagement context, GHH is unique in its focus on CBOs as the primary community partners. One reason why GHH partners closely with CBOs is the desire to enhance sustainability of GHH by building it in partnership with stable local community assets33. Lack of sustained use is a known challenge for information technology (IT) interventions that rely on voluntary motivation of users34, and is potentially addressed through a community-level approach that engages CBOs to champion the IT system.

This paper presents a conceptual model and describes how it is operationalized. Our view of operationalization comes from the social science research perspective35. We describe the conceptual underpinnings and the design of the evaluation framework. We also present baseline data for the network survey to substantiate how key outcome constructs in the framework were operationalized and measured.

Methods

Community Setting of Evaluation Study

The Washington Heights-Inwood (WAHI) neighborhood is located in upper Manhattan within New York City (NYC), north of Harlem and directly south and west of the Bronx. It is a densely populated urban area of 2.8 square miles and four zip codes with approximately 200,000 residents36. A large proportion of WAHI residents are Hispanic (71%), and 93% belong to a racial/ethnic minority group. African Americans represent 7% of the population37. Almost half (48%) of residents are foreign-born, mainly from Latin America, with 2/3 from the Dominican Republic. Many ethnic groups are represented as foreign-born residents come from a total of 55 countries38. A large proportion (39%) of residents have limited English proficiency37. The median household income was $37,460 in 2013, and 27% lived below the federal poverty level38. A large percentage (30%) of adults did not graduate from high school. The most common sources of employment are services and sales industries39. Housing quality is rated among the lowest in NYC37. Health concerns in the community are significant compared to NYC as a whole. Among WAHI residents, 27% rate their health as fair or poor, compared to 22% citywide, and 16% reported that they went without needed medical care. Obesity rate is 22%, and 22% of residents report they had no physical activity in the past month. Only 36% had influenza vaccinations37.

Approach to Defining Evaluation Framework

The overall approach includes three major components: (1) conceptual modeling, (2) selection of evaluation design and methodology, and (3) definition of operationalized measures.
The plan to evaluate process and outcomes of GHH evolved during the design stage of the project and was based on conceptual modeling of relevant constructs. Important key constructs included: “engagement” to capture active participation of users with GHH, “network connectivity” to capture a system level outcome on the level of the community as a whole, and “health-related social capital” as a construct that bridges network connectivity to community health outcomes. Detailed conceptual analysis was conducted to further break down components of each of the key constructs. Several iterations of logic models and conceptual diagrams were developed to capture the conceptual framework. Community stakeholders were engaged and provided input into the models, primarily through the GHH Steering Committee.

An evaluation design and methodology were developed based on the conceptual modeling and considerations of feasible data collection options. While recognizing the limitations of a pre-post design, an outcome evaluation design was selected to capture change in network connectivity over time in the target community. Consistent with GHH focus on CBOs, the team decided to primarily focus the evaluation on network connectivity among CBOs. Baseline data collection was conducted during early implementation of GHH, with a plan to collect similar data at later time points.

The PARTNERtool was selected as an established measure of network interactions among community agencies. It was adaptable for the purposes of the GHH evaluation. The PARTNERtool provided a foundation for defining operationalized measures of constructs that constitute “network connectivity”, as further described in the Results section below.

The process evaluation was designed to document and measure how organizations and individuals engage in GHH. Process evaluation data sources included Google Analytics for GHH, system back-end data from GHH use logs, and records of participation in GHH governance and other GHH-related activities. In the outcome evaluation design, the process measures also serve as independent variables, with the plan of measuring associations between GHH engagement and network connectivity outcomes for individual CBOs.

**Baseline Network Survey Methodology**

The primary inclusion criteria for the network survey was: non-profit organization that provides health-related services to the WAHI community and has a physical office location within the four targeted zip codes. The research team went through a systematic process of identifying the list of CBOs. This process involved decisions about what kinds of organizations to include as “health related”. The initial list of organizations was based on the researchers’ prior experience of working in the community, complimented with results from a systematic community resource mapping. The resource mapping was based on five data sources on community resources, including most recent files from the Internal Review Service listing all organizations with 501(c)(3) non-profit status within the four zip codes. We have described the resource mapping process elsewhere.

The team identified 25 qualifying non-profit organizations, including 8 multi-service CBOs, 6 senior centers, 2 nursing homes, 2 health centers/health center networks, 2 behavioral health providers, 3 child/youth service providers, 1 free-standing food pantry, and 1 HIV/AIDS service provider. The team also decided to invite three local city-government organizations to participate in the survey: local Community Board (part of city government), and 2 recreation centers (operated by city). In addition, the team decided it was important to represent the following seven entities in the survey to capture a more comprehensive picture of the community’s network: city government (libraries, health department, police department, etc.); Columbia University; Yeshiva University; K-12 public, private, and parochial schools; New York Presbyterian Hospital community programs; New York State Psychiatric Institute (located within the community); and faith-based organizations. The 25 non-profits and the three city government entities constitute the pool of 28 organizations invited to complete the survey. The 28 organizations and the seven additional entities constitute the set of 35 entities listed on the survey. Each survey respondent was asked about their connections to all other 34 entities.

We used an adapted version of the validated PARTNERtool. The PARTNERtool was designed to collect network interaction data from public health collaborations. It includes a section where the respondent is asked to indicate relationships and interactions with all listed agencies. Organizational leaders were recruited to respond to the surveys. The instructions stated that the Executive Director (or equivalent, or someone he or she designates) should fill out the survey as a representative of the organizational perspective. The survey administration was managed through the PARTER tool website.
The survey protocol was approved by the Columbia University Medical Center Institutional Review Board. The survey data were collected in summer and fall of 2016. It should be noted that GHH was already open to the public at that time, and many of the organizations in the survey already began initial use of GHH before they were recruited to participate in the survey. Thus, the baseline survey was collected during early stage implementation of GHH.

**Results**

**Evaluation Framework**

Figure 2 depicts the conceptual logic of the causal chain of outcomes we theorize for the evaluation framework. The evaluation focuses on measuring engagement in GHH as the primary independent variable and connectivity in the community resource network as the primary outcome variable. The other outcome constructs are specified to create a full picture of how and why we think GHH contributes to community wellness. In particular, the logic model defines health-related social capital as the construct that links resource network connectivity to health outcomes. Even though we do not measure them in the current evaluation, we postulate several components of social capital that the GHH intervention can improve. For example, we argue that improved connectivity will lead to efficiency in the way the community as a whole is able to deliver health programs and services.

**Figure 2:** Logic model of outcome constructs

Further conceptual analysis was conducted to elaborate on the two constructs the evaluation is designed to measure, namely, *engagement in GHH* and *connectivity*. This step is grounded in a more concrete level, looking at the actual functionality built into GHH and the actual activities of connecting among CBOs. The two bottom boxes of Figure 3 represent these two domains. The model suggests that the GHH system functionality (and participation in GHH governance) facilitate specific ways in which CBOs connect among one another. The top two boxes of Figure 3 represent how the evaluation measures the constructs in the bottom boxes. The top section of the figure captures the essential elements of the evaluation plan. Measures of the processes (and independent variable) of GHH use and participation are described on the top left, and specific network connectivity measures are described in the top right. The study design examines change in network metrics over time, but in addition, as indicated in the figure, we plan to measure associations between GHH use/participation and connectivity indicators for individual organizations.

It should be noted that the engagement of CBOs occurs within and outside of the actual online platform. In fact, in-person meetings related to governance of GHH appear to be a powerful mechanism for engaging organizations around GHH. Other engagement processes, such as sharing content or finding resources, reside more directly on the GHH website.

**Baseline Network Results**

The baseline network survey was sent to 28 organizations; with a 61% response rate. The survey listed 35 total organizations; 28 that received surveys and seven that were listed to represent selected groups or entities (e.g., faith-based organizations) that were not recruited to respond to the survey.
Organizations that responded reported that they collectively had 273 connections with other entities. The average number of connections per organization was 13 (out of a possible 34). The top seven most connected had an average of 25 connections per partner, leaving the remaining 28 organizations averaging 10 connections per partner. Table 2 shows selected network metrics for the entire network, along with the definitions of the metrics.

**Figure 3:** Key constructs and evaluation measures

**Table 1:** Selected network metrics for whole network

<table>
<thead>
<tr>
<th>Network Metric</th>
<th>Baseline Survey Score</th>
<th>Definition of Metric</th>
</tr>
</thead>
<tbody>
<tr>
<td>Density</td>
<td>39%</td>
<td>Percentage of ties present in the network in relation to the total number of possible ties in the entire network.</td>
</tr>
<tr>
<td>Degree Centralization</td>
<td>56%</td>
<td>The lower the centralization score, the more similar the members are in terms of their number of connections to others (e.g. more decentralized).</td>
</tr>
<tr>
<td>Trust</td>
<td>64%</td>
<td>The percentage of how much members trust one another. A 100% occurs when all members trust others at the highest level.</td>
</tr>
</tbody>
</table>

Figures 4 and 5 show network diagrams depicting types and frequencies of baseline connections among the entities. As shown in Figure 4, about one-third (32%) indicated that they only had awareness of one another, while 47% indicated cooperative activity connections, 14% indicated coordinated activity connections, and 7% indicated integrated activity connections with one another.
Discussion

Our evaluation framework offers a conceptual and methodological map for measuring the impact of a community-level informatics intervention on network connectivity among community organizations. We trace conceptual and operationalized definitions from engagement with functionality of the information system (GHH) to different dimensions of interactions and exchanges that constitute network connectivity. We also discuss the intervention change model that postulates that network connectivity translates to health-related social capital\textsuperscript{22}, which, in turn, is expected to contribute to positive population health outcomes\textsuperscript{23}.

The baseline results of the social network survey of CBOs, using the PARTNERtool\textsuperscript{26}, demonstrate that the survey methodology produces a rich set of network metrics for describing the state of the CBO network at baseline. Among the many available metrics are density of connections, and different types and frequencies of connections. “Trust” is an important example of a specific quality of a connection that provides a conceptually meaningful outcome in the context of our evaluation. With “Trust” at 64%, density at 39%, the community appears to have many positive connections, but with ample room for increase from baseline. Examination of the baseline visualizations shows that many of the possible binary connections are not in place at baseline, and suggests that any new edge that is added to the network in the future can be examined as evidence of change, and potentially characterized resulting from GHH facilitation. The social network methodology allows analyses of the entire network as a whole system, as well as drilling down to the level of individual organizations. Having both levels is extremely useful for an evaluation framework in a transactional context such as CBOs connecting in a community and engaging with a system designed to facilitate such connections.
Engagement is a central construct within CEHI. The design of GHH with an architecture of participation is an operationalized instance of engagement processes. In the evaluation framework, use of GHH functionality becomes not only a process evaluation metric, but also means of measuring engagement as an independent variable. An association between use of GHH and improvement over time in network connectivity measures serves as evidence that GHH has its intended impact.

The GHH evaluation highlights the role of CBOs as special kind of health-related resource. Collectively, CBOs can establish a backbone of network exchanges that translates into community-level benefits that are more than the sum of its parts. CBO networks are a natural mechanism for addressing social determinants of health. In the future, medical care systems should pursue closer connections with CBO networks and other community resources. This will be a starting point for a Learning Health System that includes community services and resources. The future vision of such Learning Health System includes sharing and flow of data across medical care and community entities.

The construct of “trust”, which we examine in our evaluation as a component of inter-organizational connectivity, will be a critical pre-requisite for data sharing and exchange to ever become reality.

**Limitations**

We only present a preliminary stage of work that needs to be supported by further data collection and analysis. However, we believe it is valuable to present a focused conceptual analysis that constructs a coherent and organized model of the phenomenon under evaluation and an operationalized way of measuring it. Our initial analysis shows that much more can be done, even within the realm of operationalizing the measures. For example, internal correlations within the baseline network survey dataset can be used to further examine validity of sub-constructs of connectivity.
The baseline dataset can also yield numerous additional network metrics both for the entire network and for its sub-units. For example, we plan to examine measures of strong ties (bonding social capital) and weak ties (bridging social capital) within the network\(^\text{40}\). We can also further specify the statistical analyses approaches to be used to measure change in network metrics over time.

A pre-post evaluation design implemented in only one community has obvious limitations, but is a valid starting point for further work. It was feasible to collect baseline data only after implementation of GHH had already begun. At present time we observe community organizations increasingly engaging around GHH. Once we collect systematic data over time we will be able to further describe mechanisms by with GHH functionality facilitates the engagement.

Operationalizing valid constructs and study designs in the context of a complex information technology poses special challenges and limitations\(^\text{41}\). It is a limitation of validity if only one system is built to “instatiate” the constructs. It should be recognized that the specific and idiosyncratic ways in which this system gets designed may account for the evaluation results.

**Conclusion**

We argue that mobilizing community-based resources through informatics is an important strategy that should be leveraged to improve population health. GHH is an instance of such strategy, linking engagement in a CEHI technology to network connectivity among entities in a geographic community.

**References**

13. Google Analytics and GHH management records
Distinction between medical and non-medical usages of short forms in clinical narratives

Sungrim Moon, PhD, Donna Ihrke, Yuqun Zeng, MD, PhD, Hongfang Liu, PhD
Department of Health Sciences Research, Mayo Clinic, Rochester, Minnesota, USA

Abstract

The short forms of medical concepts or expressions (i.e., acronyms/abbreviations) are prevalent in clinical documentation. Given the limited number of potential short forms, they are also highly ambiguous. Resolving the ambiguity of short forms is essential in clinical natural language processing (NLP). However, one prerequisite for resolving ambiguity of short forms is to have a sense inventory. This paper outlines our process of identifying 141 potential short forms with randomly sampled phrases from a large clinical corpus. We assessed various features in their ability to disambiguate medical and non-medical usages. We identified 68% of our short forms as primarily serving medical usages, whereas 12% had non-medical usages. The remaining 19% showed alternating usage based upon case form. Our short forms had an average of 3.58 senses. Usages could be distinguished using basic trigram/bigram/line information. Our initial findings will be applicable for automatic usage/sense resolution.

Introduction

The short forms of medical concepts or expressions (i.e., acronyms/abbreviations) are prevalent in clinical documentation. Given the limited number of potential short forms, one short form can have multiple senses. Resolving the high ambiguity of short forms, a special case of word sense disambiguation (WSD), is essential in clinical natural language processing (NLP) systems. There are some unique characteristics of short forms in the clinical domain. First, the appearance of short forms in clinical documents is very different from other domains (1). They are seldom defined in the text but can have diverse meanings at different settings. For example, “CA” can mean “California”, “Cancer”, and “Calcium” without any clear definition in clinical narratives. “DT” can be frequently used as “Discharge time” in Discharge summaries and “Diptheria-tetanus double vaccination” in Immunization reports. Second, “one sense per discourse” (2), a common understanding in the general English WSD community, may not hold here (3). For example, “PT” can stand for “Patient” in the subjective section and it can mean “Prothrombin time” in the assessment section. Additionally, besides multiple clinical meanings, short forms in clinical texts can also have non-medical English usages. For instance, the short form “US” in “The US shows a lesion” and “The patient retired from US army in 2011” has two different meanings where the first one means “Ultrasound”, and the second one means “United States”.

WSD in the clinical domain has been studied including the use of rules or machine learning (ML) approaches (1, 4, 5). Various textual features (such as POS tags or n-grams) have been inspected (5, 6). However, those textual features may not perform well in documents that are semi-structured (e.g., containing tables) or not grammatically well formed (7) such as Emergency Department (ED) notes. Document metadata features are also crucial but they vary across different EHR systems or institutions. Additionally, expert or distributed semantic (e.g., SNOMED concepts or topic models) features have also been utilized but they highly depend on the comprehensiveness of knowledge resources (8) and the availability of a large collection of documents in the corresponding domain. Research (5, 9) has been focused to develop individual classifier per short form to achieve reasonable performance rather than one generalized classifier for all.

However, the prerequisite towards WSD is the creation of a sense inventory which catalogs the associated meanings of those ambiguous words (10). As described, short forms can have both medical and non-medical usages in clinical documents. Existing clinical terminology resources may not capture their non-medical usages. For example, the non-medical usage of “ICE” in “ICE number” as “In Case of an Emergency phone number” is not included in existing clinical terminology resources. Additionally, some short forms have localized usages. For example, “MEA” stands for “MayoExpertAdvisor”, a short form for a decision support system at Mayo Clinic or “Minnesota Education Assembly”.

In this paper, leveraging a large collection of clinical notes in our clinical data warehouse, we investigated the usages of short forms in clinical documents, assessed potential features which can be utilized for identifying varying usages, and assembled high frequent usage patterns towards WSD tasks.
Method

Dataset and resources

For this study, we used a corpus consisting of 40 million clinical notes from 1995 to 2011 retrieved from the Mayo clinical data warehouse. The corpus includes all notes generated from a diverse range of medical specialties. The total number of our word tokens in the corpus was 13.9 billion corresponding to 61.8 million unique word tokens. For each word token, we extracted all corresponding text lines.

We used multiple publicly available resources to aggressively identify medical and non-medical usages. Specifically, we leveraged the Unified Medical Language System (UMLS)\textsuperscript{(11)} of the National Library of Medicine (NLM) as the source for concepts. The UMLS integrates various controlled vocabularies pertaining to biomedicine by assigning a concept unique identifier (CUI) to link all clinical term variants with the same meaning together. Based on our prior work\textsuperscript{(8)}, we derived an short form list, 7,092 entries, from the UMLS. Another resource is allacronyms.com\textsuperscript{(12)}(data: Aug 18~19, 2016), which is a comprehensive resource for short forms with over 3 million entries across all domains. The website offers user-driven and topic oriented short forms using online search. We also utilized the following resources: titles of Wikipedia (date: July 20, 2016), unigram tokens in the Brown corpus\textsuperscript{(13)}, and the Kunth list of American English words\textsuperscript{(14)}.

Identification of short form candidates

Since short forms can appear in diverse forms as described in our prior work\textsuperscript{(8)}, to simplify our study we used the following criteria to identify potential short forms from the word tokens gathered from the corpus:

1. The number of characters needs to be at least 2 and at most 7. For example, a word token, “NA” is one of our candidates because its length is two with all capital alphabet letters. In addition, a length of “N.A.” is four according to our criteria.
2. The token contains all capital alphabet letters with or without numbers/special symbols taking its diverse expressions into account in clinical narrative notes. We consider special symbols as “+”, “-”, “/”, “&”, and “.”. Additionally, the token has at least one capital alphabet letter. E.g., we include “N+A”, “N-A”, “N/A”, “N&A”, and “N.A.” for a short form “NA”.
3. The token may contain special symbols in the middle but not consecutively occurring. Taking as an example for a short form “NA”, we exclude “N++A” or “N+/A-”.
4. The token may end with “+”, “++” and “.”. For instance, we regard “NA+”, “NA++”, “N.A+”, and “N.A.” as valid tokens.

We used the following heuristics to identify a subset of short form candidates for our study stratified to those with both medical and non-medical usages as shown in the left side of Figure 1. Specifically, we considered those short form candidates appearing in the short form list derived from the UMLS (i.e., they have medical usages) and at the same time occurring with multiple prevalent lexical variations (i.e., potentially have non-medical usages). For a word token, if after ignoring special symbols and capitalization, it appears in the short form list derived from the UMLS, we consider it a lexical variation of that candidate. To take example on one token “(n/a)” in our clinical notes, we stripped non-alphanumeric characters in the front/end of token as “n/a” at first. Next, we removed “/” in “n/a”, and then disregard the capitals. Therefore this “(n/a)” is a lexical variation of short form “NA”.

We counted two types of frequencies for each candidate in the right side of Figure 1. One is the number of occurrences containing of all lexical variations (loose form) and one is limited to those lexical variations without lower case letters (strict form). This case-sensitive distinction was examined because the lexical forms of medical usages (specially, medical abbreviations) have a tendency to consist of capital cases (strict form) compared with those of non-medical usages. We limited to candidates with the strict form frequency of at least 100. After logarithm transformation, we computed the difference between the two frequencies. For example, we counted 705,472 occurrences of “NA” as loose form and 158,348 occurrences of “NA” as strict form. After logarithm transformation, the difference is 2 (i.e., 19 minus 17). We then randomly selected five to six short form candidates from each frequency difference.
Manual review of the selected set

For each short form candidate, we extracted 100 unique lines containing the candidate in loose form and another 100 unique lines containing the candidate in strict form after removing those capitalized lines. We asked two experts to manually annotate the following usages and the specific meaning for the candidate in the context:

- **Medical Abbreviation**: the candidate is used as a short form for a medical concept. In a sentence, “Abdominal ultrasounds to check patency of TIPS performed on 09/10/03.”, the “TIPS” is abbreviation for “Transjugular intrahepatic portosystemic shunt procedure” as medical usage.

- **Medical Concept**: the candidate itself is used for a medical concept. For example, “tips” in “She’s had some surgeries on the finger tips for wound care and ulceration and a hysterectomy.” indicates “Body Regions”.

- **Non-medical Usage**: the candidate is used as a non-medical term. For instance, “tips” represents “Advice” in the sentence, “I gave her a brochure of tips on healthy brain aging.” with regards to non-medical usage.

- **Others**: the meaning of the candidate is unknown or potential typo. For example, “IBC” was frequently used as a typo of “inferior vena cava (IVC)”. e.g., “He is status-post IBC filter”. On top of it, we counted unknown meaning for human experts in this category. We also included “Spanish” and “Template limitation” in this category.

Following the above-mentioned guideline, the export annotated “FL” as (1) “Medical abbreviation” and (2) “Full liquid diet” for the sentence “Pt is tolerating diet adv to FL.” whereas (1) “Non-medical term” and (2) “Florida” for the second sentence “Pt had the follow-up at Mayo Clinic in Jacksonville, FL.”. All annotations were validated by terms in The UMLS Terminology Services (UTS) Metathesaurus (ver. 2016AA). The inter-annotator reliability of the annotated senses was calculated with percentage agreement based on 5% of total 28,200 instances. The usage agreement of short forms was also measured based on all identified senses.
Exploratory analysis of features for usage disambiguation

In this step, we performed an exploratory analysis of features with potential of disambiguating medical usages from non-medical usages. To identify representative variations of short forms, we stripped non-alphanumeric characters in the front/end of token in our clinical documents. Specifically, we investigated the following features:

- **The number of unique lines in strict and loose form**: For instance, “OR” appeared in 581,807 unique lines as strict form and 49,908,734 as loose form in our clinical lines. If the candidate appears in loose form relatively more frequently than in strict form, it implies that the candidate has a high chance to have a non-medical usage. Oppositely, the candidate is primarily used as a short form with medical usage.

- **The number of unique lexical variations in strict and loose form**: We assume that tokens with medical usages tend to have fewer non-strict variations. For example, “CPD” always appears as one lexical variation in strict form (i.e., “CPD”), so the number of unique strict lexical variations is 1. However, it has five unique lexicon variations, “CPD”, “CPd”, “Cpd”, “cpd” and “c.p.d.” in loose form, so the number of unique loose forms is 5.

- **The number of the candidate appearing as the signal word of a sentence (in loose form)**: The first letter of the first word in a sentence is usually capitalized. We consider a lexical variation is used as the signal word if the first letter is capitalized and the remaining letters are in low case. It implies the candidate has non-abbreviation usages.

- **The number of distinct normalized bigrams and trigrams when candidate appearing in strict and loose form**: For each candidate, we obtained all the occurrences of the trigram (left word, candidate, right word) with the candidate or its variation appearing in the middle. After converting all digits to “9”, ignoring cases of the left and right words, and removing punctuations, we counted the number of distinct normalized trigrams. For instance, one phrase for “PLAT” appeared as “9 PLAT 9”, which implies a “Blood Platelets” value in Laboratory Test. (This pattern has 80% of frequency over the total frequency of twenty-seven patterns in strict forms.) We calculated the number of distinct normalized bigrams taking a phrase as a left word and target word form into account as well. Our assumption was the proximal words (especially next left word) conserved the meaning/sense.

- **Existence in well-known dictionaries and resources**: We leveraged multiple medical and non-medical resources to identify the existence in those resources. We assumed if any short form frequently appeared in non-medical or medical resources, there is a high possibility to serve as the particular usage of that domain. We explored common medical usages using [www.allacronyms.com](http://www.allacronyms.com) limited to the medical field. For non-medical usages, we used [www.allacronyms.com](http://www.allacronyms.com) limited to non-medical fields and the following general resources: Wikipedia, Brown corpus, and the Knuth’s list. We ignored case information when checking the existence in those resources. Additionally, we took case information into consideration for Wikipedia, Brown corpus and the Knuth’s list. We used the binary representation (1= existence, otherwise 0) for our features.

We used Pearson product-moment correlation coefficient to assess the predictability of the usages of those features. Each value (except the binary representation of the existence in resources) was logarithmically transformed and the differences between strict and loose forms were used to compute the correlation against usages.

**Result**

Overall, the total number of short form candidates was 297 million (2.14%) corresponding to 0.53 million unique tokens in Mayo clinical corpus. Table 1 shows the set of 141 our randomly selected candidates after the process in Figure 1. Based on line samples of these selective short forms, we collected 518 senses over medical, non-medical and other usages with 89% of them normalized to UMLS concepts. 69% of our total senses served as medical (medical abbreviations and medical concept) usages, whereas 30% of them used as non-medical usages. Also, Our short forms have the average 3.58 senses. There were total 258 typos (178 and 80 case in loose and strict form, respectively) and 337 unknown senses (134 and 203 cases in loose and strict form, respectively) over all 28,200 random instances. Table 2 represents the details of our annotation senses per domain and form. The inter-annotator agreement is 95% when assigning senses to individual instances and 97% on usages (base on total 518 senses) of short forms.
Table 1. Selected 141 short forms

<table>
<thead>
<tr>
<th>Category per Log diff (D)</th>
<th>Selected short form candidates based on logarithm difference between strict and loose forms</th>
</tr>
</thead>
<tbody>
<tr>
<td>6 ≤ D &lt; 17</td>
<td>ACHE, ANT, AS, BICARB, CAR, CUP, DAD, DIET, DIZZY, DROWSY, DRY, FIND, HOT, ICE, LIP, NEOPL, NO, ON, OR, PATH, PER, PHOS, RAW, RICH, TEA, TOP</td>
</tr>
<tr>
<td>3 ≤ D &lt; 6</td>
<td>AAV, ACE, ALPS, AME, BS, CHO, CPD, CPVT, CTS, DPT, FS, GVHD, HCT, HEP, HS, INR, IV, MAC, MAHA, MALT, NEC, OD, OK, OP, OS, OU, PLP, PNT, PR, PSA, RF, RVOT, SM, SOM, SN, SR, TIPS, TPN, TUNA, UNC, VAD, VNS, WBG</td>
</tr>
<tr>
<td>1 ≤ D &lt; 3</td>
<td>ADD, AIDS, APR, ASH, CHI, CM, DRG, FISH, GAS, GTT, HB, IE, IGE, LAP, LYMPH, MAD, MAS, NA, PLAT, PO, POTTs, RM, RT, SAT, SCN, ST, STARR, TAB, TALC, TXT, US</td>
</tr>
<tr>
<td>D &lt; 1</td>
<td>ALT, AST, AVD, CBC, CLA, CT, DD, DOB, DT, EAA, ECG, ENT, FL, GI, GU, IBC, JVP, LDL, MCHC, MD, MHR, MI, MN, MPV, MRI, OAG, OHI, OPV, PIE, PPT, PVRI, RAIR, RDW, SMP, TIBC, TMB, TMP, TPO, TSH, TSR, WBC</td>
</tr>
</tbody>
</table>

Table 2 represents prevalent medical usages and low non-medical usages as strict form (75% and 25% of sense ratios respectively) compared to loose form (68% and 31%). However, both forms primarily have senses (about 67% in strict form and 59% in loose form) serving as medical abbreviations.

Table 2. The numbers of annotated senses with regard to domain and form

<table>
<thead>
<tr>
<th>Domain</th>
<th>Number of senses</th>
<th>All forms</th>
<th>Strict forms</th>
<th>Loose forms</th>
<th>Overlap between strict and loose forms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medical Abbreviation</td>
<td>313</td>
<td>271</td>
<td>229</td>
<td>187</td>
<td></td>
</tr>
<tr>
<td>Medical Concept</td>
<td>46</td>
<td>31</td>
<td>36</td>
<td>21</td>
<td></td>
</tr>
<tr>
<td>Non-medical usage</td>
<td>155</td>
<td>100</td>
<td>119</td>
<td>64</td>
<td></td>
</tr>
<tr>
<td>Others</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>518</td>
<td>405</td>
<td>388</td>
<td>275</td>
<td></td>
</tr>
</tbody>
</table>

Table 3 summarized our selected short forms according to dominant sense ratio and relevant usages. For example, “TEA”, this short form belonged to non-medical class if “Beverages” was the prime sense (96%) among 100 random unique lines as loose form. However, if dominant usage (75%) in strict form for “TEA” was “Arthroplasty, Replacement, Elbow”, it classified as “Medical usage (Medical Abbreviation)”. According to Table 3, 86% and 72% of our short forms were medical usages as strict and loose form respectively. This prevalence of medical terminologies is unsurprising given our input corpus of clinical documents. Furthermore, 78% (strict form) and 62% (loose form) of our short forms were clinical acronyms and abbreviations. Additionally, 80% and 77% of the selected short forms have high major senses ratio (over 90% of sense dominant) in strict and loose forms respectively, which reflects the general tendency for short forms to have a dominant tendency to serve a particular domain. Sets in strict form have a tendency served as medical usages as those most likely are commonly used short forms in the medical domain.

Table 3. Distribution of annotated major senses ratio for selected short form.

<table>
<thead>
<tr>
<th>Major senses ratio%</th>
<th>Selective short forms in strict form</th>
<th>Selective short forms in loose form</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Medical usages</td>
<td>Non-medical usage</td>
</tr>
<tr>
<td></td>
<td>Abbreviation</td>
<td>Concept</td>
</tr>
<tr>
<td>100</td>
<td>52</td>
<td>6</td>
</tr>
<tr>
<td>90~100</td>
<td>40</td>
<td>2</td>
</tr>
<tr>
<td>80~90</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>50~80</td>
<td>12</td>
<td>2</td>
</tr>
<tr>
<td>0~50</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Total</td>
<td>110</td>
<td>11</td>
</tr>
</tbody>
</table>
Table 4 shows the details of short forms classified into identical usage and different usage where identical usage refers to consistency in the major usage regardless of strict or loose forms. 81% of our selected short forms show consistent usage regardless of cases form, and the majority (61% of total 141 short forms) of them are medical abbreviations.

Table 4. Identical and different usage depending on case forms

<table>
<thead>
<tr>
<th>Identical usage</th>
<th>Medical Abbreviation</th>
<th>61% (86 SF)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medical Abbreviation</td>
<td>AAV, ACE, ALPS, ALT, ANT, AST, AVD, BICARB, BS, CBC, CHO, CLA, CPD, CPVT, CT, CTS, DD, DOB, DPT, DRG, DT, EAA, ECG, ENT, FS, GI, GTT, GU, GVHD, HB, HCT, HEP, HS, ICE, IGE, INR, IV, JVP, LAP, LDL, MAC, MALT, MCHC, MD, MHR, MI, MPV, MRI, NEC, NEOPL, OAG, OD, OHI, OP, OPV, OS, OU, PATH, PHOS, PLAT, PLP, PNT, PO, PR, PSA, PVRI, RAIR, RDW, RT, RVOT, SAT, SCN, SMP, SR, TAB, TIBC, TMB, TMP, TPN, TPO, TSH, TSR, VAD, VNS, WBC, WBG</td>
<td></td>
</tr>
<tr>
<td>Medical Concept</td>
<td>7% (10 SF)</td>
<td>ACHE, DIZZY, DROWSY, DRY, GAS, HOT, LIP, LYMPH, PPT, TALC</td>
</tr>
<tr>
<td>Non-medical Usage</td>
<td>12% (17 SF)</td>
<td>AS, CAR, CM, DAD, DIET, FIND, FL, MN, NO, OK, ON, PER, POTTs, RAW, RICH, RM, SN</td>
</tr>
<tr>
<td>Others</td>
<td>1% (1 SF)</td>
<td>IBC</td>
</tr>
<tr>
<td>Different usage</td>
<td>19% (27 SF)</td>
<td></td>
</tr>
<tr>
<td>ADD, AIDS, AME, APR, ASH, CHI, CUP, FISH, IE, MAD, MAHA, MAS, NA, OR, PIE, RF, SM, SOM, ST, STARR, TEA, TIPS, TOP, TUNA, TXT, UNC, US</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* SF = short form

We categorized those 27 short forms with different usages into the following:

- Medical Abbreviation to Non-medical usage (18 cases): Majority of usage change belongs to this category. For “APR”, the major sense is “Abdominoperineal resection” in strict form. However, it changed to “April” in loose form.
- Medical Abbreviation to Medical Concept (3 cases): This category has three short forms “AIDS”, “CHI”, and “TIPS”. For “AIDS”, “Acquired Immunodeficiency Syndrome” in strict form is the major sense whereas “Hearing Aids” is the major sense in loose form.
- Medical Abbreviation to Others (3 cases): There are “AME”, “MAS”, and “SOM”. These short forms contain “Unknown” or “Typo” senses as majority senses in loose form (see discussion later).
- Medical Concept to Non-medical usage: (1 case): A short form “CUP” changes the major sense from “Acetabular Cup Prostheses” (strict from) to “Measurement” (loose form)
- Non-medical usage to Medical Abbreviation: (2 cases): We have two cases, “UNC” and “NA”. “UNC” has the major senses as “University of North Carolina” in strict form and “Intravesical reimplantation of ureter” in loose form. In case of “NA”, it changes the major sense from “Not applicable” (strict form) to “Sodium” (loose form).
Table 5 and Figure 2 illustrate the ambiguity and sense distribution in strict or loose forms. Among 141 short forms, the majority have senses as medical abbreviations (124 in strict and 118 in loose forms). The number of short forms that are ambiguous is 94 and 92 in strict and loose form respectively with 70% of them having senses from both medical and non-medical domains. The short form with the highest number of sense is “MAS” which has sixteen and eight different senses in strict and loose form, respectively. For example, “MAS” has medical abbreviations such as “Macrophage Activation Syndrome”, “Meconium Aspiration Syndrome”, “Multiple autoimmune syndrome”, “Marlo Anatomical Socket”, etc., as well as non-medical usages such as “Medical assistant”, “Initial (abbreviation)”, and “Academic degree”.

Table 5 and Figure 2. The ambiguity and annotated sense distribution for selected short forms

<table>
<thead>
<tr>
<th>Ambiguity</th>
<th>Domain</th>
<th>Strict form</th>
<th>Loose form</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not ambiguous short form</td>
<td>MA</td>
<td>41</td>
<td>40</td>
</tr>
<tr>
<td></td>
<td>MC</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>NM</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>Ambiguous short form</td>
<td>MA</td>
<td>20</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>MC</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>NM</td>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>MA/MC</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>MA/NM</td>
<td>51</td>
<td>53</td>
</tr>
<tr>
<td></td>
<td>MC/NM</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>MA/MC/NM</td>
<td>9</td>
<td>7</td>
</tr>
</tbody>
</table>

* MA = Medical abbreviation, MC = Medical Concept, NM = Non-medical usages

Results of Feature Investigation and Analysis

The number of unique lines in strict and loose form: Among our short forms, “FIND”, “CUP”, “BICARB”, “DROWSY”, “PER”, “DRY”, “ON”, “HOT”, “PHOS”, “DIZZY”, “ACHE”, “NO”, and “AS” have a high difference in frequencies (logarithm difference ≥ 7) between strict and loose forms. Most of these short forms (ten of thirteen) are belonged into medical concept or non-medical usage. Besides, short forms, having identical frequency in both forms such as "OHI", "PVRI", "RVOT", "MPV", and "MCHC", classify into medical abbreviation.

The number of unique lexical variations in strict and loose form: This feature primarily indicates the degree of diverse representations per short form in the clinical corpus. In our investigation, short forms consist of only two (or three) characters have high variations, which may imply the high sense ambiguities. For examples, “BS”, “OP”, “PO”, “US”, “IV”, “NA”, and “ST” have more than twenty lexical variations as loose form. Besides, the short form having exactly same numbers of lexical variants in both forms with low frequency (< 3) tends to have only single sense (i.e., “OHI”, “PVRI”, and “RVOT”). However, this feature may not a key factor to determine the domain usage of short forms.

The number of the candidate appearing as the signal word of a sentence (in loose form): In our sets, “NO” presents the highest frequency (63% of total) as the signal word (the major sense is “Negation” as non-medical usages). Furthermore, the words with high frequency (more than 1.7M) such as “ON”, “AS”, “PER”, and “ST” are classified into non-medical category. On the contrary, short forms which never appear as signal word, such as “AAV”, “CPVT”, “EAA”, “MHR”, “OAG”, “OHI”, “OPV”, “PVRI”, “RVOT”, “SMP”, and “TMB”, serve as medical abbreviations.

The number of distinct normalized bigrams or trigrams when candidate appearing in strict and loose form: If the short forms have the similar frequency of trigrams (or bigrams) in both forms, it serves as medical abbreviation. Our representative short forms are “AAV”, “CPVT”, “EAA”, “IBC”, “MCHC”, “MHR”, “MPV”, “OHI”, “OPV”, “PVRI”, “RVOT”, and “TSR” with less than 10 differences in our corpus. Oppositely, the short forms with the substantial difference may have a high possibility as non-medical usage (i.e., “NO”, “ON”, “OR”, and “AS” have more than 100K difference between two forms). We also find very similar phenomena in bigrams (proceeding word token with short form).

Existence in well-known dictionaries and resources: Table 6 represented the size of lexical forms and coverage of our short forms per resource. Higher numbers of lexicons and coverage ratios of available corpus may not strictly guarantee to distinguish domain usages, since the resources (Allacronyms or Wikipedia) tend to cover both domains.
extensively. As an opposite perspective, Brown and Knuth corpus have low coverage on our short forms in our clinical corpus, but contain well-known lexicons in each domain. For instance, Brown and Knuth cover “CT”, “DIZZY”, “DROWSY”, “LIP”, “LYMPH” (we categorized them as medical usages) as well as “DIET”, “FIND”, “OK”, “PER” (as non-medical usages).

**Table 6.** Total number of lexical form and coverage per resource

<table>
<thead>
<tr>
<th>Resource</th>
<th>Total number of lexicons in resource</th>
<th>Coverage ratio in our short forms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Allacronyms (medical area)</td>
<td>201,456</td>
<td>92.20% (130 short forms)</td>
</tr>
<tr>
<td>Allacronyms (non-medical area)</td>
<td>2,926,554</td>
<td>90.07% (127 short forms)</td>
</tr>
<tr>
<td>Wikipedia (original case)</td>
<td>11,856,369</td>
<td>91.49% (129 short forms)</td>
</tr>
<tr>
<td>Wikipedia (case-insensitive)</td>
<td>12,582,461</td>
<td>97.87% (138 short forms)</td>
</tr>
<tr>
<td>Brown (original case)</td>
<td>50,056</td>
<td>37.59% (53 short forms)</td>
</tr>
<tr>
<td>Brown (case-insensitive)</td>
<td>41,506</td>
<td>37.59% (53 short forms)</td>
</tr>
<tr>
<td>Knuth (original case)</td>
<td>110,573</td>
<td>32.62% (46 short forms)</td>
</tr>
<tr>
<td>Knuth (case-insensitive)</td>
<td>110,573</td>
<td>44.68% (63 short forms)</td>
</tr>
</tbody>
</table>

We selected features with any significant correlation (p < 0.05) to the usages of short forms in Table 7. Taking the difference of logarithmic transformation regarding trigram, bigram, and unique line between two forms have moderate correlations (r > 0.4). Brown/Knuths corpus, logarithmic transformations of signal word and bigram in loose form also have weak correlation (r > 0.3 and r < 0.4) Additionally, logarithmic transformations of trigram and unique lines (loose form) have weak correlations.

**Table 7.** The significant correlation between features and usages of short forms (Pearson correlation, p value <0.05)

<table>
<thead>
<tr>
<th>Features</th>
<th>Pearson coefficient (r)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Difference between logarithmic scales of normalized trigrams in strict and loose forms</td>
<td>0.451</td>
</tr>
<tr>
<td>Difference between logarithmic scales of unique lines in strict and loose forms</td>
<td>0.447</td>
</tr>
<tr>
<td>Difference between logarithmic scales of normalized bigrams in strict and loose forms</td>
<td>0.438</td>
</tr>
<tr>
<td>Existence in Brown (original case and case-insensitive)</td>
<td>0.391</td>
</tr>
<tr>
<td>The logarithmic scale of signal words in loose form</td>
<td>0.380</td>
</tr>
<tr>
<td>Existence in Knuths (original case and case-insensitive)</td>
<td>0.350</td>
</tr>
<tr>
<td>The logarithmic scale of normalized bigrams in loose form</td>
<td>0.304</td>
</tr>
<tr>
<td>The logarithmic scale of normalized trigrams in loose form</td>
<td>0.299</td>
</tr>
<tr>
<td>The logarithmic scale of unique lines in loose form</td>
<td>0.238</td>
</tr>
</tbody>
</table>
Discussion

In this study, we described the process to identify short forms from a large collection of clinical documents and investigated the usages of a randomly sampled subset of those short forms with the goal of creating a clinical sense inventory for short forms. Our analysis indicates that the majority of our short forms (68% of our 141 short forms) served as medical abbreviations or medical concepts with about 19% of them having non-medical usage. The majority of our short forms have multiple meanings. Feature analysis identifies the difference of logarithmic information regarding trigram, bigram, and unique lines between strict and loose forms can differentiate short forms primarily with medical usages from those that are not. It implies that statistical information about local context and word form obtained from a large corpus is useful in disambiguating the usages.

Note that, our institution has been governed the use of approved abbreviations, resulting in relatively low ambiguity. Therefore, 50% of ambiguous short forms have only two or three senses. According to Figure 2, 30% of our ambiguous short forms have two senses, and 20% of them have three senses. This distribution is very similar in both strict and loose forms, but short forms in upper case are frequently represented as medical abbreviation in our corpus. This implies that resolving ambiguity associated with abbreviations is relatively easier for our data.

It is important to note that the quality of our clinical notes is high with low errors or unknown senses given 95% of our clinical notes were generated through dictation followed by manual validation, so our findings may not directly applied to other health institutions. However, we identify the tendency that short forms in loose form format contains twice as many errors but fewer unknown senses compared to those in strict form format. As our samples are extracted to cover as many contextual variations as possible, for high frequent words (e.g., “has”, “some”), their misspellings (e.g., “mas” or “som”) tend to dominate our samples for the corresponding short forms (e.g., “MAS” or “SOM”) where most of those are errors introduced when transforming documents in rich text format (RTF) into plain text format (e.g., “MASS” is inserted with a line break between “MAS” and “S”).

We examined 57 senses that were not covered by the UMLS including 30 medical usages and 27 non-medical usages. Those medical usages include clinical abbreviations (e.g., “Problem-Intervention-Evaluation” for “PIE”, “Maximum Access Surgery” for “MAS”, “Anterior Posterior Repair” for “APR”, “Inversion and Eversion” for “IE”, “cycles per degree” for “CPD”, and “per patient” for “PPT”). Medical devices were also frequently abbreviated (5 cases; e.g., “Intubation Macintosh (curved) blade” for “MAC” and “Intravenous bird cage (filter type)” for “IBC”). There were also some localized medical usages (3 cases; e.g., “Mayo Age Standard” for “MAS”). Additionally, many of the non-medical usages can be localized abbreviations. For example, we have “St Mary’s” for “SM” which is the name of a hospital.

There are some limitations in our study. First, we assumed a short form served as medical usages if it has numerous occurrences with any medical sense in our random 200 samples at Mayo Clinic repository. However, the majority of our short forms were fundamentally used as medical usages since we extracted them from clinical notes. Additionally, we investigated only 200 random unique lines for one short form that may not represent the complete usages distribution of the short form. Secondly, we took the simplest way to acquire our features for exploratory analysis. Therefore, we focused on unique lines without considering any sentence detection. It significantly affected the validation of our signal words in new lines as well as normalized n-grams. Also, we disregarded syntactic or part of speech (PoS) information because we might use the fragments of clinical sentences rather than compete sentences.

In the future, we will further explore those features to achieve automatic distinction between medical and non-medical usages with the large size of short forms. As feature refinements, we will use our resources in divergent views. For example, we identified title pages of Wikipedia covered 91% of our short forms but was not useful for our usage resolutions. However, 57% (80 cases) of them represented as “Categories: Disambiguation” and 17% of them (24 cases) showed “Redirect -> Categories: Disambiguation pages”. This information can be useful to indicate the level of ambiguity for the short form. We will leverage additional well-known resources such as Corpus of Contemporary American English (COCA) and the British National (BNC) corpus. We will also consider the term frequency-inverse document frequency (TF-IDF) scores of short forms on different corpora to weight the token information for distinguishing medical and non-medical usages.
Conclusion

We have investigated the usage information of short forms mined from a large clinical corpus and discovered the prevalent use of short forms as medical abbreviations, but with a high degree of ambiguity and also including both medical and non-medical usages. The study also assessed a set of features to distinguish medical and non-medical usages. The work provides an initial feasibility assessment of creating a sense inventory of short forms and patterns for usage disambiguation.

Acknowledgements

This work was partially supported by NIH R01GM102282 and R01EB19403. The authors also thank David Chen PhD and Liwei Wang MD, PhD for insightful comments.

References

Supervised Learning Methods for Predicting Healthcare Costs: 
Systematic Literature Review and Empirical Evaluation

Mohammad Amin Morid, MS, a Kensaku Kawamoto, MD, PhD, MHS, b 
Travis Ault, c Josette Dorius, c Samir Abdelrahman, MS, PhD b 

a David Eccles School of Business, University of Utah, Salt Lake City, UT, USA 
b Department of Biomedical Informatics, University of Utah, Salt Lake City, UT, USA 
c University of Utah Health Plans, Murray, UT, USA

Abstract
An important informatics tool for controlling healthcare costs is accurately predicting the likely future healthcare costs of individuals. To address this important need, we conducted a systematic literature review and identified five methods for predicting healthcare costs. To enable a direct comparison of these different approaches, we empirically evaluated the predictive performance of each reported approach, as well as other state-of-the-art supervised learning methods, using data from University of Utah Health Plans for October 2013 through October 2016. The data set consisted of approximately 90,000 individuals, 6.3 million medical claims and 1.2 million pharmacy claims. In this comparative analysis, gradient boosting had the best predictive performance overall and for low to medium cost individuals. For high cost individuals, Artificial Neural Network (ANN) and the Ridge regression model, which have not been previously reported for use in healthcare cost prediction, had the highest performance.

Introduction
The United States’ national health expenditure (NHE) grew 5.8% to $3.2 trillion in 2015 (i.e., $9,990 per person), which accounted for 17.8% of the nation’s gross domestic product (GDP). In seeking to control these unsustainable increases in healthcare costs, it is imperative that healthcare organizations can predict the likely future costs of individuals, so that care management resources can be efficiently targeted to those individuals at highest risk of incurring significant costs. Key stakeholders in these efforts to manage healthcare costs include health insurers, employers, society, and increasingly healthcare delivery organizations due to the transition from fee-for-service payment models to value-based payment models. For any given individual, insurers generally have the most comprehensive information on healthcare costs as they pay for care delivered across various healthcare delivery organizations.

Predicting healthcare costs for individuals using accurate prediction models is important for various stakeholders beyond health insurers, and for various purposes. For health insurers and increasingly healthcare delivery systems, accurate forecasts of likely costs can help with general business planning in addition to prioritizing the allocation of scarce care management resources. Moreover, for patients, knowing in advance their likely expenditures for the next year could potentially allow them to choose insurance plans with appropriate deductibles and premiums.

Despite the importance of healthcare cost prediction, to our knowledge there has been no review of the literature on this important topic. Therefore, we conducted a systematic literature review. Moreover, in order to enable a direct comparison of approaches on a common data set, we evaluated each of the identified approaches on a health insurance data set from the University of Utah Health Plans. We also evaluated additional state-of-the-art methods not previously evaluated in the literature.

Methods

Literature Review
Adapting a search strategy from a previous systematic review, we searched Google Scholar and MEDLINE. The latest search was performed on February 21, 2017. We used a combination of the following search terms: healthcare cost prediction, medical claim cost, pharmacy claim cost; healthcare expenditure prediction; healthcare risk score prediction; and patient cost prediction.
In conducting the systematic literature review, we sought to answer the following questions. Because the answer to the first question identified that using features of prior costs to predict future costs performed as well as or better than approaches that also used clinical data for cost prediction purposes, all subsequent questions were focused on approaches that used prior cost features to predict future costs (referred to henceforth as “cost on cost prediction”).

1. What are the types of healthcare cost prediction approaches reported in the literature?
2. What are the input features that have been used for cost on cost prediction?
3. What are the supervised learning methods that have been used for cost on cost prediction?
4. What are the performance measures and evaluation results for cost on cost prediction?

**Direct Comparison of Alternative Cost Prediction Methods using a Health Insurer Data Set**

**Approach.** This study was approved by the University of Utah Institutional Review Board (Protocol # 00094358). We used a health insurance data set to directly compare the performance of cost on cost prediction approaches identified in the literature, as well as other state-of-the-art supervised learning techniques.

**Data.** Our data set consisted of 6.3 million medical claims and 1.2 million pharmacy claims from approximately 91,000 distinct individuals covered by University of Utah Health Plans from October 2013 to October 2016. Available data included demographic information (e.g., age, gender, age), clinical encounter information (e.g., place and date of service, provider information), diagnosis and procedure codes, pharmacy dispense information, and cost information (e.g., paid, allowed and billed amount). This data was filtered to individuals with insurance membership for the whole three years period, which resulted in approximately 3.8 million medical claims and 780,000 pharmacy claims from 24,000 patients.

The data set was divided into two time periods: an observation period and a result period. The former time period was from October 2013 to September 2015 (i.e., two years), which was used to predict individuals’ cost in the result period ranging from October 2015 to October 2016 (i.e., one year). Table 1 shows all input features used in this study. All features used in this study were cost related features extracted from Bertsimas et al.6, which had the largest and most complete set of cost related features among the reviewed manuscripts. If a member did not have any cost for a specific month it was considered as zero; therefore, there are no missing values in this dataset.

The range of paid amounts in the result period showed that 80% of the overall cost of the population came from only 15% of the members. Therefore, aligned with the literature on cost bucketing, to reduce the effects of extremely expensive members, the data set was partitioned into five different cost buckets. This partitioning was done so that the sum of members' costs in each bucket was approximately the same in the observation period (i.e., the total dollar amount in each bucket was the same). For instance, 84% of members are in bucket 1 with the same total cost amount as the members in bucket 5, which contains about 2% of the population.

**Classifier.** Classifiers evaluated included Linear Regression, Lasso7, Ridge8, Elastic Net9, CART10, M511, Random Forest12, Bagging13, Gradient Boosting13, SVM14, and ANN15. Except for CART, the other classifiers had not been previously evaluated for cost on cost prediction. All models were optimized on their parameters to get their best parameter setting on 30 percent of the data set. Models were evaluated with the following parameter settings: number of hidden layers, number of nodes in each layer, learning rate, and momentum were varied for the Neural Network; kernel type along with the corresponding parameters of each kernel type were varied for the Support Vector Machine; minimum split and minimum number of sample in each leaf were varied for the M5 and CART; learning rate and loss function for the Gradient Boosting; and alpha was varied for the Lasso, Ridge and Elastic Net.

A brief description of all the models used in this study (except linear regression) is provided below.

**Lasso:** This is a linear regression model enhanced with variable selection and regularization, which is given by the L1-norm (the loss function is the linear least squares error)7.

**Ridge:** This is a linear regression model where the regularization is given by the L2-norm (the loss function is the linear least squares error). L2-norm equips the model to have non-sparse coefficients, which means many coefficients with zero values or very small values with few large coefficients8.

**Elastic Net:** This is linear regression model that linearly combines the L1-norm and L2-norm penalties of the Lasso and Ridge models9.

**CART:** This is a regression decision tree, where on each node the algorithm chooses the split that minimizes the sum of squared errors for regression of the node. The important quality is that the algorithm uses the sample mean of the instances in each node for regression10.
Table 1. Features used to develop the prediction models.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
<th>Number of features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall_costs</td>
<td>The sum of medical and pharmacy costs</td>
<td>1</td>
</tr>
<tr>
<td>Overall_medical_costs</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Overall_pharmacy_costs</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Six_costs</td>
<td>Overall cost in the last 6 months of the observation period</td>
<td>1</td>
</tr>
<tr>
<td>Three_costs</td>
<td>Overall cost in the last 3 months of the observation period</td>
<td>1</td>
</tr>
<tr>
<td>Trend</td>
<td>Found by fitting a line and extracting the slope through the last monthly costs of the observation period</td>
<td>1</td>
</tr>
<tr>
<td>Acute</td>
<td>An indicator variable found by comparing the highest month with the average monthly cost. If these are significantly different, the indicator takes on the value 1. The idea is that there is a high chance that constantly high cost individuals repeat their cost in the future, while individuals who have had temporarily high cost have a lower chance.</td>
<td>1</td>
</tr>
<tr>
<td>Highest_cost</td>
<td>The cost of the highest month in the observation period</td>
<td>1</td>
</tr>
<tr>
<td>Num_above_average</td>
<td>This variable is calculated as the number of months above average and is an indicator of the shape of the cost profile. If the cost is relatively constant over the period, this variable takes on a value around six, which is an indicator for a chronic cost profile.</td>
<td>1</td>
</tr>
<tr>
<td>Monthly_costs []</td>
<td>Monthly costs of the last twelve months of the observation period (see Data section)</td>
<td>12</td>
</tr>
</tbody>
</table>

M5: Similar to CART, this algorithm is also a regression tree, where a linear regression model is used for building the model and calculating the sum of error as opposed to the mean.  

Random Forest: This is an ensemble learning algorithm that fits a number of regression decision trees on several subsamples of the data. The mean value of the outcomes of the regression tree is generated as the final prediction of the algorithm.  

Support Vector Machine: This is a support vector regression model implemented based on libsvm which uses kernels to find the regression lines.  

Bagging: This is an ensemble learning algorithm that fits each base regression model on random subsets of the data that are generated by a bootstrapping sample method. Aggregation of the individual predictors is performed by averaging to form the final prediction.  

Gradient Boosting: This is an ensemble learning algorithm, where the final model is an ensemble of weak regression decision tree models, which are built in a forward stage-wise fashion. The most important attribute of the algorithm is that it ensembles the models by allowing optimization of an arbitrary loss function. In other words, each regression tree is fitted on the negative gradient of the given loss function, which is set to the least absolute deviation.  

Artificial Neural Network (ANN): This is a large collection of processing units (i.e., neurons), where each unit is connected with many others. Neural networks typically consist of multiple layers and the goal is to solve problems in the same way that the human brain would.  

20-fold cross validation was employed as the evaluation method on 70% of the data set. For statistical significance, we first applied the Friedman’s test to verify differences among multiple classifiers. If significant at an alpha level of 0.05, pairwise comparisons were made with the Wilcoxon Signed-Rank test. This statistical approach was aligned with the method recommended by Demsar.
Results

Literature Review

1. What are the types of healthcare cost prediction approaches reported in the literature?

There are three kinds of methods that have been reported for cost prediction: rule-based, statistical and supervised learning. The disadvantage of the rule based methods (e.g. Kronick et al.15) is that they require a lot of domain knowledge, which is not easily available and is often expensive16. Although statistical models, mainly multiple regression models, are powerful tools for capturing the relationships between the predictors and the dependent variable, they have two important challenges 18. One is that working with several independent variables often causes multicollinearity, which is caused by the presence of significant correlations among predictors. Moreover, their performance is challenged by the skewed nature of healthcare data, where cost data typically feature a spike at zero, distributions are strongly skewed with a heavy right-hand tail19, and extreme values can be present, all of which make them inefficient in small to medium sample sizes if the underlying distribution is not normal. Although several advanced statistical methods have been proposed to accommodate the skewness observed in healthcare data, this type of prediction method is not able to outperform supervised learning methods20. Therefore, this paper is devoted to the use of supervised learning methods for cost prediction, and the remainder of the literature review excludes other types of prediction methods.

There are generally three types of literature that use supervised learning for cost prediction. In the first type, the goal is to predict cost using medical predictors. In this type of literature, the main goal is to show the effect of medical factors such as chronic disease score on cost prediction21. In the second type of literature (which is limited), cost predictors with or without medical predictors are used to predict cost. In the last type of literature, researchers bucket individuals’ costs and predict an individual’s cost bucket rather than his or her actual costs. This last type of research applies nominal predictive models rather than numerical predictive models.

Cost prediction using non-cost predictors. Lee et al.22 provided one of the earliest works on predicting cost by using non-cost predictors. They selected a small sample of 492 patients from a hospital in Korea and compared the performance of ANN and a classification and regression tree for cost prediction. Demographic information, diagnosis codes, number of laboratory tests, the number of admissions and number of operations were the predictors of their analysis. The results showed the superiority of ANN.

Powers et al.23 evaluated several regression statistical modeling approaches for predicting prospective total annual health costs (medical plus pharmacy) of health plan participants using Pharmacy Health Dimensions (PHD), a pharmacy claims-based risk index. Their models included ordinary least squares (OLS) regression, log-transformed OLS regression with smearing estimator, and 3 two-part models using OLS regression, log-OLS regression with smearing estimator, and generalized linear modeling (GLM), respectively. The results showed that most PHD drug categories were significant independent predictors of total costs. The OLS model had the lowest mean absolute prediction error and highest R². The main conclusion was that the PHD system derived solely from pharmacy claims data can be used to predict future total health costs.

Analyzing the impact of multimorbidity (i.e., co-occurrence of more than three chronic disease conditions) on healthcare costs, König et al.21 interviewed 1,050 randomly selected primary care patients aged 65 to 85 years suffering from multimorbidity in Germany. A conditional inference tree algorithm was used as the classifier. The results showed that Parkinson’s disease and cardiac insufficiency were the most influential predictors for total costs, and that the high total costs of Parkinson’s disease were largely due to costs of nursing care.

Cost bucket prediction. Lahiri et al.4 predicted the rise in patient care costs as a binary classification problem. They used a data set with more than 114,000 patients for a span of three years (2008-2010) to investigate which patients experienced increases in inpatient expenditures between 2008 and 2009. Using stacked generalization, they ensembled six classifications algorithms including gradient boosting machine, conditional inference tree, neural networks, SVM, logistic regression and Naive Bayes. This achieved 80% recall, 78% accuracy and 76% precision. One of the contributions of the paper was that they initially had 12,400 features, most of them arising out of diagnosed conditions and drugs taken, and selected 44 of them according to their information gain. This helped the authors to identify major factors which were crucial in determining whether an individual was going to incur higher healthcare expenditure going forward. In a similar study, Guo et al.24 tried to predict patients’ transition from one cost bucket to another bucket in the following year. To do so, they applied multiple methods (each for a single type of transition) to improve the prediction performance. The results showed that they could improve the performance for 21% comparing to baselines. Moreover, they found that the proposed method can help health care entities
achieve efficient resource allocation while improving care quality. Reviewing all papers in this category, we found no studies on categorical cost prediction that used cost-based features as the input.

Cost prediction using cost predictors (cost on cost prediction). Bertsimas et al.\textsuperscript{6} provided one of the first evaluations in the area of health cost prediction using supervised learning techniques. They used a combination of medical, demographic and cost related features from August 2004 to July 2006 as the input and applied regression decision tree and clustering to predict total patient costs in 2007, as measured by insurance payments including medical and pharmacy payments. The results showed that utilizing just 22 cost related features as input and a CART regression decision tree as the classifier gave almost the same performance as adding the medical and demographic information (total of around 1500 features) or applying clustering techniques. Performance was reported in terms of Mean Absolute Error, Hit Ratio, $R^2$ and a penalty based evaluation designed by the authors. Bucketing was also used to evaluate the prediction results to assess the accuracy. This evaluation showed that while the method is strong at predicting low cost buckets, it had a weak performance on higher cost buckets.

Following the above study, Sushmita et al.\textsuperscript{18} evaluated the use of a regression tree, M5 model tree and random forest for cost prediction and showed that M5 had the best performance. The results also confirmed that prior healthcare costs alone can serve as a good indicator for future healthcare costs. To predict patients’ cost for the next year, they used the Medical Expenditure Panel Survey (MEPS) data set coming from responses to panel surveys given to households and their employers, medical providers, and insurance providers over two year periods.

Duncan et al.\textsuperscript{2} compared several different supervised learning and statistical models to predict patients’ cost including M5, Lasso and boosted trees. They applied their experiments on 30,000 patients where the information from 2008 was used for training and the total allowed amounts in the claims from 2009 were used for testing. They involved a variety of predictors as input including the previous year’s total cost, total medical cost, total pharmacy cost, demographic information, total visits and chronic conditions (83 different conditions). The results showed that boosted trees and M5 were the most effective classifier in terms of $R^2$ and Mean Absolute Error (MAE) respectively, and that cost predictors were the strongest predictors. Moreover, confirming previous literature results, this paper showed that statistical methods are not as good as supervised learning techniques.

Kuo et al.\textsuperscript{25} attempted to show the significance of pharmacy-based metrics as opposed to diagnosis-based morbidity measures in predicting patients’ costs and outpatient visits. They used data from 2006 to predict patients’ billed costs in 2007. To achieve this, they applied linear regression on the data set. Evaluation was done based on Mean Absolute Error and $R^2$. Although the purpose of the study was to explore the capability of the pharmacy-based metric in cost prediction, the results confirmed that using cost based features for cost prediction has almost the same accuracy as adding other types of features to the input. This paper did not incorporate sophisticated cost features and just used a single cost feature from 2006. Frees et al.\textsuperscript{26} studied the ability of linear regression to predict individuals’ costs in terms of healthcare insurance payments. They used self-rated physical health and self-rated mental health, provided by participants, using demographic and survey-based information as their input. Getting a reasonable performance (i.e., $R^2=0.27$), they found that cost, self-rated mental health and self-rated physical health are the most important predictors.

Collectively, and in particular in the study by Bertsimas et al.\textsuperscript{6}, these studies found that cost on cost prediction can match the performance of predictions made using clinical input factors or clinical plus cost input factors.

2. What are the input features that have been used for cost on cost prediction?

Input features are one of the essential parts of a supervised learning task. Numeric cost prediction studies have benefited from a variety of features as input, which are summarized in Table 2. As seen, Bertsimas et al.\textsuperscript{6} evaluated a wide range of cost inputs and reported the performance of cost inputs separately. Their results showed that prediction using a superset of 1542 features, including clinical features, had the same performance as using just the 21 cost predictors. This finding was confirmed by other researchers in subsequent work.\textsuperscript{5,18}

3. What are the supervised learning methods that have been used for cost on cost prediction?

There are a variety of supervised learning methods that have been used in this area. Table 3 summarizes all different methods that have been reported as successful methods for cost on cost prediction. These methods include Lasso, which is a type of linear regression, gradient boosting on regression decision trees, M5 regression decision tree, random forest, linear regression and CART regression tree. Table 3 also shows the target type of the cost that was studied in each paper. Billed amount is the total amount that is charged by the health care provider and the paid amount is the amount that is paid by the insurance company.
Table 2. Input features used for cost on cost prediction in the literature

<table>
<thead>
<tr>
<th>Paper</th>
<th>Number of Cost Inputs</th>
<th>Cost Inputs</th>
<th>Non Cost Inputs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bertsimas (2008)</td>
<td>21</td>
<td>Monthly cost (12), Total pharmacy cost, Total medical cost, Total cost,</td>
<td>Age, Sex, Diagnosis groups, Count of claims with diagnosis codes from each group,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Total cost in last 6 months, Total cost in last 3 months, Trend, Acute,</td>
<td>Procedure groups, Drug groups, Count of members' diagnoses, procedures, Drugs,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Months above average, Cost of highest month</td>
<td>Gender, Age</td>
</tr>
<tr>
<td>Duncan (2016)</td>
<td>4</td>
<td>Professional costs, Pharmacy costs, Outpatient costs, Inpatient costs</td>
<td>Age, Sex, Diagnose codes grouped into existing condition categories, Total visit</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>count, Hospital admission count, Primary care provider visits count</td>
</tr>
<tr>
<td>Sushmita (2015)</td>
<td>1</td>
<td>Total previous cost</td>
<td>Age, Sex, Diagnosis Groups, Procedure groups, Comorbidity scores</td>
</tr>
<tr>
<td>Kuo (2011)</td>
<td>1</td>
<td>Previous medication cost</td>
<td>Age, Sex, Elixhauser’s index, Pharmacy-based metrics</td>
</tr>
<tr>
<td>Frees (2013)</td>
<td>1</td>
<td>Total previous cost</td>
<td>Sex, Race, Region, Education, Job, Marriage, Income level, Self-rated physical</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>health, Self-rated mental health</td>
</tr>
</tbody>
</table>

Table 3. Supervised learning methods used for cost on cost prediction in literature

<table>
<thead>
<tr>
<th>Paper</th>
<th>Method</th>
<th>Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Duncan (2016)</td>
<td>Gradient Boosting DT, Lasso, M5</td>
<td>Paid amount</td>
</tr>
<tr>
<td>Sushmita (2015)</td>
<td>M5, RandomForest, CART</td>
<td>Billed amount</td>
</tr>
<tr>
<td>Frees (2013)</td>
<td>Linear regression</td>
<td>Paid amount</td>
</tr>
<tr>
<td>Kuo (2011)</td>
<td>Linear regression</td>
<td>Billed amount</td>
</tr>
<tr>
<td>Bertsimas (2008)</td>
<td>CART</td>
<td>Paid amount</td>
</tr>
</tbody>
</table>

4. What are the performance measures and evaluation results for cost on cost prediction?

**MAE:** This shows the average error of the model on prediction of the actual cost values and is calculated as follows:

$$MAE = \frac{\sum |a_i - p_i|}{n}$$

where $a_i$ and $p_i$ are the actual and predicted costs of member $i$ in the result period respectively.

**Mean absolute percentage error (MAPE)**: This is a modified version of absolute error in which the MAE is divided by the mean of the cost, so that the MAE could be compared across the models with different means of cost:

$$MAPE = \frac{\sum |a_i - p_i|}{\frac{n}{\bar{a}}}$$
MAE is dependent on the data set, such that different models from different studies cannot be directly compared using that measure. MAPE is a relative measure and does not have this limitation.

$R^2$: This shows the Pearson correlation between the actual and predicted cost values:

$$R^2 = 1 - \frac{\sum_i(a_i - p_i)^2}{\sum_i(a_i - \bar{a})^2}.$$  

*Hit Ratio*: This measure shows the percentage of the members for whom a model forecasts the correct cost bucket:

$$\text{Hit Ratio} = \frac{\text{Number of members with correct predicted bucket}}{\text{Total number of members}}$$

*Penalty Error*: This is a performance measure for cost prediction based on domain knowledge. Penalty error penalizes models for underestimating high cost members more than overestimating low cost members, which is motivated by the estimated opportunity loss. Table 4 shows the penalty table for the five-cost-bucket scheme. The final value of the penalty error is calculated from the average forecast penalty per member of a given sample.

Table 5 summarizes the evaluation measures used in different papers. The reported performance measures in this table correspond to the whole data set used in each study. This study reports the experimental results in terms of all five performance measures except MAE, which is not reported given the sensitivity of absolute cost data.

**Direct Comparison of Alternative Cost Prediction Methods using a Health Insurer Data Set**

Tables 6 to 9 show the performance comparison between different supervised learning models on training and validation data sets. As seen, Gradient Boosting had the highest performance in terms of all measures in all buckets except bucket five. Here, ANN was superior. Also, the Ridge model showed a comparable performance compared to ANN, especially for low cost buckets.

### Table 4. Penalty table based on the predicted and actual cost buckets

<table>
<thead>
<tr>
<th>Predicted Bucket</th>
<th>Actual Bucket</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
</tr>
</tbody>
</table>

### Table 5. Performance measures and outcome for cost on cost prediction in literature

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>MAE (S)</td>
<td>2,214</td>
<td>3,104</td>
<td>8,112</td>
<td>507</td>
<td>2,705</td>
</tr>
<tr>
<td>MAPE</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.16</td>
<td>0.20</td>
<td>-</td>
<td>0.75</td>
<td>5.25</td>
</tr>
<tr>
<td>Hit Ratio</td>
<td>84.6</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Penalty Error</td>
<td>0.38</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

1311
Table 6. Performance comparison among different supervised learning models for numeric measures on the training data set. Models that are annotated with ($) have been used in the cost on cost prediction literature before (see Table 3), while those annotated with (n) are new to this study.

<table>
<thead>
<tr>
<th>Method</th>
<th>MAPE</th>
<th>R²</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>All 1 2 3 4 5</td>
<td>All 1 2 3 4 5</td>
</tr>
<tr>
<td>Gradient Boosting ($)</td>
<td>0.62 0.74 0.61 0.58 0.53</td>
<td>0.48 0.07 0.14 0.18 0.16</td>
</tr>
<tr>
<td>ANN (n)</td>
<td>0.67 0.83 0.67 0.64 0.50 0.42</td>
<td>0.46 0.04 0.10 0.14 0.27 0.46</td>
</tr>
<tr>
<td>Ridge ($)</td>
<td>0.69 0.83 0.67 0.65 0.50 0.41</td>
<td>0.44 0.04 0.09 0.14 0.29 0.45</td>
</tr>
<tr>
<td>SVM (n)</td>
<td>0.75 1.01 0.70 0.66 0.56 0.50</td>
<td>0.43 0.04 0.09 0.13 0.22 0.37</td>
</tr>
<tr>
<td>Elastic Net (n)</td>
<td>0.77 1.01 0.70 0.67 0.58 0.50</td>
<td>0.42 0.04 0.09 0.13 0.19 0.33</td>
</tr>
<tr>
<td>Lasso ($)</td>
<td>0.80 1.13 0.71 0.67 0.58 0.51</td>
<td>0.42 0.04 0.08 0.13 0.19 0.34</td>
</tr>
<tr>
<td>M5 ($)</td>
<td>0.80 1.13 0.71 0.68 0.56 0.51</td>
<td>0.42 0.04 0.08 0.13 0.19 0.33</td>
</tr>
<tr>
<td>Linear Regression ($)</td>
<td>0.80 1.14 0.72 0.67 0.58 0.51</td>
<td>0.42 0.04 0.08 0.13 0.18 0.34</td>
</tr>
<tr>
<td>Random Forest ($)</td>
<td>0.90 1.14 0.87 0.74 0.73 0.55</td>
<td>0.41 0.03 0.06 0.14 0.07 0.37</td>
</tr>
<tr>
<td>Bagging (n)</td>
<td>0.90 1.14 0.85 0.77 0.65 0.57</td>
<td>0.40 0.02 0.06 0.10 0.08 0.36</td>
</tr>
<tr>
<td>CART ($)</td>
<td>0.95 1.17 0.94 0.80 0.74 0.62</td>
<td>0.32 0.02 0.05 0.04 0.05 0.21</td>
</tr>
</tbody>
</table>

Table 7. Performance comparison among different supervised learning models for categorial measures on the training data set. Models that are annotated with ($) have been used in the cost on cost prediction literature before (see Table 3), while those annotated with (n) are new to this study.

<table>
<thead>
<tr>
<th>Method</th>
<th>Hit Ratio (%)</th>
<th>Penalty Error</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>All 1 2 3 4 5</td>
<td>All 1 2 3 4 5</td>
</tr>
<tr>
<td>Gradient Boosting ($)</td>
<td>94.8 98.7 74.3 64.9 52.7 37.5</td>
<td>0.17 0.10 0.63 0.89 1.07 1.10</td>
</tr>
<tr>
<td>ANN (n)</td>
<td>91.5 96.1 71.8 56.8 56.9 51.8</td>
<td>0.20 0.13 0.67 0.95 0.97 0.96</td>
</tr>
<tr>
<td>Ridge ($)</td>
<td>91.6 94.9 71.5 58.9 55.8 51.9</td>
<td>0.20 0.13 0.69 0.91 0.97 0.95</td>
</tr>
<tr>
<td>SVM (n)</td>
<td>91.1 94.7 69.7 57.9 54.2 50.7</td>
<td>0.20 0.13 0.67 0.90 1.06 1.09</td>
</tr>
<tr>
<td>Elastic Net (n)</td>
<td>90.9 94.7 68.9 57.1 52.3 49.9</td>
<td>0.20 0.13 0.67 0.91 1.11 1.14</td>
</tr>
<tr>
<td>Lasso ($)</td>
<td>90.8 94.7 68.5 57.1 51.9 49.7</td>
<td>0.20 0.13 0.67 0.91 1.10 1.15</td>
</tr>
<tr>
<td>M5 ($)</td>
<td>90.6 94.6 67.9 56.3 51.9 49.7</td>
<td>0.20 0.13 0.67 0.91 1.10 1.15</td>
</tr>
<tr>
<td>Linear Regression ($)</td>
<td>90.1 94.2 67.9 56.1 51.9 47.5</td>
<td>0.20 0.13 0.66 0.91 1.10 1.14</td>
</tr>
<tr>
<td>Random Forest ($)</td>
<td>88.8 93.7 62.9 48.9 50.9 46.6</td>
<td>0.23 0.15 0.71 1.05 1.08 1.10</td>
</tr>
<tr>
<td>Bagging (n)</td>
<td>86.7 93.6 62.7 52.1 47.6 43.6</td>
<td>0.23 0.15 0.71 0.98 1.11 1.15</td>
</tr>
<tr>
<td>CART ($)</td>
<td>86.1 93.2 62.5 53.0 44.4 41.6</td>
<td>0.25 0.16 0.82 1.05 1.19 1.31</td>
</tr>
</tbody>
</table>

Table 8. Performance comparison among different supervised learning models for numeric measures on the validation data set. Models that are annotated with ($) have been used in the cost on cost prediction literature before (see Table 3), while those annotated with (n) are new to this study.

<table>
<thead>
<tr>
<th>Method</th>
<th>MAPE</th>
<th>R²</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>All 1 2 3 4 5</td>
<td>All 1 2 3 4 5</td>
</tr>
<tr>
<td>Gradient Boosting ($)</td>
<td>0.65 0.76 0.63 0.60 0.59 0.54</td>
<td>0.46 0.04 0.11 0.15 0.13 0.32</td>
</tr>
<tr>
<td>ANN (n)</td>
<td>0.7 0.84 0.69 0.66 0.52 0.45</td>
<td>0.44 0.02 0.07 0.11 0.25 0.44</td>
</tr>
<tr>
<td>Ridge ($)</td>
<td>0.71 0.85 0.70 0.67 0.51 0.44</td>
<td>0.41 0.02 0.07 0.11 0.27 0.43</td>
</tr>
<tr>
<td>SVM (n)</td>
<td>0.78 1.00 0.72 0.68 0.58 0.52</td>
<td>0.41 0.02 0.07 0.12 0.20 0.36</td>
</tr>
<tr>
<td>Elastic Net (n)</td>
<td>0.8 1.06 0.73 0.68 0.60 0.53</td>
<td>0.40 0.02 0.07 0.12 0.16 0.30</td>
</tr>
<tr>
<td>Lasso ($)</td>
<td>0.83 1.14 0.74 0.68 0.60 0.53</td>
<td>0.40 0.02 0.07 0.12 0.16 0.30</td>
</tr>
<tr>
<td>M5 ($)</td>
<td>0.83 1.15 0.74 0.69 0.58 0.55</td>
<td>0.40 0.02 0.07 0.12 0.16 0.31</td>
</tr>
<tr>
<td>Linear Regression ($)</td>
<td>0.83 1.16 0.74 0.68 0.60 0.53</td>
<td>0.40 0.02 0.07 0.12 0.16 0.31</td>
</tr>
<tr>
<td>Random Forest ($)</td>
<td>0.91 1.17 0.9 0.77 0.75 0.58</td>
<td>0.40 0.02 0.05 0.13 0.08 0.34</td>
</tr>
<tr>
<td>Bagging (n)</td>
<td>0.9 1.16 0.88 0.80 0.68 0.55</td>
<td>0.39 0.01 0.04 0.09 0.09 0.34</td>
</tr>
<tr>
<td>CART ($)</td>
<td>0.98 1.23 1.01 0.83 0.77 0.66</td>
<td>0.29 0.01 0.03 0.02 0.03 0.18</td>
</tr>
</tbody>
</table>
Table 9. Performance comparison among different supervised learning models for categorical measures on the validation data set. Models that are annotated with (l) have been used in the cost on cost prediction literature before (see Table 3), while those annotated with (n) are new to this study.

<table>
<thead>
<tr>
<th>Method</th>
<th>Hit Ratio (%)</th>
<th>Penalty Error</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>All</td>
<td>1</td>
</tr>
<tr>
<td>Gradient Boosting (l)</td>
<td>92.9</td>
<td>96.4</td>
</tr>
<tr>
<td>ANN (n)</td>
<td>89.2</td>
<td>94.0</td>
</tr>
<tr>
<td>Ridge (n)</td>
<td>89.1</td>
<td>93.9</td>
</tr>
<tr>
<td>SVM (n)</td>
<td>88.9</td>
<td>93.7</td>
</tr>
<tr>
<td>Elastic Net (n)</td>
<td>88.5</td>
<td>93.5</td>
</tr>
<tr>
<td>Lasso (l)</td>
<td>88.4</td>
<td>93.4</td>
</tr>
<tr>
<td>M5 (l)</td>
<td>88.4</td>
<td>93.4</td>
</tr>
<tr>
<td>Linear Regression (l)</td>
<td>88.4</td>
<td>93.4</td>
</tr>
<tr>
<td>Random Forest (l)</td>
<td>85.6</td>
<td>90.9</td>
</tr>
<tr>
<td>Bagging (n)</td>
<td>85.6</td>
<td>90.9</td>
</tr>
<tr>
<td>CART (l)</td>
<td>85.3</td>
<td>90.8</td>
</tr>
</tbody>
</table>

Discussion

Summary of findings. This study reviewed the literature of healthcare cost prediction and found that cost on cost prediction performs as well or better than cost prediction using clinical data or clinical data plus cost data. Moreover, supervised learning methods were found to be superior in predictive ability. Moreover, we found that gradient boosting provides the best cost on cost prediction models in general, with ANN providing superior performance for higher cost patients. The evaluations show consistency between training and validation results.

Strengths. An important strength of this study is that we combined both a systematic literature review and a head-to-head empirical evaluation of different supervised learning methods reported in the literature. An additional strength is that we evaluated state-of-the-art supervised learning methods not previously evaluated in the literature for cost on cost prediction in health care.

Limitations. The main limitation of this study is that we used one data set. More experiments on different data sets from different institutions and regions could provide more solid evidence on the comparative performance of different algorithms. The second limitation of this study is that we used cost features. Although previous studies showed that medical features did not improve the performance of the cost models, we could potentially still benefit from such features for two reasons. One is that the new supervised machine learning methods may benefit from the medical features. Second is that the medical features have more explanatory power that may help decision makers understand the root causes of members’ costs.

Future studies. This study was devoted to the paid amount of the medical claims. An interesting venue of research would be analyzing the billed amount as well as the out-of-pocket amount paid by patients to see which approaches work best for each type of cost metric. Another future research direction would be to explore the use of more advanced supervised learning methods such as deep learning and structure analysis to improve the performance of cost prediction methods. Finally, adding medical features and benefiting from their predictive and explanatory power can be another future research direction, which has already been started in our team.

Conclusion

The literature indicates that the preferred approach to healthcare cost prediction is cost on cost prediction using supervised learning methods. Empirical analysis of alternate approaches using data from a single health insurer found that gradient boosting provides the best cost on cost prediction models in general, with ANN providing superior performance for higher cost patients.
References


The Role of Surface, Semantic and Grammatical Features on Simplification of Spanish Medical Texts: A User Study

Partha Mukherjee, Ph.D1, Gondy Leroy, Ph.D1, David Kauchak, Ph.D2, Brianda Armenta Navarrete, BS1, Damian Y. Diaz, MA1, Sonia Colina, Ph.D1
1University of Arizona, Tucson, AZ; 2Pomona College, Claremont, CA

Abstract

Simplifying medical texts facilitates readability and comprehension. While most simplification work focuses on English, we investigate whether features important for simplifying English text are similarly helpful for simplifying Spanish text. We conducted a user study on 15 Spanish medical texts using Amazon Mechanical Turk and measured perceived and actual difficulty. Using the median of the difficulty scores, we split the texts into easy and difficult groups and extracted 10 surface, 2 semantic and 4 grammatical features. Using t-tests, we identified those features that significantly distinguish easy text from difficult text in Spanish and compare with prior work in English. We found that easy Spanish texts use more repeated words and adverbs, less negations and more familiar words, similar to English. Also like English, difficult Spanish texts use more nouns and adjectives. However in contrast to English, easier Spanish texts contained longer sentences and used grammatical structures that were more varied.

Introduction

Providing simplified text to patients and health information consumers facilitates understanding. Using clear and understandable language is especially relevant in the medical domain as better text comprehension helps create a health-literate patient group. Easier texts help patients remember medical information1 and motivates them to read and understand the text2. Well informed patients can manage their conditions better.

Readers’ education level, language skills and reasoning abilities play a role in their ability to understand, digest and act on information3. Studies that evaluated readability grade levels, presentation and format, and education levels4, 5 show that the education level influences understanding. Demographic differences also impact patient understanding of health education and instructions6; there are significant differences in health literacy with respect to age, gender, academic background and household income, with higher educational attainment and higher household income associated with more adequate health literacy. Advanced age had a negative correlation with adequate health literacy.

There exist several approaches for improving patient health literacy7, however, providing text-based materials is one of the most cost- and time-efficient solutions. Albright et al.8 recommend providing patients and caretakers with written information as it is vital in reinforcing verbal communication between patients and doctors. However, written materials for patient education must be carefully matched to patient reading levels9. Different text features affect text difficulty10 and different approaches have been suggested to increase comprehension. For example, studies advise replacing technical jargon with more common terms or improving the organization of the text11: avoiding long sentences, keeping length to no more than eight to ten words12; organizing the content in question and answer format; and creating a logical flow of information13. Interdisciplinary team collaboration to write and provide information for patients in plain language has also been suggested14. Finally, the use of information technology tools has been advocated to assist with this process and improve the understanding of clinical information and data15, 16.

In this study, we examine how different text features interact with text difficulty in Spanish medical texts. Most previous studies have focused on English texts, even though Spanish is the second most used language in the United States17. We leverage our previous work on English features to motivate our feature selection and to help understand how Spanish text simplification might differ from English simplification. Identifying parallel simplification approaches is important for our final goal, which is the development of a multi-lingual text simplification toolkit. To evaluate our set of potentially interesting features, we conducted a user study to estimate the difficulty of texts. We separated the texts into ‘easy’ and ‘difficult’ based on perceived and actual difficulty scores. We then examined feature occurrence in the texts and report on those that were significantly different between the easy and difficult texts. We also compare the Spanish text features with those found important for English text simplification.
Literature Review

Text Simplification

Text simplification is a procedure that transforms text into a clearer and more understandable version while preserving the meaning of the text\(^8\). Simplifying texts can facilitate readability for people with disabilities\(^9,10,11\), low literacy\(^12,13\), non-native backgrounds\(^14,15\) or non-expert knowledge\(^16,17\). Text simplification may also help improve the performance of many natural language processing (NLP) tasks, such as parsing\(^18\), summarization\(^19,20\), semantic role labeling\(^21\) and machine translation\(^22\).

Formulas to measure text readability include Reading Ease Score (RES), SMOG (Simple Measure of Gobbledygook), Flesh-Kincaid score and Gunning Fog index\(^23\). However, these formulas rely on simple, surface-level features such as sentence length, percentage of familiar words and word length and, although they have been used in readability research over the past two decades\(^24\), there is little evidence that there is a relationship with actual text difficulty.

We have developed a new feature to measure text difficulty, term familiarity\(^2\), which can be used to both measure text difficulty and guide text simplification. Term familiarity is measured based on a term/word’s frequency in a large corpus, with low frequency terms being more difficult. A text can be simplified by replacing difficult words with synonyms that are less difficult, as measured by term familiarity\(^25,26\). Similar to term familiarity, we developed grammar familiarity as a new feature, measured by the frequency of the parse tree structure of the sentence in a large corpus. Grammar familiarity has a similar effect to term familiarity: more commonly used grammatical structures (i.e. higher frequency in the corpus) are considered more familiar and have been shown to be easier to understand\(^27\).

We have also demonstrated that different types of negation are used in easy and difficult texts: morphological negation is used more frequently in difficult text\(^28\). Furthermore, we have shown that splitting long noun phrases is beneficial for improving perceived difficulty, but only affects actual difficulty when it can be done while maintaining the natural flow of the sentence\(^29\).

Even though Spanish is one of the most common languages in the US, there is little information about how to determine text difficulty and simplify text in Spanish. Even traditional readability formulas do not entirely fit Spanish text\(^30\). Some previous Spanish studies have focused on structural simplification such as sentence splitting, lexical substitution of functional multi-word units\(^31\) and re-ordering of syntactic units\(^32\). For example, the Simplext project focuses on simplifying vocabulary and syntactic structures\(^33\) for people with cognitive disabilities. Others focus on different aspects of the text such as the treatment of idioms and collocations\(^34\), the way that the syntactic structure looks\(^35\), and the font and colors\(^36\).

Actual and Perceived Difficulty

Two components play a role in the difficulty of a text: 1) whether the text looks easy for readers (perceived difficulty) and 2) whether the reader can understand it (actual difficulty). Perceived and actual difficulty independently influence intention to read and the comprehension of materials\(^37\). Perceived difficulty is often measured on a 5-point Likert scale by asking participants how difficult the text ‘looks’. Actual difficulty requires measuring understanding and has been accomplished with cloze tests, multiple-choice tests and other measures of comprehension\(^38\). A text conveys information effectively when people both want to read it and can understand the content. The willingness to read is related to a text’s perceived difficulty, while comprehension is related to its actual difficulty and specific characteristics. Previously we showed that lexical simplification reduces perceived difficulty, while coherence enrichment reduces actual difficulty\(^39\).

Evidence for distinguishing between perceived and actual difficulty comes from two research models. The Health Belief Model (HBM) explains and predicts health-related behaviors. It is defined in terms of perception of four constructs: 1) perceived susceptibility, 2) perceived severity, 3) perceived benefits, and 4) perceived barriers. These concepts were proposed to account for people’s ‘readiness to act’\(^40\). It was assumed that diverse demographic, sociopsychological and structural variables might, in any given instance, affect an individual’s perception and thus indirectly influence health-related behavior. The second model is the Theory of Planned Behavior (TPB), which is an extension of the Theory of Reasoned Action (TRA)\(^41\). TPB is used to explain deliberate and planned behavior. According to this theory, human action is guided by three kinds of considerations: 1) behavioral beliefs, 2) normative beliefs and 3) control beliefs. Human actions are guided by three concepts that particularly impact willingness to change behavior. The general rule is that the greater the perceived control, the stronger a person’s intention to perform the behavior in question.
Both models highlight the importance of perceived text difficulty. Many readers do not read a text if they feel that it is difficult\textsuperscript{48}. Naturally, perceived difficulty does not tell the entire story. It is insufficient for a text to be perceived as easy; actual difficulty plays a major role in the resulting comprehension of information. We want to both reduce the perceived difficulty and increase text comprehension. Adjusting text based on reading levels will encourage people to read more\textsuperscript{45}.

**Methods**

To evaluate which text features are important for simplifying Spanish text, we conducted a user study to measure both perceived and actual difficulty of the texts. We then split the texts into easy and difficult based on the perceived and actual difficulty scores and evaluated which text features are indicative of difficult and easy texts.

**Text and Content Questions**

To analyze the characteristics of medical texts, we created a corpus with 15 texts. The texts are paragraphs from abstracts of scientific articles related to seven different diseases that are the most frequent causes of death in the world\textsuperscript{49}: asthma, cancer, diabetes, hepatitis, hypertension, influenza, leukemia. We chose scientific articles since the content is more difficult to understand\textsuperscript{50}. Each disease was discussed in at least one text, with some diseases represented in as many as six.

For each text, we generated four multiple-choice questions. The first three multiple-choice questions tested content knowledge in the article and participants had to select from four answer options. The fourth multiple-choice question is a conclusion question that asked participants to select the sentence that would best finish the text. We removed the actual last sentence of the text and included this as an option, along with three other incorrect options. We used Amazon Mechanical Turk (MTurk) to recruit participants, with the restrictions that they were US residents and had a 95% approval rate on tasks previously performed for other requesters.

After the study, we pre-processed the text and extracted the different features. The word parts-of-speech (POS) were identified using the Freeling parser\textsuperscript{51} and the full parse trees were generated using the Stanford parser\textsuperscript{52}. Word frequencies were looked up in the LEXESP\textsuperscript{53} corpus which contains approximately 120,000 words. The frequency of words not found in LEXESP was assumed to be zero.

**Metrics**

To measure perceived difficulty, we used a 5-point Likert scale and asked people “How difficult is the text to read?” The answers were quantified using a 0-4 range, with 0 representing the most difficult and 4 representing the least difficult.

To measure actual difficulty, we calculated the average accuracy of the answers to the multiple-choice questions.

**Features**

We identified surface, semantic and grammatical features motivated by prior work in English\textsuperscript{37, 54}. The surface and semantic features are averaged over the number of words per text, while the grammatical features are aggregated over the number of sentences per text. “Proportion” below denotes the number of words that have that feature (e.g. words that are nouns) divided by the total number of words in a text.

- **Surface features** (10):
  - Number of sentences.
  - Average words per sentence.
  - Function word ratio: proportion of words in the text that are function words (determiners, prepositions and conjunctions).
  - Punctuation ratio: proportion of words in the text that are punctuation characters.
  - Number ratio: proportion of words in the text that are numerals.
  - Negation ratio: proportion of words in the text that are negation words. The negation words are identified by the Freeling parser.
  - Noun/adjective/verb/adverb ratio: proportion of words in the text that are nouns/adjectives/verbs/adverbs in the text, each as a separate feature.

- **Semantic features** (2):
  - Repeated words ratio: proportion of the words in the text that are repeated.
  - Term familiarity: We use the frequency of content bearing terms (i.e., nouns, adjectives, verbs, adverbs) in the LEXESP\textsuperscript{53} database as an approximation of term familiarity.
Grammatical features (4): We denote the “grammar structure” of the sentence as the top two levels of the parse tree of a sentence.37

- Average grammar frequency: Number of different sentence grammar structures in a text, averaged over the texts.
- Average grammar question frequency: Average number of different grammar structures for the multiple-choice question per text.
- Average minimum edit distance: Edit distance is the minimum number of edit operations (deletions, substitutions and insertions) to transform one grammar structure into another. We calculate the edit distances of the grammar structures for all adjacent sentence pairs in the text and calculate the average of these values. This measure quantifies how dissimilar (i.e., distant) two grammar strings are, so lower values indicate more similarity. We consider this measure complementary to the average grammar frequency since it also measures grammatical diversity in a text.
- Average cosine similarity: We created frequency vectors for each of the constituents in a grammar structure. We then calculated the cosine similarity between these vectors for all adjacent sentences and averaged these values per text.

Participants

A total of 66 workers participated in our study with 12 participants per text on average. They were paid $1.50 for completing each text and the associated questions. As is customary with MTurk studies, each participant could choose to complete from 1 to all 15 texts. Overall, an average of 3.9 minutes was needed to read the texts and answer the associated questions. The shortest time spent to complete reading the text and answering the questions was 1.4 minutes and the longest was 14 minutes.

Table 1 provides the participants’ demographic information as self-identified by the participants. Most participants (95%) were less than 50 years old. The majority were with 1) gender: female (52%), 2) race: white (80%), and 3) ethnicity: Hispanic or Latino (70%). Most had moderate education: 20% have a high school diploma, 26% an associate’s degree and 36% a bachelor’s degree. Thirty-nine percent of the participants spoke half Spanish at home and 35% mostly Spanish.

Results

Perceived and Actual Difficulty

To analyze the text features, the 15 texts were classified into three categories, easy, middle, and difficult using the median perceived and actual difficulty scores which are calculated based on the participant responses to the questions. Texts with scores higher than the median are considered easy; texts with scores less than the median are difficult. The texts with scores that were nearly the same as the median (i.e., median ± 0.05) were considered in the middle category. To get a clear distinction between ‘easy’ and ‘difficult’ texts, we excluded the middle category from our analysis, which included two texts. Table 2 shows the descriptive statistics for the easy and difficult texts: 6 texts were easy and 7 texts were difficult based on perceived difficulty and 7 texts were easy and 6 texts difficult based on actual difficulty. A t-test confirmed that the easy and difficult splits were significantly different. For the perceived difficulty split, the easy texts had significantly lower perceived difficulty scores with an average difference of almost a full point (p < 0.001). For the actual difficulty split, the easy texts had significantly better accuracy with the easy text resulting in question scores that were 23% (absolute) better (p < 0.001).

Example of sentences from easy and difficult texts are:

1. Perceived difficulty
   - Easy: “El manejo de la diabetes tipo 1 en la infancia y adolescencia ha evolucionado en los últimos años, como consecuencia de la intensificación del tratamiento insulínico y la aceptación de nuevos objetivos en el tratamiento.”
   - Difficult: “La primeroinfección por los virus herpes simple (VHS), varicela-zóster (VVZ), citomegalovirus (CMV), herpesvirus humano 6 y virus de Epstein-Barr (VEB) ocasiona hepatitis generalmente leve y autolimitada en pacientes inmunocompetentes.”

2. Actual difficulty

---

• Easy: “Aproximadamente el 75% de las mujeres con cáncer de mama avanzado presenta metástasis óseas, lo que les ocasiona una importante morbilidad y deterioro en su calidad de vida.”
• Difficult: “Con la creciente posibilidad de ocurrencia de una pandemia de influenza en las próximas décadas, es necesario que los niveles subnacionales de gobierno (estados, provincias, municipios) estén adecuadamente preparados con planes operativos basados en los lineamientos generales establecidos por organizaciones multinacionales o gobiernos nacionales.”

**Table 1.** Demographic information (N=66)

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>N (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td></td>
</tr>
<tr>
<td>Less than 30 years</td>
<td>30 (45.4)</td>
</tr>
<tr>
<td>31 – 40 years</td>
<td>16 (24.2)</td>
</tr>
<tr>
<td>41 – 50 years</td>
<td>17 (25.7)</td>
</tr>
<tr>
<td>51 – 60 years</td>
<td>2 (3)</td>
</tr>
<tr>
<td>61 – 70 years</td>
<td>1 (1.5)</td>
</tr>
<tr>
<td>Greater than 70 years</td>
<td>-</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>32 (48.5)</td>
</tr>
<tr>
<td>Female</td>
<td>34 (51.5)</td>
</tr>
<tr>
<td>Race</td>
<td></td>
</tr>
<tr>
<td>American Indian/Alaska Native</td>
<td>5 (7.5)</td>
</tr>
<tr>
<td>Asian</td>
<td>1 (1.5)</td>
</tr>
<tr>
<td>Black</td>
<td>6 (9)</td>
</tr>
<tr>
<td>Native Hawaiian /Pacific Islander</td>
<td>1 (1.5)</td>
</tr>
<tr>
<td>White</td>
<td>53 (80.3)</td>
</tr>
<tr>
<td>Ethnicity</td>
<td></td>
</tr>
<tr>
<td>Hispanic or Latino</td>
<td>46 (69.7)</td>
</tr>
<tr>
<td>Not Hispanic or Latino</td>
<td>20 (30.3)</td>
</tr>
<tr>
<td>Education</td>
<td></td>
</tr>
<tr>
<td>Less than high school</td>
<td>1 (1.5)</td>
</tr>
<tr>
<td>High School diploma</td>
<td>13 (19.7)</td>
</tr>
<tr>
<td>Associate’s degree</td>
<td>17 (25.7)</td>
</tr>
<tr>
<td>Bachelor’s degree</td>
<td>24 (36.3)</td>
</tr>
<tr>
<td>Master’s degree</td>
<td>10 (15.1)</td>
</tr>
<tr>
<td>Doctorate degree</td>
<td>1 (1.5)</td>
</tr>
<tr>
<td>Language spoken at home</td>
<td></td>
</tr>
<tr>
<td>Never Spanish</td>
<td>2 (3)</td>
</tr>
<tr>
<td>Rarely Spanish</td>
<td>8 (12.1)</td>
</tr>
<tr>
<td>Half Spanish</td>
<td>26 (39.3)</td>
</tr>
<tr>
<td>Mostly Spanish</td>
<td>23 (34.8)</td>
</tr>
<tr>
<td>Only Spanish</td>
<td>7 (10.6)</td>
</tr>
</tbody>
</table>

**Feature Comparison**

**Features that impact perceived difficulty**

To evaluate the differences in feature occurrences between easy and difficult texts, we carried out t-tests for all features. Since the easy/difficult document sets were different for perceived and actual difficulty, we conducted a separate analysis for each. The t-test results for the perceived difficulty partitioning are shown in Table 3. We used Bonferroni correction with significance level ($\alpha = 0.05$) to reduce Type 1 errors. With 17 features, the level required for statistical significance then becomes $0.05/17 = 0.003$.

Table 3 shows that there are several significant differences between the perceived easy and difficult texts. For surface features, we show, as has been seen in English text, that easy documents contain more adverbs while difficult documents contain more adjectives$^{11}$. Interestingly, unlike in English text, we did not see any significant differences between the use of function words, nouns and verbs. We also saw similarities with English text for the semantic and grammatical features, with easy documents containing more frequent words and using fewer unique
grammatical structures. We also saw more repeated words, punctuation and numbers in easy texts, while difficult texts had more negations.

Table 2. Classification of texts

<table>
<thead>
<tr>
<th></th>
<th>Using Perceived Difficulty</th>
<th>Using Actual Difficulty</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Easy</td>
<td>Difficult</td>
</tr>
<tr>
<td>Number of texts</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>Average Perceived Difficulty</td>
<td>3.130</td>
<td>2.206</td>
</tr>
<tr>
<td>Average Actual Difficulty</td>
<td>NA</td>
<td>NA</td>
</tr>
</tbody>
</table>

Table 3. The t-test results for easy and difficult texts based on perceived difficulty (*' indicates significance with Bonferroni adjustment)

<table>
<thead>
<tr>
<th>Feature Type</th>
<th>Feature Name</th>
<th>Easy</th>
<th>Difficult</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Surface</td>
<td>Average number of sentences</td>
<td>2.814</td>
<td>4.07</td>
<td>0 (*)</td>
</tr>
<tr>
<td></td>
<td>Average words per sentence</td>
<td>34.741</td>
<td>25.8</td>
<td>0 (*)</td>
</tr>
<tr>
<td></td>
<td>Functional words ratio</td>
<td>0.402</td>
<td>0.403</td>
<td>0.924</td>
</tr>
<tr>
<td></td>
<td>Punctuations ratio</td>
<td>0.121</td>
<td>0.095</td>
<td>0 (*)</td>
</tr>
<tr>
<td></td>
<td>Numbers ratio</td>
<td>0.017</td>
<td>0.002</td>
<td>0 (*)</td>
</tr>
<tr>
<td></td>
<td>Negations ratio</td>
<td>0</td>
<td>0.003</td>
<td>0 (*)</td>
</tr>
<tr>
<td></td>
<td>Nouns ratio</td>
<td>0.303</td>
<td>0.301</td>
<td>0.892</td>
</tr>
<tr>
<td></td>
<td>Adjectives ratio</td>
<td>0.121</td>
<td>0.162</td>
<td>0 (*)</td>
</tr>
<tr>
<td></td>
<td>Adverbs ratio</td>
<td>0.029</td>
<td>0.015</td>
<td>0 (*)</td>
</tr>
<tr>
<td></td>
<td>Verbs ratio</td>
<td>0.101</td>
<td>0.101</td>
<td>0.986</td>
</tr>
<tr>
<td>Semantic</td>
<td>Repeated words ratio</td>
<td>0.12</td>
<td>0.102</td>
<td>0 (*)</td>
</tr>
<tr>
<td></td>
<td>Term familiarity ratio</td>
<td>1344.225</td>
<td>796.768</td>
<td>0 (*)</td>
</tr>
<tr>
<td>Grammatical</td>
<td>Average grammar frequency</td>
<td>2.815</td>
<td>3.47</td>
<td>0 (*)</td>
</tr>
<tr>
<td></td>
<td>Average grammar question</td>
<td>0.350</td>
<td>0.621</td>
<td>0 (*)</td>
</tr>
<tr>
<td></td>
<td>Average minimum edit distance</td>
<td>3.451</td>
<td>3.463</td>
<td>0.327</td>
</tr>
<tr>
<td></td>
<td>Average cosine similarity</td>
<td>0.473</td>
<td>0.464</td>
<td>0.848</td>
</tr>
<tr>
<td>Other</td>
<td>Percent accuracy</td>
<td>0.652</td>
<td>0.571</td>
<td>0.024</td>
</tr>
</tbody>
</table>

Features that impact actual difficulty

Table 4 shows the analysis when splitting the text based on actual difficulty. Many of the same patterns appear, with numbers and adjectives occurring more in easy texts and negations more frequently in difficult texts, though a number of the features are not significant in this split. Difficult texts do use more nouns, though, which has been seen in English as well\textsuperscript{11}. Interestingly, while the easy texts had fewer structures (not significant), the structures were more similar in the difficult texts.

Discussion

Many features are significantly different between easy and difficult texts, particularly when perceived difficulty is used to make the distinction. The surface features such as the proportions of numerals, negations and adjectives are significant for both the perceived and actual difficulty and follow the same direction in terms of high and low between easy and difficult text, while average words per sentence follows the opposite direction. Relating to perceived difficulty, the easy texts are shorter in length but have longer sentences and difficult texts have more different grammar structures. Surprisingly, for actual difficulty, sentences in the easy texts have more dissimilar grammar structures. This means the sentences of difficult texts have more similar grammar structures. Neither of the
similarity measures plays an important role in identifying difficulty of Spanish texts from the perceived difficulty perspective.

Comparing these findings with those for English texts\(^7\), we see that from the perceived difficulty perspective easier Spanish texts have more punctuation and longer sentences. Similar to English, difficult Spanish texts contain more negations\(^8\). In English more nouns are seen in difficult texts\(^5\), but the proportion of adjectives is significantly higher in difficult Spanish texts. Surprisingly, function word use is not significantly different between easy and difficult text. In contrast, in English texts the higher frequency of function words identifies the text as easy\(^9\). Furthermore, easy texts in Spanish contain more adverbs, while easier English texts have more verbs\(^5\). Regarding semantic features, easy Spanish texts have a higher term familiarity, which is similar to easier English texts. Unlike English, for Spanish texts the grammar frequency does not have any effect on evaluating how easy a sentence is to understand (i.e., actual difficulty)\(^37\).

Some of these features differences between Spanish and English can be partially accounted for because of language differences. Though Spanish word order is similar to English (Subject-Verb-Object), Spanish generally places words that are emphasized at the end of the sentence. This may result in more grammatical variation. Long noun groups (modifier-noun-qualifier) are commonly used in English text, but are troublesome in Spanish\(^5\). To express the same meaning, Spanish sentences use prepositions or other avenues instead that may partially account for longer sentences.

Although preliminary, some initial practical observations can be made. Many of the practices used already in English can be continued in Spanish, e.g. use repetition, avoid negations and use familiar words. However, not all practices can be directly followed. For example, it is common in English to shorten sentence (e.g. by splitting long sentences) when simplifying text. In Spanish, we find that longer sentences are easier. Relatedly, extra caution should be used when using readability metrics, since they often incorporate sentence length, with a bias towards shorter sentences.

Looking forward, we plan to incorporate our findings into a tool to assist writers. Directly, the features above can be measured before and after manual simplification to help the writer understand the changes they’ve made and how the changes correlate with these findings. Additionally, suggestions can be made such as identifying more familiar (i.e. higher frequency) synonyms.

**Table 4.** The t-test results for easy and difficult texts based on actual difficulty (\(*\)* indicates significance with Bonferroni adjustment)

<table>
<thead>
<tr>
<th>Feature Type</th>
<th>Feature Name</th>
<th>Easy</th>
<th>Difficult</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Surface</td>
<td>Average number of sentences</td>
<td>3.510</td>
<td>3.750</td>
<td>0.106</td>
</tr>
<tr>
<td></td>
<td>Average words per sentence</td>
<td>27.735</td>
<td>34.024</td>
<td>0.002 (*)</td>
</tr>
<tr>
<td></td>
<td>Functional words ratio</td>
<td>0.392</td>
<td>0.396</td>
<td>0.523</td>
</tr>
<tr>
<td></td>
<td>Punctuations ratio</td>
<td>0.114</td>
<td>0.107</td>
<td>0.222</td>
</tr>
<tr>
<td></td>
<td>Numbers ratio</td>
<td>0.013</td>
<td>0.003</td>
<td>0 (*)</td>
</tr>
<tr>
<td></td>
<td>Negations ratio</td>
<td>0.001</td>
<td>0.003</td>
<td>0 (*)</td>
</tr>
<tr>
<td></td>
<td>Nouns ratio</td>
<td>0.288</td>
<td>0.309</td>
<td>0 (*)</td>
</tr>
<tr>
<td></td>
<td>Adjectives ratio</td>
<td>0.133</td>
<td>0.155</td>
<td>0 (*)</td>
</tr>
<tr>
<td></td>
<td>Adverbs ratio</td>
<td>0.024</td>
<td>0.022</td>
<td>0.39</td>
</tr>
<tr>
<td></td>
<td>Verbs ratio</td>
<td>0.106</td>
<td>0.104</td>
<td>0.523</td>
</tr>
<tr>
<td>Semantic</td>
<td>Repeated words ratio</td>
<td>0.114</td>
<td>0.104</td>
<td>0.009</td>
</tr>
<tr>
<td></td>
<td>Term familiarity ratio</td>
<td>1127.095</td>
<td>1035.268</td>
<td>0.207</td>
</tr>
<tr>
<td>Grammatical</td>
<td>Average grammar frequency</td>
<td>3.202</td>
<td>3.434</td>
<td>0.111</td>
</tr>
<tr>
<td></td>
<td>Average grammar question</td>
<td>0.524</td>
<td>0.482</td>
<td>0.119</td>
</tr>
<tr>
<td></td>
<td>Average minimum edit distance</td>
<td>3.877</td>
<td>2.939</td>
<td>0 (*)</td>
</tr>
<tr>
<td></td>
<td>Average cosine similarity</td>
<td>0.49</td>
<td>0.527</td>
<td>0.328</td>
</tr>
<tr>
<td>Other</td>
<td>Perceived difficulty</td>
<td>0.671</td>
<td>0.634</td>
<td>0.236</td>
</tr>
</tbody>
</table>
Conclusion

We measured perceived and actual difficulty separately in our study and evaluated the effect of different linguistic features that are indicative of easy and difficult Spanish text. Difficult texts have more nouns and adjectives, while easy texts have more adverbs and longer sentences. A Spanish text perceived as less difficult has higher term familiarity. Features identifying difficult texts exhibit somewhat similar behavior for both Spanish and English texts.

Acknowledgement

Research reported in this publication was supported by the National Library Of Medicine of the National Institutes of Health under Award Number R01LM011975. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

52. Stanford CoreNLP – a suite of core NLP tools Stanford, USA: Stanford CoreNLP; [Available from: http://stanfordnlp.github.io/CoreNLP/].
Identification of Clinically Meaningful Plasma Transfusion Subgroups Using Unsupervised Random Forest Clustering

Che Ngufor, Ph.D., Matthew A. Warner, M.D, Dennis H. Murphree, Ph.D., Hongfang Liu, Ph.D., Rickey Carter, Ph.D., Curtis B. Storlie Ph.D, and Daryl J. Kor, M.D.

Mayo Clinic, Rochester, MN

Abstract

Statistical techniques such as propensity score matching and instrumental variable are commonly employed to “simulate” randomization and adjust for measured confounders in comparative effectiveness research. Despite such adjustments, the results of these methods apply essentially to an “average” patient. However, as patients show significant heterogeneity in their responses to treatments, this average effect is of limited value. It does not account for individual level variabilities, which can deviate substantially from the population average. To address this critical problem, we present a framework that allows the discovery of clinically meaningful homogeneous subgroups with differential effects of plasma transfusion using unsupervised random forest clustering. Subgroup analysis using two blood transfusion datasets show that considerable variablilities exist between the subgroups and population in both the treatment effect of plasma transfusion on bleeding and mortality and risk factors for these outcomes. These results support the customization of blood transfusion therapy for the individual patient.

Keywords: Plasma transfusion, bleeding, unsupervised learning, subgroup analysis.

Introduction

Numerous studies and published guidelines encourage the appropriate use of fresh frozen plasma (FFP) and recommend specific circumstances for FFP transfusions. Although there is some variation about the definition of appropriate FFP transfusion, most guidelines suggest a cutoff in the international normalized ratio (INR) of 1.5 (i.e. Prothrombin time > 1.5 × normal).1 However, the documented compliance to these guidelines is poor.2–4 Data suggest that inappropriate FFP transfusion varies from institution to institution and ranges from about 10% to 83%.5 6 Moreover, FFP transfusion puts the patient at risk of a variety of outcomes. FFP transfusion is associated with high risk of transfusion-associated circulatory overload (TACO),7 transfusion-associated lung injury (TRALI),8 perioperative bleeding,8–10 multi-organ failure,7 infectious complications, and increase in health resource utilization. Therefore, strategies that can safely reduce the need for FFP transfusion bear high potential for improving patient outcomes. Recognizing that inappropriate plasma transfusions should be avoided, the literature is however not clear about the remaining percentage for which it might be beneficial. Currently, no study or best-practice guidelines exist regarding either patient subpopulation or specific characteristics of those patients who might benefit from FFP transfusion. Identifying the subgroup(s) with the most beneficial; wasteful; harmful, or futile prospect of FFP transfusion can provide an efficient means to improve patient outcomes, reduce unnecessary exposure to treatment adverse effects, and save resources.

Subgroup analysis is an important task in comparative effectiveness research where assessing the effect of a treatment on an outcome is of critical interest. Large observational health care databases provide potentially rich sources of information for data mining and machine learning methods to help research on heterogeneity in patient response to treatments and to guide care-givers’ decisions. Because of the large sample sizes, heterogeneous patient population, and real-world settings, they are suitable for studying either patient-specific or group-specific characteristics with respect to a clinical measure. However, comparative effectiveness research based on observational data is challenged by both selection bias and potential for unmeasured confounding. In usual care settings, many patient and physician factors influence whether a patient is selected for a treatment or not, thus any comparison between treatment groups is subject to bias. Through classical statistical methods such as propensity score matching and instrumental variables, it is possible to adjust for measured confounding and obtain unbiased estimates of treatment effects. These methods however suffer from several known weaknesses described below.

Traditionally, treatment effect is commonly estimated by a regression model where the outcome is regressed against patient covariates and the treatment. The effect is then read off as the corresponding regression coefficient of the treatment variable. However, as patients can show significant heterogeneity in response to a treatment, this “average” effect is not appropriate for describing individual level differential effects. Average superiority of one treatment
over another does not necessarily mean the treatment will remain superior for each patient. As a result of heterogeneity in patient characteristics such as genetics, phenotypic, pharmacokinetic, environmental, and socio-economic factors, clinical outcomes for some patients may deviate considerably from the population average. The important relationship between treatment effect and patient heterogeneity has been well investigated; however, comparative effectiveness researchers still rely on inefficient and non-robust classical regression and propensity score methods for estimation of treatment effects in observational studies.

In this study, we provide a three stage framework that allows the discovery of stable, robust and clinically meaningful homogeneous subgroups with differential effects of plasma transfusion on important patient outcomes. In the first step, our proposed framework makes use of the unsupervised random forest (URF) algorithm to derive a “proximity” or dissimilarity matrix between data points in a mixed-type (continuous and categorical) high dimensional covariate space. In the second step, we use the dissimilarity matrix in a hierarchical clustering algorithm to identify highly similar patient subgroups. Compared to classical parametrically derived propensity scores, the URF subgroup membership represents a more robust covariate balancing score. Thus, treatment effect estimates within subgroups of well-matched clinically homogeneous patients are then conditionally unbiased. In the final step, we applied the doubly robust targeted maximum likelihood estimation (TMLE) method to estimate the effect of FFP transfusion on bleeding and mortality in each subgroup. The TMLE further insures against any potential confounding that may still exist in the subgroups.

The framework was applied to two datasets from a single academic institutional blood transfusion datamart to discover subgroups of patients with differential responses to pre-operative or pre-procedural plasma transfusion (PPT) on two important patient outcomes: intra-operative or intra-procedural bleeding and mortality. Using only pre-operative or pre-procedural patient information, a cluster validation technique based on the predictive strength of cluster memberships and treatment assignment indicated that the first dataset consisting of patients undergoing non-cardiac surgery (NCS) can be clustered into six homogeneous subgroups while the second dataset consisting of patients undergoing interventional radiology (IR) procedures can be clustered into five subgroups. With respect to clustering the NCS data set, we found two clusters with harmful effect, two clusters with beneficial effect, and a cluster with no effect of PPT on bleeding. Three clusters showed no effect of PPT on mortality and two clusters showed harmful effects. Similar results were obtained for the IR data set. Compared to previous studies that have shown the population wide harmful effects of plasma transfusion, the findings in this study suggest the need to consider individualized and/or subgroup effects of plasma transfusion.

To further characterize phenotypes of patients within these subgroups, we applied a random forest feature contribution technique to determine which patient characteristics most strongly predict bleeding or mortality at both the population and individual levels. The feature contributions showed that considerable variabilities exist between population level risk factors and individual/subgroup level risk factors.

Method

Study Population

This is a retrospective observational cohort study conducted under the approval of the Mayo Clinic Institutional Review Board (Rochester, MN) before initiation. The protocol was reviewed and approved by institutional review board as a minimal risk study and informed consent was not required. Screening for potential study participants was performed using the perioperative datamart, an institutional resource that captures clinical and procedural data for all patients who are admitted to an acute care environment including procedural suites, operating rooms, ICUs, and progressive care units at the study’s participating institution. This robust data warehouse also contains information on baseline demographic and clinical characteristics, fluid and transfusion therapies, perioperative/periprocedural medications and laboratory values, postoperative/postprocedural outcomes, and lengths of stay. Two different cohorts of patients were extracted from the datamart: the first comprising patients undergoing non-cardiac surgery and the second made up of patients undergoing percutaneous invasive image-guided intervention (i.e., inpatient or outpatient procedures performed by the Division of Vascular and Interventional Radiology).

Non-Cardiac Surgery Data

The non-cardiac surgery data (NCS) was originally extracted to study the association between preoperative plasma transfusion and perioperative bleeding complications for patients with elevated INR. To be considered for study
participation, patients must meet the following criteria: age $\geq 18$ years, non-cardiac surgery and an INR $\geq 1.5$ in the 30 days preceding surgery. Between January 1, 2008 and December 31, 2011, a total of 1,233 patients were identified and comprised the study population. Plasma transfusion was offered to 139 patients. To expand the work in\(^9\) that was based on traditional propensity score and matching techniques, we used the same data set and kept the same exclusion and inclusion criterion.

**Baseline Variables.** Baseline patient demographics include age, height, weight, gender and the ASA physical status classification. Disease conditions included myocardial infarction, congestive heart failure, cerebrovascular disease, dementia, chronic pulmonary disease, diabetes mellitus, etc. Preoperative laboratory values included INR, hemoglobin, platelet counts, creatinine, albumin, and APTT (activated partial thromboplastin time). A total of 51 predictors were considered for inclusion in the analyses.

**Outcomes and Treatment.** Two main outcomes were considered: perioperative bleeding and mortality. Bleeding was taken as the World Health Organization (WHO) grade 3 bleeding events, defined as the need for early perioperative red blood cell (RBC) transfusion.\(^9\) Mortality was death during surgery or death within 30 days post-surgery (typically in the ICU). The treatment variable PPT indicates if a patient was offered plasma transfusion after INR test and 24 hours before surgery. As a guard against residual confounding, all RBC transfusions cases within this interval were dropped.

**Interventional Radiology Data**

The interventional radiology (IR) data set has been used in\(^{10}\) to study the association between prophylactic plasma transfusion and periprocedural RBC transfusion rates (or bleeding) in patients with elevated INR (INR $\geq 1.5$ ). Similar to the NCS study, the IR study was based on traditional propensity and matching methods and this study seeks to expand those results through application of advanced machine learning methods. As with the NCS study, the same inclusion and exclusion criteria were used. Between January 1, 2009 and December 31, a total of 1,902 patients met the inclusion criterion with 190 receiving plasma transfusion. Similar groups of baseline predictors as for the NCS data were used for the analysis.

To handle missing values in both data sets, we applied the random forest imputation method *missForest*\(^{20}\) implemented in the R statistical programming language to impute variables with less than 35% missing observations. Predictors with greater missingness were removed from the data.

**Unsupervised Random Forest**

The goal of clustering the blood transfusion data is to discover internal structure in the data by breaking it down into groups without any prior knowledge about the groupings. The idea is that once these groups are identified and proven robust, the clusters can aid in the determination of the effects of plasma transfusion. Not only are the clusters expected to balance the covariates (mitigate confounding) and account for patient heterogeneity, we also expect them to be clinically meaningful. A clustering technique known to be able to produce accurate and clinically meaningful clusters is the unsupervised random forest (URF) clustering.\(^{13;21}\) URF clustering has the additional attractive property that it can handle mixed type of variables. The NCS and IR data sets contain both continuous (e.g. age) and categorical (e.g. race) variables making the use of classical clustering methods such as hierarchical or k-means clustering based on Euclidean or binary type distance measures inappropriate. While many researchers in health sciences have mainly used the random forest\(^{13}\) method for supervised learning in the context of classification, regression, or feature selection, many are unaware of its utility in unsupervised learning.\(^{22}\)

*Random Forest (RF):* RF is an ensemble learning method where multiple decision trees are constructed on a bootstrap sample of the training data and the predictions combined by averaging or majority vote. The left-out cases of the bootstrap also called out-of-bag (OOB) sample and consisting of around 37% of the data, are not used for tree construction but are used to validate the performance of the tree. The splitting criterion in RF is based on selecting a random subset of predictors or *mtry* and the predictor yielding the best split within this set is chosen to perform the split. An important output of a RF analysis is the “proximity” matrix, a similarity matrix of size $n \times n$, where $n$ is the number of observations. This matrix constitutes the fraction of times in which two observations are placed in the same terminal node of a tree. The intuition is, if two observations end up in the same terminal node, then they are naturally similar and their proximity or similarity score is increased by one. This is done for all observations and trees in the forest and the proximities are normalized by dividing by the number of trees. Computation of the proximity matrix is not
required for classification or regression problems, but crucial for URF. The proximity matrix can be easily transformed into a dissimilarity matrix: if $s_{ij} \in [0,1]$ is the proximity of the $i$ and $j$ observations, then the distance between them is given by $d_{ij} = \sqrt{1 - s_{ij}}$. The dissimilarity matrix can then be used for unsupervised learning such as clustering and multidimensional scaling.

**URF Clustering:** URF clustering consists of two steps. In the first step, a RF classification model is generated to distinguish between the original data labeled as class 1 and a synthetic data of the same size as the original data and labeled as class 0. One way to generate the synthetic data is to take independent random samples from each dimension according to the empirical distribution of the corresponding dimension of the true data and a second approach is to simply permute each dimension. The supervised learning step attempts to distinguish the true data from a random version of the data, thus if there exists any underlying structure in the true data, the OOB error will be small, showing that the synthetic data destroyed that structure. An OOB error of about 50% indicates the original data is not very different from the synthetic data and possibly contains no informative structure. Thus the OOB error provides a natural data driven way to determine if interesting patterns exist in the data. In the second step, the proximity matrix between the true data points is extracted and passed to a clustering algorithm such as hierarchical clustering. For clustering the NCS and IR data, we use the permutation strategy to generate synthetic data and used the agglomerative hierarchical clustering method with the Ward’s minimum variance criterion to identify clinically relevant subgroups with differential effects of plasma transfusion.

**Cluster Validation:** Cluster validation is the process of evaluating the quality of a clustering result, and is vital to the success of clustering applications. The robustness or stability of the clustering as well as the optimal number of clusters can be determined using “internal” or “external” validation measures.23 Internal measures used only information available to the data without reference to any external information. As our goal is to generate homogeneous clusters with respect the treatment a patient receives and an outcome, we evaluated the homogeneity of the clustering using an external validation technique that make use of the treatment and outcome not used to generate the clusters. Our external measure is based on the assumption that similar patients receiving the same type of treatment are expected to experience the same outcome. Thus, clusters are validated by measuring the area under the ROC curve (AUC) for a logistic regression model that predicts the true outcome (Bleeding and Mortality) based on cluster memberships and treatment (PPT). For a stable clustering application, the more homogeneous the clusters, the higher the AUCs, which implies that the AUC increases with the number of clusters. We select the optimal number of clusters by plotting the AUC by the number of clusters and apply the “elbow criterion”. The elbow is the point on the graph where addition of a cluster does not lead to a significant gain in AUC and corresponds to the optimal number of clusters (see Figure 1 (a) and (b)). As the elbow method can be ambiguous, we also apply principle of parsimony when selecting the number of clusters.

**Random Forest Feature Contributions**

An important component of the RF algorithm is its capability to produce a variable ranking score or variable importance based on association of a given predictor with other predictors and with the outcome variable. This score can help in interpretation of the model or for dimension reduction. The variable importance is a population measure, and as an average score, it does not indicate the relative influence of each patients’ feature value in predicting the outcome. A variable might be an important risk factor for the overall population, but not a risk factor for a given individual or subgroup of patients. Recently, the RF variable importance measure has been extended to an importance score or feature contribution19 for each individual patient in the training set. The feature contributions characterize the relative contribution of a patient’s baseline variables towards predicting an outcome value or class. Another attractive property of the feature contribution method is that it can be predicted for new patients. This offers a way to further validate the random forest model: when the average feature contributions of the training and test set matches, then the model can generalize well. For classification problems, a zero value for feature contribution indicates that the variable is irrelevant with respect to assigning a patient to a given class. Positive values indicate that the variable is influential towards classifying the patient to a reference class.

**Treatment Effect of Plasma Transfusion**

The standard approach to investigate the causal relationship between a treatment or exposure and an outcome is to construct statistical regression models in which the outcome is regressed against baseline covariates and the treatment variable. The attributable effect of the treatment is then read off as the corresponding regression coefficient. This
study takes a different approach and estimates the treatment effect through application of machine learning methods. The theory of causal inference or technical details of the considered estimation procedure are beyond the scope of this study. The interested reader is referred to\textsuperscript{16,24,25} for more details. However, for the purposes of this study, a brief discussion of the data structure required to compute these estimators is presented next.

\textit{Data structure and likelihood.} The observations for each patient in the data set can be written as $O = (X, Y, Z)$ where $Z \in \{0, 1\}$ is the treatment indicator with $Z = 1$ if patient was treated and $Z = 0$ if patient was not treated. $X$ is a vector of baseline covariates that records information specific to each patient prior to treatment. $Y$ is the outcome such as bleeding or mortality. The relationship between the observed variables in $O$ can be written in a factorize data likelihood as

$$
Pr(X, Y, Z) = Pr(Z|X) Pr(Y|Z, X) Pr(X).
$$

Pr$(X)$ and Pr$(Y|X, Z)$ are referred to as the $Q$ component of the likelihood while Pr$(Z|X)$ is the $g$ component. $g(Z|X)$ represents the propensity or the causal disposition of the treatment to produce some outcome. Let $Q_0(Z, X) = E[Y|Z, X]$ be the true potential outcome conditional on the observed characteristics. Estimates of $g$ and $Q_0$ can be obtained by standard regression or machine learning methods.

For a binary outcome and in the presence of no confounding variables, the treatment effect can be easily computed by taking the expectations $\psi_1 = E[Y_{Z=1}]$ and $\psi_0 = E[Y_{Z=0}]$: where $E[Y_{Z=1}]$ is the mean of $Y$ assuming every patient in the population was exposed at level $Z = 1$. These two statistics can then be combined in useful ways to assess the effect of different levels of the treatment. Two commonly reported summary statistics include the Additive Treatment Effect : $ATE = \psi_1 - \psi_0$ and the Risk Ratio : $RR = \psi_1 / \psi_0$.

The $ATE$ quantifies the additive effect of every patient being exposed to the event versus not being exposed. Thus, a meaningful interpretation of $ATE = 0.05$ could read: “offering a patient plasma transfusion versus not increases the risk of bleeding/mortality by 5%”.\textsuperscript{16} The $RR$ quantifies the multiplicative effect of being exposed versus not. A $RR$ of 5 can be interpreted as: “offering a patient plasma transfusion versus not would lead to a 5 times increase in the risk of bleeding/mortality”.

\textit{Targeted maximum likelihood estimation.} In observational studies, estimators of treatment effect need to account for possible confounding, i.e situations where the (apparent) effect of the treatment is actually the effect of another characteristic that is associated with both the treatment and the outcome. Several methods have been proposed for the estimation of $ATE$ and $RR$ in a way that can mitigate the effects of confounding (and model misspecification), e.g. G-computation formula, propensity score matching, inverse probability of treatment weighting (IPTW), and doubly-robust estimation. See\textsuperscript{24,25} for more in-depth discussion of these estimators. In this study, the targeted maximum likelihood estimation (TMLE)\textsuperscript{16,17} method is considered because of its double robustness and bias reduction properties. TMLE is a two stage doubly robust semi-parametric estimation methodology designed to minimize the bias of the parameters of interest. The first stage of the method estimates the density of the data generating distribution (specifically $Q_0$) while the second stage solves an efficient influence curve estimating equation. The influence curve describes the behavior of the target parameter under slight changes of the initial density estimates.

In TMLE, if either $g$ or $Q_0$ are consistently estimated, then the TMLE estimator is guaranteed to be asymptotically unbiased. However, TMLE will not return consistent estimates of the parameter of interest when both $g$ and $Q_0$ are misspecified. Thus it is important to avoid overfitting these measures.

As discussed above, estimating the two statistics $\psi_1$ and $\psi_0$ allows for calculating any of the causal effects $ATE$ and $RR$. The TMLE estimate of $\psi_z (z \in \{0, 1\})$ is given by

$$
\hat{\psi}_z = \frac{1}{n} \sum_{i=1}^{n} \hat{Q}_0^*(z, x_i)
$$

where $\hat{Q}_0^*(z, x_i)$ is an update of $\hat{Q}_0(z, x_i)$. The targeting step for updating $\hat{Q}_0(z, x_i)$ is done by fluctuating $\hat{Q}_0(z, x_i)$ through a parametric sub-model of the form: $logit(\hat{Q}_0^*(z, x)) = logit(\hat{Q}_0(z, x)) + \varepsilon \hat{H}(z, x)$, where $\varepsilon$ is the fluctuation parameter, $\hat{H}(z, x) = I(Z = z)/(\hat{g}(z, x))$ is the efficient influence curve equations, and $I$ is the indicator function. The MLE of $\varepsilon$ is obtained by a logistic regression of $Y$ on $\hat{H}(z, x)$ with offset logit$(\hat{Q}_0(z, x))$. Confidence intervals and p-value for TMLE can be obtained through the variance of the influence curve.
TMLE can use initial estimates of $Q_0$ and $g$ from any fixed parametric model such as generalized linear models (GLM) (e.g., logistic regression). However, most parametric models require a functional form for the predictors, and some assume distributions for the outcome and predictors variables, which are often not realistic such that model misspecification is difficult to avoid. It is therefore recommended to use machine learning methods that make little or no assumptions and are able to estimate complex relationships between the outcome and observed variables.

Results

This section presents the main results: estimates of treatment effect of PPT on bleeding and mortality for the complete data and for each subgroup. For the calculations of ATE and RR, we estimate $Q_0$ and $g$ using five models: generalized boosting machine (GBM), random forest, support vector machine (SVM), logistic regression, and extreme logistic regression (ELR) and select the best model through 5-fold cross-validation. First, we present the cluster validation analysis, then the treatment effect analysis, and end the section with analysis of the population or median feature contributions over all patients in the cohort and the individual feature contributions for two random patients.

Cluster validation and selection of optimal number of clusters

A robust and stable clustering procedure generates homogeneous clusters such that patients within clusters are similar with respect to baseline characteristics. This implies the procedure has identified hidden structure in the data. A way to determine if the URF method identified underlying structure in the NCS and IR datasets is to look at the OOB error rates. Specifically, the OOB error rate for the NCS and IR data clustering problems were 11.95% and 1.20% respectively, indicating that the synthetic data destroyed existing structure in the data and the RF model was able to capture this information with high accuracy.

Ward’s Minimum Variance: URF can identify structure in the data, but obtaining good clusters crucially depends on the clustering algorithm. We employ the agglomerative hierarchical clustering algorithm with the Ward’s minimum variance criterion. Figure 1 (c) and (d) shows the dendrogram of the algorithm for the NCS and IR data sets respectively. The dendrogram represents the similarity relationships between patients in a tree-like form. Agglomerative hierarchical clustering starts by assuming that each patient is a cluster, and successively merges similar clusters to form larger clusters. Because of the hierarchical structure, different number of clusters can be obtained by cutting the tree at different heights. We used the Ward’s method, which minimizes the sums of squares between clusters to merge similar clusters together. However, we also tried other merging algorithms such as the single, complete, and average linkage methods, but all produced unstable and sparse clusters. In contrast, the Ward’s method produced stable and equal sized clusters.

Predictive Ability of Cluster Memberships: The clustering partitioned the data space into non-overlapping regions, where each region is associated to a given level of the treatment and outcome. In other words, if the regions are sufficiently homogeneous, then similar outcomes are expected for patients if offered the same treatment. As a consequence, a classification model can efficiently discriminate between the classes based on treatment status and subgroup allocations, and the discriminative power increases the more homogeneous the groups become.

Figure 1 (a) and (b) shows AUCs (averaged over 5-fold cross-validation) of a logistic regression model predicting bleeding (red curve) and mortality (blue curve) based on cluster memberships and plasma transfusion plotted against the number of clusters. Clearly, the cluster memberships are predictive as can be seen by the rise in AUC as the number of clusters increases (solid lines) compared to the poor and unstable performance of a randomly generated clusters (dotted lines). There is a big jump in performance from 2, 3 and 4 clusters to 5 or 6 clusters. After the 5'th or 6'th cluster, the relative increase in AUC reduces and becomes somewhat stable. We choose the number of clusters as that corresponding to the elbow or turning point of the curve. At the elbow, adding another cluster to the logistic regression model does not lead to any appreciable performance gain. Thus, 6 clusters are optimal for the NCS data and 5 for IR. Close observation of the curves will indicate that 7 or 8 clusters can equally be selected. However, these higher cluster numbers produced very sparse distributions of the observed treatment and outcome events in some of the clusters. In our implementation, we set the minimum number of events observed in each cluster to 10.

Effect of Plasma Transfusion on Bleeding and Mortality

Population Effect: Table 1 presents the ATE and RR quantifying the population effect of PPT on bleeding and mortality for the NCS and IR datasets. The GBM algorithm performed best for the NCS data, while the ELR offered the best
performance for the IR data. To save on space, discussions will be restricted to the ATE summary statistics; interpretations for RR can be similarly made. Overall, the estimates from TMLE confirmed previous findings that population wise, PPT increases the risk of bleeding. Specifically, for the population of NCS and IR patients considered in this study, PPT significantly increases the risk of bleeding by 14% (p-value = 0.00) and 12% (p-value = 0.00) respectively (95% confidence intervals are shown in brackets). With respect to mortality, PPT marginally increases risk by 4% for NCS and has no effect for IR populations.

Table 1: Population Effect of PPT on Bleeding and Mortality

<table>
<thead>
<tr>
<th>Data</th>
<th>Outcome</th>
<th>ATE</th>
<th>p-value</th>
<th>RR</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>NCS</td>
<td>Bleeding</td>
<td>0.14 (0.08, 0.20)</td>
<td>0.00</td>
<td>1.438 (1.25, 1.65)</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>Mortality</td>
<td>0.04 (0.00, 0.09)</td>
<td>0.06</td>
<td>1.523 (1.06, 2.00)</td>
<td>0.02</td>
</tr>
<tr>
<td>IR</td>
<td>Bleeding</td>
<td>0.12 (0.09, 0.15)</td>
<td>0.00</td>
<td>2.01 (1.68, 2.39)</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>Mortality</td>
<td>-0.01 (-0.03, 0.02)</td>
<td>0.53</td>
<td>0.92 (0.72, 1.19)</td>
<td>0.54</td>
</tr>
</tbody>
</table>

Subgroup Effect: Table 2 presents estimates of ATE and RR within each cluster identified for the NCS and IR data sets. With respect to the NCS clusters, we found: (a) One cluster with 551 patients where PPT increases the risk of bleeding and mortality by 32% and 6% (p-value = 0.00 and 0.05) respectively. (b) One cluster where PPT has no effect on bleeding and mortality. This cluster may represent patients where the administration of prophylactic plasma
transfusion is wasteful. (c) Two clusters of sizes 150 and 127, where PPT reduces the risk of bleeding by 9% in each
cluster. Correspondingly, PPT increases the risk of mortality in one cluster and has no effect in the second. (d) The
last cluster with 171 patients show harmful effect of PPT on bleeding and no effect on mortality.

Overall, the effect PPT on bleeding for the NCS data was beneficial in two subgroups and none showed any beneficial
effect with respect to mortality. It is therefore interesting to investigate the characteristics of patients in these sub-
groups. This information can help reduce the inappropriate use of plasma products as only the patients who will truly
benefit from plasma transfusion are considered for treatment.

For the IR data clustering, we found roughly similar results. One group with 187 patients having beneficial effect
of PPT on bleeding (11%. p-value=0.00). The IR data set however contains two subgroups with beneficial effect
of PPT on mortality. The NCS and IR clustering problems contain subgroups (n = 75 and 451 respectively) where
the observed number of patients with bleeding/mortality and PPT events was less than 10, and no treatment effect
estimates were computed for these groups.

<table>
<thead>
<tr>
<th>Data</th>
<th>Cluster</th>
<th>ATE (95% CI)</th>
<th>p-value</th>
<th>RR (95% CI)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>NCS</td>
<td>Cluster 1 (n=171)</td>
<td>Bleeding: -0.02 (-0.11, 0.08), 0.73</td>
<td>0.98 (0.84, 1.13), 0.74</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mortality: -0.03 (-0.12, 0.06), 0.53</td>
<td>0.89 (0.60, 1.30), 0.53</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cluster 2 (n=551)</td>
<td>Bleeding: 0.32 (0.23, 0.41), 0.00</td>
<td>2.06 (1.71, 2.47), 0.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mortality: 0.06 (0.00, 0.13), 0.05</td>
<td>1.67 (1.06, 2.59), 0.02</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cluster 3 (n=150)</td>
<td>Bleeding: -0.09 (-0.18, -0.01), 0.04</td>
<td>0.49 (0.27, 1.14), 0.09</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mortality: 0.74 (0.65, 0.82), 0.00</td>
<td>26.26 (10.07, 68.48), 0.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cluster 4 (n=171)</td>
<td>Bleeding: 0.13 (0.06, 0.20), 0.00</td>
<td>2.22 (1.42, 3.49), 0.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mortality: -0.01 (-0.02, 0.01), 0.30</td>
<td>0.00 (0.00, 0.00), 0.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cluster 5 (n=127)</td>
<td>Bleeding: -0.01 (-0.02, -0.01), 0.32</td>
<td>0.00 (0.00, 0.00), 0.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mortality: -0.01 (-0.02, -0.01), 0.32</td>
<td>0.00 (0.00, 0.00), 0.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cluster 6 (n=75)</td>
<td>Bleeding: -0.09 (-0.18, -0.01), 0.00</td>
<td>0.49 (0.29, 0.81), 0.01</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mortality: -0.01 (-0.02, -0.01), 0.32</td>
<td>0.00 (0.00, 0.00), 0.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IR</td>
<td>Cluster 1 (n=650)</td>
<td>Bleeding: 0.37 (0.33, 0.42), 0.00</td>
<td>4.92 (3.75, 6.44), 0.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mortality: -0.04 (-0.07, -0.01), 0.02</td>
<td>0.60 (0.39, 0.92), 0.02</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cluster 2 (n=390)</td>
<td>Bleeding: 0.06 (-0.01, 0.13), 0.11</td>
<td>1.45 (0.94, 2.23), 0.09</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mortality: 0.01 (-0.06, 0.07), 0.89</td>
<td>1.03 (0.62, 1.75), 0.90</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cluster 3 (n=451)</td>
<td>Bleeding: -0.12 (-0.21, -0.03), 0.01</td>
<td>1.50 (1.11, 2.03), 0.01</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mortality: -0.11 (-0.18, -0.04), 0.002</td>
<td>0.47 (0.28, 0.81), 0.01</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cluster 4 (n=224)</td>
<td>Bleeding: -0.11 (-0.18, -0.04), 0.003</td>
<td>0.49 (0.29, 0.81), 0.01</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mortality: 0.02 (-0.06, 0.10), 0.62</td>
<td>1.11 (0.73, 1.70), 0.62</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Feature Contributions

We report only the results for the NCS complete data. Result for the IR data and all subgroups can be obtained by
contacting the authors. Figures 2 (a) and (b) shows the median feature contributions for all patients averaged over
five-fold cross-validation and the corresponding feature contributions for two random patients. The two patients (P1
and P2) were both offered plasma transfusion but experience different levels of the bleeding/mortality outcome: P1
bled/died and P2 did not bleed/die. Preoperative hemoglobin levels and PLT test value (platelet count) are the two
most contributing variables towards predicting bleeding and mortality. Though appearing in different order, the same
9 variables appear among the top 10 most predictive variables for all patients in the two models. Considering the
median feature contribution, these top 9 variables are somewhat equally contributive towards predicting whether a
patient will bleed or not. However, these variables contribute more towards predicting the death status of a patient
compared to predicting alive status.

The plots clearly show that significant variabilities exist between population level (median) feature contributions and
the individual level. For example, while plasma transfusion is a strong contributing factor towards predicting bleeding
for the random patients, the effect of the variable is almost zero when we consider the population level. Similarly, while
the population level contributing factor of connective tissue disease is almost zero, it however contributes significantly
towards predicting the death status of the random patient P1.
Conclusion

The most common reason cited for plasma transfusion is the correction of an elevated pre-operation/pre-procedural international normalized ratio (INR) for the prevention of bleeding complications, despite lack of evidence to support such practices. The decision to offer plasma transfusion to patients with abnormal coagulation factors still remains largely controversial. Current recommendations are mostly based on expert opinion and a precautionary approach to correct abnormal laboratory test results and there is widespread variation in the practice with respect to plasma transfusion. Many studies, including randomized control trials, have shown no significant benefit for prophylactic and therapeutic use of fresh frozen plasma (FFP) across a range of indications. Moreover, majority of these studies report the inappropriateness and harmful effect of prophylactic plasma transfusion. However, almost all the studies have evaluated the effect of plasma transfusion at the population level. Despite accounting for parameters such as treatment selection bias and potential confounding in observational studies, those results apply essentially to the average patient. Given that the critically ill patient population can be highly heterogeneous in their responses to treatments, the average effect of a treatment is of limited value, as it ignores individual patient level variabilities of the treatment, which often deviate substantially from the population average. Furthermore, except for the work in, most of the published work on the effect of plasma transfusion have traditionally used classical regression, propensity score, and matching methods, which often make unrealistic and difficult to satisfy assumptions about the patient population.

This study takes a different approach and applied subgroup analysis based on efficient and robust machine learning methods and identified several homogeneous subgroups exhibiting differential effects of plasma transfusion on bleeding and mortality. Specifically, using the unsupervised random forest (URF) clustering method and the doubly robust targeted maximum likelihood estimation (TMLE) method, we identified stabled and clinically meaningful subgroups with beneficial, harmful, and no effect of plasma transfusion on bleeding and mortality. Recognizing the widespread inappropriate use of FFP and the lack of evidence to support the use of plasma transfusion to prevent bleeding, the results from this study suggest that researchers should reconsider evaluation measures based on the overall population, and strongly support the fact that blood transfusion therapy should be customized for the individual patient. Further, analysis of the subgroup characteristics can help shed light on the much needed evidence to support the use of plasma transfusion to correct prolonged prothrombin time and prevention of bleeding complications.

References


Casey L. Overby, Ph.D. 1,2,3, Phillip Thompkins, B.S.2, Harold Lehmann, M.D., Ph.D.3, Christopher G. Chute, M.D., M.P.H.1,2,3, Jeanne S. Sheffield, M.D.4

1Division of General Internal Medicine; 2Institute for Clinical & Translational Research; 3Division of Health Sciences Informatics 4Division of Maternal & Fetal Medicine; Johns Hopkins University School of Medicine, Baltimore, MD, USA

Abstract

In order to better understand the potential value of genetics-informed drug dose guidance to obstetric healthcare providers at Johns Hopkins we administered a web-based needs assessment survey. The survey included questions about: 1) experience with adjusting drug doses during pregnancy; 2) comfort prescribing medications to pregnant women with chronic conditions; 3) awareness and use of genetics-informed dosing guidance; and 4) perceived value of access to services to provide genetics-informed dosing guidance. Among thirty-one respondents, 81% indicated an interest in access to genetics-informed drug dose guidance, particularly a mobile or electronic health record (EHR) application. It was indicated, however, that genetics is one of many characteristics that influence dose adjustments during pregnancy. This study motivates future research to help obstetric healthcare providers tailor drug dose to individual patients based upon models integrating multiple patient characteristics, including genetics.

Introduction

While obstetric healthcare providers counsel pregnant women to limit medication use during pregnancy unless clinical necessity warrants, prescription drug use during pregnancy is on the rise. Mitchell et al 1 conducted a 33-year study finding that nearly half of women between 1997 and 2003 used prescription medications and less than 5% took four or more. Use increased from 2003 to 70% of pregnant women using any medication, and the use of four or more medications rose to over 15%. Women today are delaying pregnancy 2 resulting in older women with higher rates of obesity and chronic medical conditions becoming pregnant. This increase in chronic medical conditions often requiring medications has highlighted the need for more research on medication use and safety in pregnancy 3.

Despite these challenges, there is a growing body of pharmacogenomics data on medications in general that can be used to find the most effective and safe doses for pregnant women. In addition, the Food and Drug Administration (FDA) has included pharmacogenomics information in the labeling of several medications that are commonly prescribed in pregnant women 4. Thus, the development of clinical dosing models that bring together clinical characteristics, physiologic parameters in pregnancy, and pharmacogenomics measures has been proposed as a way forward to eventually guide obstetric healthcare providers to individualize drug therapy in pregnant patients 5,6.

The first stage of designing a new software product is to specify the needs for a setting and the intended users of that product 7. One common way to complete this evaluation and to establish design requirements is to use a needs assessment survey. Given our interests in understanding the needs of obstetric health care providers, we conducted a needs assessment survey. The primary goal for this survey was to understand the value of genetics-informed dosing guidance to obstetric healthcare providers.

Methods

Recruitment

An initial recruitment email containing a link to a web-based survey was sent to approximately 80 obstetric healthcare providers associated with the Johns Hopkins Hospital (JHH). Two follow-up emails were sent to the distribution list in subsequent weeks. Survey responses were collected anonymously, and participants who agreed to be contacted about future research were entered into a contest to win $100 USD.

Survey development

Four co-authors developed a twenty-four-item survey instrument with the primary goal of understanding the potential value of genetics-informed drug dose guidance to obstetric healthcare providers (CO, HL, CC, JS). The survey included questions about: experiences with adjusting drug doses in pregnant women (five multiple choice, and one free response question, e.g., “How often do you have to seek consultation to confirm drug doses in pregnant women?”);
comfort in prescribing nine antiretroviral, four antihypertensive, and fourteen antidepressant drugs to pregnant women (Likert scale questions e.g., “How comfortable are you in treating pregnant women with the following antihypertensive medications?”); awareness and use of genetics-informed dosing guidance (five Likert scale and one free response questions, e.g., “Do you know of and use any resources for clinical guidance on genetics-informed dose use in pregnancy?”); perceived value of tools and consult services to provide genetics-informed drug dosing guidance for pregnant women (two multiple choice questions, e.g., “Would it be valuable to have access to a tool that considers genetics in drug dose use in pregnancy?”); and demographic information such as gender, years practicing, and clinical caregiver status (six questions) (see Appendix 1). For questions about comfort with prescribing, one co-author (CO) selected an initial list of medications used to treat HIV, hypertension, and depression for which drug exposure is known to be influenced by genetics according to PharmGKB8. Another co-author with expertise in obstetrics (JS), then narrowed down that list of medications to those prescribed in pregnant women.

Data analysis
Survey data was collected using Qualtrics (Qualtrics, Provo, UT) and analyzed using Stata (StataCorp, 2013. Stata Statistical Software: Release 13. College Station, TX: StataCorp LP). We estimated the proportion of respondents indicating that genetics-informed drug dose guidance was of high importance by gender, experience, specialty, and years since completing training. We then evaluated factors that may influence the perceived value of accessing guidance including experiences with adjusting drug dose in pregnant women, comfort with adjusting drug doses in pregnant women taking antiretroviral, antihypertensive and antidepressant medications, the perceived importance of genetics-informed dosing guidance in prescribing those medications, and awareness and access to existing guidance for genetics-informed drug dosing. Furthermore, we identified which modalities obstetric healthcare providers would like to use when seeking guidance in prescribing. Only descriptive statistics are reported given that our population is relatively small (obstetric healthcare providers from one institution).

Results
Data collection initiated on October 11, 2016 and concluded on December 23, 2016. Thirty-one out of 80 obstetric healthcare providers completed the survey for a response rate was 39%. The results of this survey are accurate at the 90% confidence level plus or minus 12%. All survey respondents were employed by Johns Hopkins University, Johns Hopkins Hospital, and/or Johns Hopkins Community Physicians. In response to questions about the value of accessing genetics-informed dosing guidance, with one exception, 50% or more of the respondents indicated that they would value access to genetics-informed drug dose guidance across all demographic categories (for more details see “Value of Genetics-Informed Dosing Guidance for Obstetric Healthcare Providers”). The one exception was that one male respondent indicated that it a consult service would be valuable. Table 1 summarizes the demographic data of our study sample.

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Healthcare Providers (N=31) N (%)</th>
<th>% Value access to genetics-informed drug dose guidance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Yes, as an app</td>
<td>Yes, as a consult service</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>27 (87%)</td>
<td>78%</td>
</tr>
<tr>
<td>Male</td>
<td>4 (13%)</td>
<td>100%</td>
</tr>
<tr>
<td>Experience</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Resident</td>
<td>10 (32%)</td>
<td>90%</td>
</tr>
<tr>
<td>Attending</td>
<td>17 (55%)</td>
<td>76%</td>
</tr>
<tr>
<td>Other</td>
<td>4 (13%)</td>
<td>75%</td>
</tr>
<tr>
<td>Specialty</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Obstetrics Healthcare Provider</td>
<td>29 (94%)</td>
<td>83%</td>
</tr>
<tr>
<td>Gynecology practice only</td>
<td>2 (6%)</td>
<td>50%</td>
</tr>
<tr>
<td>Years since completing training</td>
<td>Median 10 (1-36)</td>
<td>Median 5 (1-36)</td>
</tr>
</tbody>
</table>

1336
Experiences with Adjusting Drug Dose in Pregnant Populations

Survey respondents indicated several patient characteristics that they have considered when making dose adjustments: twenty-six (81%) indicated that they use body weight, ten (31%) use height, fifteen (47%) use body surface area, twenty-seven (84%) use organ system function, five (16%) use genetics, and five (16%) indicated use of other characteristics. Among the twenty-seven respondents indicating that they use organ system function to adjust drug dose, free-text descriptions of specific organs included: kidney, liver, GI, respiratory (e.g., h/o asthma), blood volume, and brain/cognitive function. For the five respondents indicating that they have used other patient characteristics, free-text descriptions included: patient history, blood levels for some medications (e.g., anti-epileptics), medical co-morbidities and other medication use.

Dose adjustments are relatively common in pregnant women with nine (30%) of survey respondents indicating that dose adjustments occurred greater than or equal to 50% of the time when including changes due to pregnancy progression. When excluding dose changes due to pregnancy progression, only two (7%) indicated that that dose adjustments occurred greater than or equal to 50%. In instances where drug dose adjustments do occur, it is uncommon for drug doses to be modified beyond normal ranges, with twenty-three (74%) of respondents indicating that they modified drug dose beyond normal ranges less than 10% of the time (45%, or fourteen, indicated that they did so less than 1% of the time). However, in the instances where drug dose changes are made, seeking outside consultation is modest, with twenty-six (84%) of respondents indicating that they sought consultation less than 25% of the time (10%, or three, indicated that they never seek consultation).

Comfort in Prescribing Medications and Opinions about Using Genetics when Prescribing Medications

Obstetric healthcare providers were asked about their comfort in prescribing medications used to treat HIV, hypertension, and depression in pregnant women for which drug exposure is known to be influenced by genetics. For those who prescribe those medications, comfort in prescribing was split with more respondents indicating high comfort levels across all three categories of medications.

For nine antiretroviral medications, 35-45% do not prescribe, 26-29% are very uncomfortable or uncomfortable prescribing, and 26-39% are very comfortable, comfortable, or neither comfortable or uncomfortable with prescribing those medications (See Figure 1). The majority of survey respondents (58%, eighteen) also indicated a belief that genetics-informed dose adjustments was important or very important to achieve good outcomes when treating pregnant women with HIV. Thirteen (42%) respondents indicated that genetics-informed dose adjustments were moderately important, slightly important or not at all important.

Figure 1. Comfort in Prescribing Antiretroviral Drugs
For four antihypertensive medications, survey respondents indicated that 3-23% do not prescribe, 6-16% are very uncomfortable or uncomfortable prescribing, and 68-90% are very comfortable, comfortable, or neither comfortable or uncomfortable with prescribing those medications (See Figure 2). Slightly fewer than half of survey respondents (48%, fifteen) indicated a belief that genetics-informed dose adjustments was important or very important to achieve good outcomes when treating pregnant women with hypertension. Sixteen (52%) respondents indicated that genetics-informed dose adjustments were moderately important, slightly important or not at all important.

Figure 2. Comfort in Prescribing Antihypertensive Drugs

For fourteen antidepressant medications, among obstetric healthcare provider survey respondents 10-41% do not prescribe, 6-30% are very uncomfortable or uncomfortable prescribing, and 31-84% are very comfortable, comfortable, or neither comfortable or uncomfortable with prescribing those medications (See Figure 3). Slightly fewer than half of survey respondents (48%, fifteen) indicated a belief that genetics-informed dose adjustments was important or very important to achieve good outcomes when treating pregnant women with depression. Sixteen (52%) respondents indicated that genetics-informed dose adjustments were moderately important, slightly important or not at all important.

Figure 3. Comfort in Prescribing Antidepressant Drugs
**Awareness of and Access to Genetics-Informed Dosing Guidance**

Few survey respondents, ten (37%), were aware of any existing resources for clinical guidance on genetics-informed drug dosing. Of those indicating an awareness of any existing resources, only four indicated that they had used any. One respondent indicated as a free-text response that they had used FDA resources. For the six respondents indicating that they were aware of resources for guidance on genetics-informed dosing guidance, but that they had not used any, the following resources were listed as free-text responses: Micromedex®, REPROTOX®, UpToDate®, OMIM®, and PubMed. One respondent indicated that a patient’s lab report can contain guidance. Another respondent indicated that genetics-informed dosing guidance can be provided by genetic counselors.

Among ten respondents indicating an awareness of resources for genetics-informed dosing guidance, seven indicated a belief that guidance on genetics-informed drug dose use in pregnancy was moderately or slightly convenient to obtain. Three indicated that such guidance was not at all convenient. One respondent indicated resources for genetics-informed guidance in pregnancy are very useful, seven indicated that the resources are moderately or slightly useful, and two indicated that the resources are not at all useful.

**Value of Genetics-Informed Dosing Guidance for Obstetric Healthcare Providers**

In response to questions about access to genetics-informed dosing guidance, twenty-five (81%) indicated that they would like to have access to an app when seeking such guidance, four (13%) indicated that they may like to have access to an app, and two (6%) indicated that they would not like to have access to an app. Among twenty-five respondents indicating that they would like to have access to an app, they also indicated preferred modalities (they were able to select more than one): nine (36%) as an app that can be installed on a computer, seventeen (68%) as an app that is embedded in the EHR, and sixteen (64%) as a mobile app that can be installed on a tablet or cell phone.

Approximately one-half (48%-58%) of respondents perceived genetics-informed dose guidance to be important to achieve good outcomes when treating pregnant women with depression, hypertension or HIV (see Table 2). Regardless of the perceived importance, with one exception, 50% or more of the respondents indicated an interest in having access to genetics-informed drug dose guidance as an app or as a consult service. The one exception was that 44% healthcare providers perceiving genetics-informed drug dose guidance to be of low importance for anti-hypertensive medications indicated that it a consult service would be valuable (see Table 2).

**Table 2. Obstetric healthcare providers that value access to genetics-informed drug dose guidance, according to perceived importance**

<table>
<thead>
<tr>
<th>Perceived importance of genetics-informed dosing guidance</th>
<th>Healthcare Providers (N=31) N (%)</th>
<th>% Value access to genetics-informed drug dose guidance</th>
<th>Yes, as an app</th>
<th>Yes, as a consult service</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anti-depressant medications</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High</td>
<td>15 (48%)</td>
<td>93%</td>
<td>60%</td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>16 (52%)</td>
<td>69%</td>
<td>50%</td>
<td></td>
</tr>
<tr>
<td>Anti-hypertensive medications</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High</td>
<td>15 (48%)</td>
<td>93%</td>
<td>67%</td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>16 (52%)</td>
<td>69%</td>
<td>44%</td>
<td></td>
</tr>
<tr>
<td>Anti-microbial medications</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High</td>
<td>18 (58%)</td>
<td>94%</td>
<td>55%</td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>13 (42%)</td>
<td>61%</td>
<td>54%</td>
<td></td>
</tr>
</tbody>
</table>

The interest in having access to an app was higher among respondents indicating that genetics-informed dosing guidance was of high importance (important or very important) compared those indicating a low importance (moderate importance to not at all important, see Table 2 and Figure 4). When asked about interest in being able to access a consult service when seeking genetics-informed dosing guidance for pregnant women, sixteen (52%) indicated that they would like to have access and thirteen (42%) may like to have access. One respondent indicated that (s)he would not like to have access, and another indicated that (s)he already has access to a consult service. The interest in having access to a consult service is similar to or slightly higher than among respondents indicating that genetics-informed dosing guidance was of high importance compared to those indicating a low importance (see Table 2 and Figure 5).
Figure 4. Percent of Obstetric Healthcare Providers Indicating An Interest in Having Access to an App

![Bar Chart]

- Respondents indicating that genetics-informed dosing was of low importance
- Respondents indicating that genetics-informed dosing was of high importance

Figure 5. Percent of Obstetric Healthcare Providers Indicating an Interest in Having Access to Consult Services

![Bar Chart]

- Respondents indicating that genetics-informed dosing was of low importance
- Respondents indicating that genetics-informed dosing was of high importance

Discussion and Conclusion

This study obtained feedback from a sample of thirty-one Johns Hopkins obstetric healthcare providers. We found that dose adjustments are relatively common in pregnant women, and they are made using a combination of factors. Most obstetric healthcare providers use body weight and organ system function. Height and body surface area are used less often, and genetic data is rarely used. Also, obstetric healthcare providers who report making dose adjustments, infrequently report using outside consultation. We believe that as more evidence to support the use of genetics becomes available and readily accessible as clinical decision support, physician education will improve and genetic information use will increase.

When asked about comfort in prescribing medications for which drug exposure is known to be influenced by genetics, obstetric healthcare providers indicated that they do not prescribe many of the medications we asked about. For those indicating that they do use the medications, around half were uncomfortable with prescribing antiretroviral medications. The majority of those indicating that they used the antidepressants and antihypertensive medications were comfortable with prescribing them. However, over a quarter of respondents were uncomfortable with prescribing many antidepressant medications. While the first and second line therapies are widely used, and may be working well at the time a patient becomes pregnant, dosing adjustments may still be needed. Furthermore, even those physicians that are comfortable with individual drugs, may need assistance with prescribing specific drug combinations. Genetics-informed dose adjustments may help guide dose adjustments for individual drugs and combination regimens.
When asked about clinical guidance on genetics-informed dosing, few respondents indicated that they were aware of or had ever used any. Most would, however, indicated an interest in having access to genetics-informed dosing information as an app embedded in an EHR and as an app that can be installed on a tablet or cell phone. In addition, around half of respondents indicated an interest in having access to a consultation service for guidance on genetics-informed dosing.

Findings from this study on genetics-informed drug dose guidance are concordant with other studies in identifying a need for more clinical guidance on using pharmacogenomics data\(^{4,16}\). One study indicated that healthcare providers had a preference for a pharmacogenomics educational resource to be electronic and to include information on how to interpret pharmacogenomic test results, recommendations for prescribing, population subgroups most likely to be affected, and contact information for laboratories offering pharmacogenomic testing\(^{16}\). Similarly, obstetric healthcare providers also had a preference for guidance to be made available electronically. Though it is not yet standard of practice, several institutions have established programs to perform prospective pharmacogenomics testing and also provide the results of those tests in the medication treatment context with clinical decision support\(^{17,22}\). The current study provides insights into the clinical decision support needs of obstetric healthcare providers that use pharmacogenomics testing.

Our finding that dose adjustments are relatively common and most often occur due to pregnancy progression, suggests that pharmacogenomics should be considered secondarily to other more common characteristics that also influence drug exposure in pregnant patients (e.g., physiological changes). However, there may be some uncertainty to what will be the added value of presenting pharmacogenomics in the context of other characteristics given that this rarely occurs today. While there remains a lack of research on pharmacogenetics changes during pregnancy\(^{23}\), pharmacogenomic testing developed in other populations can potentially help guide its use in pregnant women. There are already a few clinical decision support tools to consider patient genetics that are targeted to obstetrics healthcare providers\(^{24}\). None of these tools to these authors knowledge, however, provide drug dosing guidance based upon models that consider pharmacogenomics. Thus, there is a particular need for dosing models that integrate multiple patient characteristics in this special population.

A large number of pregnant women who take medications could potentially benefit from improved drug dosing guidance. One study conducted in 2004 indicated that 49.9% of their sample of 152,531 pregnant women used medications from Food and Drug Administration (FDA) categories C, D, or X at some point during their pregnancy\(^{25}\). Categories C, D, and X were previously used to denote potential or unknown risks to a human fetus. Another study examining the use of medications during pregnancy in a sample of 578 women found that 59.7% used prescription medications, 92.6% used at least one over-the-counter medication, and 45.2% used at least one herbal medication\(^{4}\). Further indicating a need for improved guidance, in 2014 the letter-based category system was retired in favor of the Pregnancy and Lactation Labeling (PLL) Rule\(^ {26,27}\). The PLL rule summarizes risks and provides information about testing, contraception, and infertility to be used by the health care providers. The new labels, however, do not provide the detailed information needed to make prescribing decisions, which “will require focused efforts on the part of multiple stakeholders.”

The safety of medications is an even greater concern for pregnant women given the potential risks that some medications pose to the fetus. Research by Adam et al\(^ {28}\) found that, of 172 drugs tested for teratogenicity (harmfulness for a fetus), 168 (97.7%) could not be definitively determined to pose a risk for the fetus, and 126 (73.3%) of those drugs did not even have enough data to begin assessment. While outside of the scope of our survey focusing on maternal pharmacogenomics, prenatal pharmacogenomics studies are on the rise. As the influence of pharmacogenomics on fetal exposure to medications is better understood, that data could also begin to be made available to obstetric healthcare providers for use to manage their patients\(^ {29}\).

We found that obstetric healthcare providers at Johns Hopkins would find access to genetics-informed dosing guidance to be valuable. This work thus motivates future work to develop electronically accessible software to offer drug dosing guidance targeted to obstetric healthcare providers (or designated parties that provide consultation) based upon a range of patient characteristics, including genetics. The work of others has also indicated that pharmacogenomics data alone is not always sufficient to change test ordering and prescriber decisions\(^ {30,31}\), further supporting this need to consider multiple patient characteristics.

Despite the perceived value of access to genetics-informed dosing guidance, there is also the potential for confidence in prescribing decisions to be impacted. Previous research of one co-author, for example, found that physicians indicated a lowered confidence in prescribing decisions with access to pharmacogenomics clinical decision support\(^ {22}\). This finding was also reflected in the proportion of physicians who changed doses toward doses supported by
published evidence. Thus, another area for future research will be to monitor prescribing patterns to see what is the impact of providing access to genetics-informed dosing guidance. Findings from self-reports of comfort prescribing in this work have potential to provide some insight into prescribing pattern observations. However, we believe that ultimately new approaches to enable assessing the clinical utility of having a more complete patient picture in drug dosing decisions are needed. Indeed, there are circumstances where pharmacogenomics differences become more critical for prescribing decisions (when pregnancy significantly changes pharmacokinetics), and other circumstances where factors having nothing to do with the pharmacogenomics of these drugs influence those decisions (e.g., weight gain as a drug side effect).

Furthermore, like many needs assessment studies, the generalizability of findings is limited to the group studied. Our goal, however, was to understand the value of genetics-informed dosing guidance to obstetric healthcare providers at Hopkins so that in follow-up work we can design clinical decision support with potential to support obstetric healthcare practices more broadly. There are some parallels from our previous work investigating opinions of junior physicians (cardiology and oncology fellows)\(^2\). Work here, however, includes physicians with a range of experience (obstetrics residents and attending physicians). Differences in responses and prescribing patterns between the residents and the attending physicians in the obstetrics clinical domain will be another interesting area to explore.

References


Appendix A. Survey questions

1. Which of the following do you take into account in modifying drug doses for pregnant women?
2. How often do you have to seek consultation to confirm drug doses in pregnant women?
3. How often do you modify doses beyond the norm in pregnant women?
4. How often do you need to make dose adjustment in patients due to pregnant progress?
5. How often do you need to make dose adjustment in pregnant patients (excluding reasons due to pregnancy progress)?
6. Any other comments about experience and need for dose adjustment in pregnant patients?
7. How comfortable are you in treating pregnant women with the following antimicrobial drugs known to be influenced by genetics?
8. To what degree do you think genetics-informed dose adjustment is critical for good outcomes for the treatment of pregnant women with HIV in your practice?
9. How comfortable are you in treating pregnant women with the following antihypertensive drugs known to be influenced by genetics?
10. To what degree do you think genetics-informed dose adjustment is critical for good outcomes for the treatment of pregnant women with hypertension in your practice?
11. How comfortable are you in treating pregnant women with the following antidepressants known to be influenced by genetics?
12. To what degree do you think genetics-informed dose adjustment is critical for good outcomes for the treatment of pregnant women with depression in your practice?
13. Do you know of and use any resources for clinical guidance on genetics-informed drug dose use in pregnancy?
14. How convenient is it to obtain guidance on genetics-informed drug dose use in pregnancy? (Select one)
15. How useful are currently available information resources and tools on genetics-informed drug dose use in pregnancy?
16. Would it be valuable to have access to a tool that considers genetics in drug dose use in pregnancy?
17. Would it be valuable to have access to a consult service to provide genetics-informed dosing guidance for pregnant women?
18. Any other comments about available pharmacotherapy resources for genetics-informed drug dose use?
19. What is your gender?
20. What is your status as a clinical caregiver?
21. How many years since completing your training?
22. What is your specialty?
23. Where are you located?
24. What is your institution?
Evaluation of Semantic Web Technologies for Storing Computable Definitions of Electronic Health Records Phenotyping Algorithms

Václav Papež1,2,*, MSc, Spiros Denaxas1,2,*, PhD, Harry Hemingway1,2, FRCP
1 Institute of Health Informatics, University College London, London, UK
2 Farr Institute of Health Informatics Research, University College London, London, UK

Abstract

Electronic Health Records are electronic data generated during or as a byproduct of routine patient care. Structured, semi-structured and unstructured EHR offer researchers unprecedented phenotypic breadth and depth and have the potential to accelerate the development of precision medicine approaches at scale. A main EHR use-case is defining phenotyping algorithms that identify disease status, onset and severity. Phenotyping algorithms utilize diagnoses, prescriptions, laboratory tests, symptoms and other elements in order to identify patients with or without a specific trait. No common standardized, structured, computable format exists for storing phenotyping algorithms. The majority of algorithms are stored as human-readable descriptive text documents making their translation to code challenging due to their inherent complexity and hinders their sharing and re-use across the community. In this paper, we evaluate the two key Semantic Web Technologies, the Web Ontology Language and the Resource Description Framework, for enabling computable representations of EHR-driven phenotyping algorithms.

Introduction

Electronic Health Records (EHR) are structured, semi-structured and unstructured data that are generated during routine interactions of patients with primary care, hospital care and tertiary healthcare or as a byproduct of those interactions for billing or administrative purposes1. Structured EHR are recorded using controlled clinical terminologies while unstructured data include clinical text and narrative. Semi-structured EHR data often loosely follow a data specification (e.g. prescription events, medical imaging reports) but this varies greatly across information systems, clinical specialties and healthcare providers. High-throughput genotyping and increased availability of EHR data are giving scientists the unprecedented opportunity to exploit routinely generated clinical data to advance precision medicine at scale. EHR data can fundamentally alter the manner in which genetic association studies are performed and enable scientists to examine the association of genetic variants and traits in larger sample sizes and phenotypic breadth2.

A primary use-case of EHR data is the creation of phenotyping (or “case finding”) algorithms3, computational algorithms that identify patients that have (or have not) been diagnosed with a particular condition4 (e.g. acute myocardial infarction, prostate cancer, or anxiety etc.) and where applicable the disease onset and severity. Phenotyping algorithms tend to use clinical information such as diagnoses, laboratory tests, symptoms, clinical examination findings, prescriptions, referrals and other EHR data elements. While the term phenotype is traditionally defined as the physical manifestation of a particular trait, in the context of EHR research, phenotypes are broadly (but not exclusively) as the presence or absence of a particular clinical condition. In EHR resources linked with genetic data, such as the Electronic Medical Records and Genomics (eMERGE) consortium5, these phenotypes can enable large-scale genomic association studies which have been traditionally limited to a small set of traits. Phenotyping however is a challenging and time-consuming process since often data been collected for care, auditing or administrative purposes and not for research. The contents of EHR data sources are an indirect representation of the true patient state as skewed by the underlying healthcare process e.g. clinical guidelines, information systems, data standards6.

Defining and validating EHR phenotyping algorithms is challenging and time-consuming. Challenges are amplified by the lack of a common definition standard for algorithms, making their sharing across the scientific community problematic. Despite the fact that phenotype components are structured and often annotated by controlled clinical terminology terms, phenotype definitions, and their underlying logic are usually expressed as free-text which is not readily machine-readable. The translation from this narrative to programming code used to identify and extract patients (e.g. implementing a phenotyping algorithm using Structured Query Language for use in a relational database management system) can be problematic due to potential ambiguities in the manner in which the algorithm was expressed or potential ways of implementing it using local data. There is a clear and urgent need to develop and
evaluate a computable, standards-driven format to facilitate the systematic creation, sharing and re-use of EHR-derived phenotyping algorithms.

**Semantic Web Technologies (SWT)**

Semantic Web Technologies are a collection of World Wide Web Consortium (W3C) standards (https://www.w3.org) for annotating and sharing data using Web protocols and can potentially address some of these challenges and be utilized to define computable EHR-derived phenotypes. A key advantage of SWT is that they were specifically built to facilitate the automated integration and reuse of data in a machine-readable manner while in parallel enabling standard Web-driven user interaction in human-friendly formats. The Ontology Web Language (OWL) (https://www.w3.org/OWL/) is a specification of an ontology language based on a description logic. An ontology in OWL is mostly modeled as an RDF (Resource Description Framework) graph (https://www.w3.org/RDF/), a specification to describe data/resources in form of *triples* `<subject-predicate-object>`, where subject and predicate are resources and object is also a resource or a literal (value). Resources, uniquely identified by Internationalized Resource Identifiers (IRIs) which in turn are a generalization of Uniform Resource Identifiers (URIs), represent nodes (subjects and objects) and edges (relationships) of the graph structure and leaf nodes are created by literals (values) as for example in Figure 3.

The RDF Schema (RDFS) provides a set of basic predefined semantics (i.e. resources) for RDF graphs including data types and elements for the construction of classes and subclasses. OWL provides extended semantic expressions and constructs such as set operations, class disjoints, class equivalencies, and cardinality constraints. Semantic expressiveness (and thus a set of OWL constructs) is supplied by the OWL dialect and profile and influence the behavior of the *semantic reasoners*, engines able to automatically infer consequences and associations within the ontology which are not explicitly specified. The querying of RDF is enabled by the SPARQL Protocol and RDF Query Language (SPARQL) semantic language. SPARQL (https://www.w3.org/TR/rdf-sparql-query/) has a similar structure to Structured Query Language (SQL) which is widely used in relational database systems and enables queries for data stored as RDF through query processing engines called SPARQL end-points.

The use of SWT has been previously evaluated in the field of EHR phenotyping. The work by Pathak and colleagues was focused on the transformation of EHR data from multiple heterogeneous sources into RDF graphs. The created RDF graphs which were then queried via a single SPARQL end-point in order to assess research cohort. The research outlined in this paper is complementary to their work in that it is focused on the explicit description of the phenotype algorithm itself. EHRs are included into the process in the second step and outcome cohorts (or their parts) are inferred automatically by semantic reasoners. Compared to previous work, our approach describes how a phenotype algorithm and its logic can be defined by OWL constructs and queries can be implemented in SPARQL.

In this paper, we implement an OWL/RDF ontology-based approach for storing a deterministic EHR-derived diabetes phenotyping algorithm. We validate our approach against a pre-defined list of desiderata developed previously. We used diabetes as a case study since it exemplifies many of the associated challenges but our findings are generalizable to other diseases and syndromes.

**Methods**

**CALIBER**

We used a deterministic EHR phenotyping algorithm for diabetes developed and validated in the CALIBER resource. CALIBER is a translational research platform which links national, structured primary care, hospital care, disease registry, mortality data and socioeconomic information in the UK for ~10m patients. Primary care data are provided by the Clinical Practice Research Datalink, an anonymized national cohort of longitudinal data for all individuals registered with a general practitioner. Secondary care data are recorded in Hospital Episode Statistics (HES), a national database of administrative data used for hospital reimbursement. Finally, mortality and socioeconomic data are collected and curated by the Office of National Statistics (ONS).

Primary care data include diagnoses, referrals, symptoms, laboratory tests and clinical examination findings recorded using the Read controlled clinical terminology (a subset of SNOMED-CT, http://www.snomed.org/snomed-ct). Medication prescriptions are organized using the British National Formulary (https://www.bnf.org/), a structured resource for classifying all therapeutic agents prescribed in UK healthcare. Hospital care data include ranked diagnoses recorded using ICD-10 terms and interventional procedures recorded using the OPCS Classification of Interventions and Procedures version 4 (OPCS4) terms which are essentially semantically equivalent to the Current Procedural
Terminology (CPS) terms used in the United States. Mortality data are recorded using ICD-9 and ICD-10 and include the underlying cause of death and up to 15 contributory causes of mortality.

Diabetes phenotyping algorithm

Diabetes affects approximately 25 million people in the US and is one of the most common non-communicable diseases globally associated with a significant burden to patients and healthcare systems. Diabetes has a number of environmental (e.g. body mass index) and genetic risk factors and is one of the leading causes of heart disease and stroke. In addition, diabetes complications such as neuropathy, nephropathy, retinopathy and amputation have a substantial impact on the quality of life of patients. The ability of researchers to undertake studies in diabetes using electronic health records potentially linked to genetic data is of high value.

We used a previously validated EHR-derived phenotyping algorithm for identifying and classifying patients with diabetes into four distinct, non-overlapping groups: 1) patients with type 1 diabetes, 2) patients with type 2 diabetes, 3) patients with unspecified diabetes and 4) patients that are not diabetic. The algorithm (Figure 1) combines clinical information from specific diagnostic codes for type 1 and type 2 diabetes with less specific codes for 'insulin dependent diabetes' (IDDM) and 'non-insulin dependent diabetes' (NIDDM) recorded across both primary and secondary care. Patients with diacoses of both type 1 and type 2 diabetes are classified as patients with diabetes of unspecified type. The algorithm was designed to primarily identify patients with type 2 diabetes that can be on a variety of medications so it does not make use of medication data explicitly. Further refinements would be required to identify patients with type 1 diabetes with greater recall.

![Diagram of the CALIBER diabetes phenotyping algorithm](image)

**Figure 1.** CALIBER diabetes phenotyping algorithm: patients are classified in four distinct categories: type 1 diabetes, type 2 diabetes, diabetes of uncertain type and non-diabetic on the basis of diagnostic codes found in primary care EHR or hospital administrative data.

Individual phenotype components are stratified by data source - e.g. `dm_gprd` represents a diagnosis of diabetes from primary care, `dm_hes` a diagnosis from secondary care. Within each component, diagnostic terms from controlled clinical terminologies have been grouped by clinicians in terms of certainty e.g. `historical diagnoses, possible`
diagnosis, confirmed diagnosis. All individual components and definitions are made available through an open-access Data Portal (https://www.caliberresearch.org/portal/show/phenotype_diabetes).

Desiderata

We used the desiderata defined by Mo and colleagues in order to evaluate RDF/OWL for storing computable representations of phenotyping algorithms. In their work, the authors reviewed a series of EHR phenotyping algorithms (e.g. dementia, Crohn’s disease, cataract, HDL) developed as part of eMERGE and a series of authoring tools such as the Measure Authoring Tool (MAT) (https://www.emeasuretool.cms.gov/), i2b2 (https://www.i2b2.org/), and the SHARPn PhenotypePortal (http://phenotypeportal.org/). The authors propose a list of recommendations for desired features (desiderata) for computable phenotype representations models:

i. Support both human-readable and computable representations
ii. Implement set operations and relational algebra
iii. Represent phenotype criteria using structured rules (e.g. nested logical structure, Boolean logic, comparative operations, aggregative operations, and negation)
iv. Support defining temporal relations between clinical events
v. Utilize standardized controlled clinical terminologies and facilitate reuse of value sets
vi. Provide interfaces for external software algorithms or data components such as support for Natural Language Processing (NLP) operations for extracting clinically significant markers from unstructured EHR
vii. Maintain backwards compatibility to accommodate for temporal changes in the underlying healthcare process model and EHR data specifications.

Results

Phenotype ontology

As part of our work, a proof of concept application (Figure 2) for generating OWL phenotype ontologies was developed in Java using the OWL API, the Hermit reasoner (http://www.hermit-reasoner.com/) and the Protégé, software (http://protege.stanford.edu/). Phenotype ontologies were serialized into OWL/XML format by the semantic reasoners. We utilized SPARQL for defining and performing queries and additional operations, but the main algorithmic logic is defined by OWL constructs.

Figure 2. Diagram of the ontology-based phenotyping solution
We developed the following process to incrementally generate and extend the phenotype ontology using OWL DL:

1) **Generic structural core**: A generic ontology structure is predefined which essentially defines the generic structure for phenotyping algorithms.

2) **Phenotype components**: Individual components of phenotypes, such as lists of diagnostic terms associated with a diagnosis of diabetes in primary care or in administrative data in hospitals is processed. This results in the creation of phenotype element classes (i.e. component, category, diagnostic term) that will hold this information. Diagnostic term lists (*code lists*) associated with phenotype components (e.g. diagnoses of type 2 diabetes in primary care) are currently stored using a very simple bespoke format in CALIBER which enables their automatic processing and addition to the ontology as new classes and individuals. Terms from controlled clinical terminologies used in the algorithm specified as the literals can be defined using IRIs from external ontology/terminology repositories such as the Bioportal, the repository of biomedical ontologies (http://bioportal.bioontology.org/), which provides ontologies derived from e.g. SNOMED-CT, Read or ICD-10/9 vocabularies.

3) **Phenotype algorithm logic**: In this stage, the logic associated with a phenotyping algorithm is manually specified enabling the algorithm to be decomposed into sets and set operations. The phenotype algorithm is based on narrowing the initial set of patients to a final subset representing patients with type 1 diabetes, type 2 diabetes, and unknown type diabetes. Therefore, the algorithm can be described as a graph where nodes represent sets of patients (classes from the ontology perspective) and edges represent set operations and constraints. For example, a class defined as *subjects which have asserted at least 1 Read code of type 1 diabetes intersected with subjects which have asserted at least 1 Read code of type 2 diabetes* is a subclass of class of *subjects with unknown type diabetes*. Union of all subclasses of *unknown type diabetes* is equivalent to *subject with unknown type diabetes* class. With this step, the OWL phenotype description can be serialized. Class names are generated automatically as a composition of the original element names as they are presented in CALIBER and suffixes (and/or prefixes) of general class names.

4) **Subject classification**: In this final stage, structured EHRs are automatically transformed into RDF triples `<patient-obtained-code>` and added to the graph (Figure 3). Individual patients are asserted to the class `Subject` and individual diagnostic terms are asserted to the class `Code`. As the code and patient id are unique, Internationalized Resource Identifiers (IRIs) of those individuals are concatenated from ontology IRI and patient’s id/code. This step does not describe the phenotype itself, but it uses the phenotype for subject classification.
Figure 3. Part of the phenotype ontology including two imported subjects (patient pseudo-identifiers 10642 and 10643) and consequence inferences as generated from the reasoner, generated using Protégé 5. C100100 is the Read identifier for the term “Diabetes mellitus, adult onset, no mention of complication” and E10 is the ICD-10 term for “Type 1 diabetes mellitus”. DM GPRD is diagnosis of diabetes in primary care, DM HES is a diagnosis of diabetes in hospital care. Category 3 is “Insulin dependent diabetes” and Category 4 is “Non-insulin dependent diabetes”. Algorithm implementation details available on CALIBER Data Portal: https://www.caliberresearch.org/portal/show/dm_hes.

Once these steps are complete, the automatic semantic reasoner (e.g. Hermit, Fact++) can be applied over the ontology and new individual assertion axioms are inferred and stored in separate RDF graphs. From these inferences (Figure 4 shows inferred assertions of patients contained within a yellow rectangle), individual groups of classified patients (cohorts) can be extracted using SPARQL queries in which the additional constraints can be specified if required. OWL supports constructs specifying constraints as disjoint classes, complement classes, class closures or zero individual occurrence, but semantic reasoners operate under an open world assumptions e.g. they assume the truth value may be true irrespective of whether or not it is known to be true. As a result, to reach the exact patients subset which is constrained by non-existence of some elements, e.g. patients obtained no diagnostic term from secondary health care, additional restrictions into SPARQL queries has to be added. As an example, the semantic reasoner can infer all patients that have a diabetes type 1 diagnostic term and all patients that have both diabetes type 1 and type 2 diagnostic terms. Patients with only diabetes type 1 diagnostic terms can be query as an intersection of the previously mentioned classes and this intersection can be easily specified within the WHERE clause of SPARQL (Figure 6).
Figure 4. Part of the diabetes phenotype ontology. Blue boxes represent the core; green boxes represent the automatically imported elements; red boxes represent manually added elements; yellow box shows the inferred instances, which are otherwise asserted to the class subject

(i) Support both human-readable and computable representations

OWL/RDF directly supports various machine-readable serialization formats with XML-based serialization into OWL/XML (RDF/XML respectively) being the most common approach. A major advantage is the implicit compatibility with common XML parsers for integrating into external applications. Turtle is another widely used serialization format, which is designed specifically for RDF graphs (Figure 5). In contrary to OWL/XML, Turtle is more compact and easily readable for humans without requiring additional interpretation. Manchester syntax represents both human and machine readable serialization. While RDF/OWL represents named graphs, they can be easily expressed to humans in graphical representation (Figure 3).

```
:phenotype_diabetes_code rdf:type owl:Class ;
  owl:equivalentClass [ rdf:type owl:Class ;
    owl:unionOf ( :diabdiag_gprd_code
                :dm_gprd_code
                :dm_hes_code
    )]
  rdfs:subClassOf :code .
```

Figure 5. Example of Turtle serialization: part of the diabetes ontology and the definition of the diabetes code class used to denote terms from controlled clinical terminologies and their individual components. The class to hold diagnostic terms associated with diabetes diagnoses is a union between three classes of diagnostic terms specifically for primary and secondary care (e.g. diabdiag_gprd_code, dm_gprd_code, dm_hes_code).
(ii) Implement set operations and relational algebra

OWL natively supports set operations over defined classes and subjects. OWL allows us to define an operation of union, intersection and complement and to specify equivalent or disjoint classes and properties. SPARQL, a query language designed to query RDF graphs which grammar is similar to SQL supports relational algebra11. Figure 5 presents an example of OWL set operations of how the diabetes phenotype diagnostic terms are equivalent to a union of terms from three individual phenotype components.

(iii)-(iv) Structured and temporal phenotype rule representation

Nested logical structures can be created by merging individual structured RDF graphs. Comparative and aggregative operations are not implicitly supported by OWL itself, nevertheless SPARQL does support both as well as Boolean logic and negation and as a result, OWL constraints coupled with SPARQL and reasoners meet this criterion. As an example, the SPARQL query returning a set of patients with a type 1 diabetes diagnosis is illustrated in Figure 6. OWL/RDF does not directly support temporal relations, but since it natively provides a data-time datatype, the model for temporal relations could be designed and validity of the relation could be checked within individual SPARQL queries.

(v) Utilize standardized controlled clinical terminologies and facilitate reuse of value sets

One of the main driving principles of ontologies is reusability. The majority of controlled clinical terminologies enable the unambiguous reference of particular terms by a unique identifier and RDF graphs are designed to reference these resource elements by an IRI. Additionally, many controlled clinical terminologies (e.g. Read, ICD-10, ICD-9) which do not natively support IRI can be obtained in a form of ontology on e.g. Bioportal (https://bioportal.bioontology.org/). Presence of such ontologies allows us not only to use a standardized controlled clinical vocabulary, but use them in the context of RDF natively. In our use-case, the diabetes phenotype algorithm uses Read terms for primary healthcare diagnosis and ICD-10 terms for hospital diagnosis. Currently, terms are added to the ontology as the literals however they can be also added as IRIs to their external resources. For example, the ICD-10 term Insulin-dependent diabetes mellitus (E10) has an IRI in the Bioportal ICD-10 terminology resource http://purl.bioontology.org/ontology/ICD10/E10.

(vi) External software interfacing

Our solution does not natively support an interface for external software algorithms or data components as it only contains an explicit specification without the executive components. However commonly used APIs like OWL API provide powerful tools to build such interfaces. Moreover, any external software algorithms implemented in an API proposed for OWL/RDF could be applied on the phenotype in OWL/RDF. NLP is not natively supported by OWL/RDF however, NLP rules and lexicons can be described by and extended by the phenotype ontology graph12.

(vii) Maintain backwards compatibility
Backward compatibility on the level of versions is assured and OWL 2 is backward compatible to OWL 1. Additionally, as OWL/RDF are serialized into XML-based files, version control systems like Git (https://git-scm.com/) can potentially provide an effective manner to capture and control changes over time.

**Conclusion**

In this paper, we evaluated the use of SWT components such as OWL and RDF for storing computable representation of EHR-derived phenotyping algorithms and a proof of concept application using the OWL API and the Hermit reasoner was created. In our case study, we applied OWL in DL dialect on a deterministic diabetes EHR phenotype algorithm from the CALIBER resource. We propose a semi-automatic approach to constructing an ontology which serializes a phenotype description into a machine-readable XML-based file.

OWL/RDF has the potential to be a sufficient resource for storing phenotyping algorithms as it is versatile enough to meet most of the desiderata. The main significant limitation occurs in the case of supporting NLP methods for extracting clinical markers from unstructured text, which is not natively supported. On the other hand, NLP rules and lexicons for NLP engines could be specified in OWL and thus extend current phenotype ontology. Additionally, after ontology extension by structured EHRs in RDF, common reasoners like Hermit can automatically infer consequences and classify patients according to their diagnosis.

**Future work**

Future work includes evaluating the framework in additional phenotyping algorithms including complex phenotypes that make use of combinations of common EHR data such as diagnostic codes, lab measurements, prescription information or that rely on external data sources and integrate multimodal sources of clinical information (e.g. raw text found in patient charts or notes or medical images). Formal evaluation in terms of the algorithm implementation and it's accuracy in identifying patient cohorts should be undertaken and a comparison between other knowledge representation specifications and standards (e.g. the Guideline Interchange Format, GLIF13) would be very useful. As the automatic inferencing strongly depends on an algorithm logic specification which has to be added manually, a human-friendly interface should be implemented and formally evaluated to enable researchers to specify algorithmic logic easily and through a graphical user interface. Researchers should be allowed to specify algorithm logic (i.e., to create rules between classes) and ideally queries using SPARQL even without intimate knowledge of a programming language. Development of the user-friendly interface addresses one of many shortcomings of the semantic web14 and thus detailed investigation of these shortcomings (e.g. existence of competitive ontologies, performance loss for large datasets) and their influence on this work is highly beneficial for the future progress.

**References**


Tracking Health Related Discussions on Reddit for Public Health Applications

Albert Park, PhD¹, Mike Conway, PhD¹
¹Department of Biomedical Informatics, University of Utah, Salt Lake City, Utah

Abstract

We use Reddit to demonstrate social media’s potential for public health applications. First, we employ a lexicon-based approach to track the prevalence of keywords indicating public interest in Ebola, electronic cigarette, influenza, and marijuana. Second, to better understand the public reactions, we use the Latent Dirichlet Allocation algorithm, to identify either the general themes or motivations for extreme changes in the volume of discussion over time. We observe that discussions related to Ebola and influenza, infectious diseases of public health interests, surged when the first case of Ebola was diagnosed and a new strain of H1N1 influenza virus was confirmed in the United States. We also observed that discussions of a controversial health topic like marijuana increased with the announcement of a major change in United States federal policy. Discussions of electronic cigarette highlighted opportunities for better health education. Lastly, we discuss the implications of our findings for utilizing Reddit data for public health applications.

Introduction

Nearly two-thirds of American adults (65%) use social media: a nearly a tenfold increase in the past 10 years¹. Social media provides a platform for users to freely express their thoughts and provides an opportunity to interact with geographically dispersed likeminded individuals. These social media users discuss a wide variety of topics ranging from ordinary details of their daily life to information about infectious diseases of public health interest like Ebola². Due to the popularity and ubiquitous nature of social media, researchers advocate for utilizing social media for public health applications³–⁵. Public health agencies are in an early adoption stage of using social media for information distribution⁶. In addition to the substantial potential for using social media as a disease surveillance tool³–⁵ and means of information distribution⁶, social media also has the potential to provide other opportunities to improve public-health practice.

Studying the reactions or opinions of a population has traditionally involved nationally distributed data collection, such as surveys from government agencies. However, these methods are expensive, and perhaps more importantly, time consuming. Some researchers suggest that mining social media data can provide opportunities to reduce time and expense when understanding the reactions or opinions of a population on health issues⁷–¹⁰. For example, social media allows for accessing first person accounts of experiences⁷,⁸, public sentiments⁹, public knowledge¹⁰, and public attitudes¹⁰ that may help public health agencies and researchers to develop policies that improve public health outcomes. Moreover, social media can provide the contextual information and prevalence of public interests more efficiently than traditional public health methods. Tracking the prevalence of public interests and understanding the general public reactions and opinions on various health issues have the potential to expand the scope of public-health practice.

In this paper, we report on findings derived from social media data gathered from Reddit for the purpose of tracking the prevalence of public interests and understanding public reactions towards infectious diseases of public health interests like Ebola and influenza as well as controversial health issues, such as electronic cigarettes and marijuana. In fact, although Reddit is one of the most popular public social media platforms, it has been underutilized for public health applications. Reddit’s size and range of topics make it difficult to make use of the data without any knowledge of how the platform is used in practice. Thus, we aim to fill this gap in the literature with the current study and answer the following two research questions (RQ):

- (RQ1) Is Reddit an effective source for tracking the prevalence of public interests on infectious diseases (i.e., Ebola and influenza) and controversial health related issues (i.e., electronic cigarette and marijuana) over time?
- (RQ2) What do Reddit members discuss regarding these health issues (a) in times of elevated discussion volume or (b) in general, if the issues have a steady level of discussions?

The work described in this paper was exempted from review by the University of Utah's Institutional Review Board (IRB) [ethics committee] (IRB 00076188).
Background

A growing body of research has demonstrated the successful use of social media for public health applications\textsuperscript{11-13}. Often referred to as \textit{digital disease detection}\textsuperscript{1}, \textit{Infoveillance}\textsuperscript{4}, and \textit{digital epidemiology}\textsuperscript{5}, many studies have used Twitter data for applications in public health, primarily due to the real-time nature of the data. For example, Twitter data have been used to monitor or estimate influenza\textsuperscript{12,14}, seasonal allergies\textsuperscript{12}, alcohol sales and consumption\textsuperscript{15}, cholera outbreaks\textsuperscript{16}, earthquake\textsuperscript{17}, and smoking behavior\textsuperscript{18}, as well as to examine sentiment towards marijuana use\textsuperscript{19}. Although Twitter is highly popular and tweet analysis has performed well with the aforementioned topics, tweets provides relatively limited context due to a length limitation of 140 characters.

Other social media data, such as Facebook and online health community data, have also been mined to, for example, characterize and predict postpartum depression\textsuperscript{20}, classify opioid addiction phrases\textsuperscript{21} and predict adverse drug reactions\textsuperscript{22}. Google search queries allowed researchers to provide timely estimation of influenza rates\textsuperscript{23}. However, a previous study suggested that Facebook users are reluctant to discuss certain negative topics on Facebook, due to users’ desire to convey positive images of themselves\textsuperscript{24}. Online health communities can provide rich details of first person accounts of experiences\textsuperscript{25}, however, online health communities typically are single topic focused groups, often with a small number of members and attracting a substantial number of “lurkers”\textsuperscript{26} (i.e., individuals who participate without posting) and dropouts\textsuperscript{27}. Google search queries can be useful and timely, however, search queries are relatively limited in providing context and have been shown to overestimate disease rates, due (in part) to heightened media coverage\textsuperscript{28}.

Recently, Reddit, due to the availability of a public Application Programming Interface (API)\textsuperscript{29}, the capability of providing contextual information, and the support for throwaway accounts, has become a widely studied social media platform for controversial discussions. For example, using Reddit data, researchers have found empirical evidence that Reddit members openly discuss and exchange information support for potentially stigmatized issues like mental health illnesses\textsuperscript{30}, detected increases in suicidal content following reports of several celebrity suicides\textsuperscript{31}, identified distinct markers of shifts to suicidal ideation from mental illnesses\textsuperscript{32}, explored the relationship between social feedback and community participation\textsuperscript{33}, identified distinctive linguistic characteristics that are associated with mental illnesses\textsuperscript{34}, characterized smoking and drinking problems\textsuperscript{35}, and examined user experiences with different tobacco products\textsuperscript{36}. Thus, in this study, we explore Reddit’s utility as a data source for public health applications for tracking and understanding public opinions and reactions to health issues.

Data: Social Media Site

The data for this study is hosted in the popular social media platform, Reddit (http://www.reddit.com). We use Reddit to track and understand discussions of Ebola, influenza, electronic cigarettes, and marijuana for the following three reasons. First, Reddit is a highly active social media platform that had 83 billion page views from over 88,000 active sub-communities (subreddits) in 2015. Members of Reddit made over 73 million individual posts with over 725 million associated comments in the same year\textsuperscript{37}. Second, Reddit allows for throwaway and unidentifiable accounts that are suitable for controversial discussions, such as thoughts and feelings on electronic cigarettes and marijuana as well as epidemic concerns like Ebola and influenza that may be inappropriate or sensitive for identifiable accounts. Third, Reddit content is publicly available, in contrast to other health focused social media platforms like Facebook Groups or specifically health-focused online communities like PatientsLikeMe, where the content is typically not available on the open web.

Reddit members converse via a forum like platform. Reddit discussion consists of posts (i.e., a submission that starts a conversation) and associated comments (i.e., a submission that replies to posts or other comments) in various topical focused subreddits. Members who have achieved a certain status within the community are able to create new subreddits. For this study, we used a dataset\textsuperscript{38} released by a Reddit member. The dataset has been used in previous studies\textsuperscript{34,39,40}. The dataset for the current study is comprised of 239,772 (including both active and inactive) subreddits, 13,213,173 unique member IDs, 114,320,798 posts, and 1,659,361,605 associated comments that were made from October 2007 to May 2015.

Methods

RQ1. Is Reddit an effective source for tracking the prevalence of public interests on infectious diseases and controversial health related issues over time?

We used a lexicon-based approach to track discussions on Ebola, electronic cigarettes, influenza, and marijuana from all subreddits available in Reddit. First, we identified key terms associated with the topics of our interests. A
summary of key terms for each issue is shown in Table 1. Second, we preprocessed the entire dataset, which included converting text to lower case and removing punctuation. Third, to extract submissions (i.e., posts and comments) containing key terms from all available 239,772 subreddits, we employed a lexicon-based approach and extracted timestamps, comment or post IDs, member IDs, and subreddit IDs of the submissions. We extracted and included any partial matches in this process to cover a wide variation of terms. For example, a partial match of ‘cig’ can cover a variation of ‘cig’, ‘cigs’, ‘cigarette’, and ‘cigarettes’ for electronic cigarette. Fourth, we counted unique member IDs, subreddits, posts, and comments containing key terms. Fifth, we normalized the frequencies over time by dividing the frequency counts by the total number of the respective variables from all available subreddits for that period. Since the total number of submissions in Reddit generally increases over time, we report normalized frequencies over time counts.

RQ 2. What do Reddit members discuss on these health issues (a) in times of elevated activities or (b) in general, if the issues have a steady level of discussions?

Based on results of RQ 1, we created two scenarios deciding which time periods to further investigate for understanding the discussions on Ebola, electronic cigarette, influenza, and marijuana. (a) If the issue has a sudden elevated level of discussion, we investigated the time period in which the elevation occurs along with prior discussions of the same temporal length to understand the underlying causes for these sudden changes in public interest. Similar methods that contrast to prior time periods have been used to detect emerging topics\textsuperscript{41,42}. (b) If the issue has a steady level of discussions, we investigated the entire discussions on the issue to understand the main themes.

We used natural language processing (NLP) and language modeling for this research question. Due to the size of the dataset and range of topics discussed on Reddit, we used automated methods. Similar automated methods have been used in the health care domain to extract information and analyze data, and to enhance the personal health care experiences\textsuperscript{43–45}. First, we preprocessed the entire dataset as we did in RQ1. Second, to improve the language modeling results, we removed the URLs and comments and posts with less than 5 words, and then extracted nouns using Python Natural Language Toolkit (NLTK) package\textsuperscript{46}. The extracted nouns were used to create language models—a set of topics generated from document-level word co-occurrences for a given set of documents—using Latent Dirichlet Allocation\textsuperscript{47} (LDA) for the time period of our interests. We elected to use LDA, an unsupervised algorithm, due to the lack of a ground truth dataset. We considered each post and its associated comments as a single document.

One advantage of using LDA as opposed to other unsupervised clustering techniques is that the algorithm considers each document with multiple topics. A previous study of online health discussions suggested that discussions could have multiple topics due to topic drift\textsuperscript{48}. Thus, we employed LDA for this study. One disadvantage of using LDA is, however, it requires a pre-determined number of topics. After experimenting with varying numbers of topics, we generated 50 topics to understand Ebola, electronic cigarette, influenza, and marijuana related issues. We used the Python package genism\textsuperscript{49} to conduct LDA analysis. We then present the main topics and their top 50 associated words as the word cloud overview using the Python package wordcloud\textsuperscript{50}. Despite its simplicity, word cloud overview remains one of the more preferred and user-friendly visualizations that can also scale to different data sizes\textsuperscript{51}. We then manually investigated the identified topics and their associated words to thoroughly examine the LDA results.

Lastly, we performed two types of validity checks. First, for health issues with a sudden elevated level of discussion, we verified the LDA results via a systematic analysis of news at the time of the change. LDA results reflect motivations for the extreme changes, thus news can be an effective source for a validity check. Second, we extracted URLs using regular expressions and categorized the results. A previous study concerning electronic cigarettes—a product with few marketing restrictions in the US until recently—suggested that up to 90 percent of social media (in this case, Twitter) content could be related to product marketing\textsuperscript{52}. Thus, because marketing content can skew our result, we used URLs as a proxy to marketing content and reported the percentage of posts with URLs. We also manually examined several extracted URLs to ensure the quality of the validation process.

<table>
<thead>
<tr>
<th>Issues</th>
<th>Key terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ebola</td>
<td>ebola</td>
</tr>
<tr>
<td>Electronic cigarette</td>
<td>e cig, elec cig, electronic cig</td>
</tr>
<tr>
<td>Influenza</td>
<td>flu, influenza, H1N1</td>
</tr>
<tr>
<td>Marijuana</td>
<td>weed, marijuana, ganja, cannabis, bong, spliff, Mary Jane</td>
</tr>
</tbody>
</table>

Table 1. Key terms used in the lexicon-based approach
Results

RQ1. Is Reddit an effective source for tracking the prevalence of public interests on infectious diseases and controversial health related issues over time?

The lexicon-based approach identified Reddit posts, comments, and members discussing Ebola, electronic cigarette, influenza, and marijuana from October 2007 to May 2015 (Table 2). The most discussed matter was influenza, followed by marijuana, electronic cigarettes, and then Ebola. The raw counts of discussions and members who mentioned each topic generally increased with time.

Table 2. The total number and average normalized count of posts, comments, members, and subreddits identified using the lexicon-based approach

<table>
<thead>
<tr>
<th>Issues</th>
<th>Total posts and comments (n)</th>
<th>Average normalized count of posts and comments (%)</th>
<th>Total members (n)</th>
<th>Average normalized count of members (%)</th>
<th>Number of subreddits containing the key terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ebola</td>
<td>252,243</td>
<td>7.18E-05</td>
<td>113,546</td>
<td>9.68E-04</td>
<td>6,039</td>
</tr>
<tr>
<td>Electronic cigarette</td>
<td>355,839</td>
<td>2.17E-04</td>
<td>176,252</td>
<td>3.75E-03</td>
<td>4,454</td>
</tr>
<tr>
<td>Influenza</td>
<td>6,876,684</td>
<td>4.48E-03</td>
<td>1,443,223</td>
<td>0.06</td>
<td>30,856</td>
</tr>
<tr>
<td>Marijuana</td>
<td>4,809,337</td>
<td>3.31E-03</td>
<td>968,892</td>
<td>3.75E-02</td>
<td>18,236</td>
</tr>
</tbody>
</table>

We identified one notable increase in discussion each for Ebola, influenza, and marijuana using the normalized frequencies over time (Figure 1). First, the normalized count on marijuana almost doubled from the previous month in February 2009. The heightened level of discussions continued for two months then slowly dropped back to the previous level. Second, in April of 2009, the normalized count on influenza almost doubled from the previous month. Third, October 2014 accounts for the Ebola discussions. The discussions on Ebola showed the most increase, jumping more than five times from the previous month. The number of members discussing each issue increased in a similar manner (Figure 1). The Discussions on electronic cigarette was relatively steady from October 2007 to May 2015.

Figure 1. The Line Graphs of normalized frequencies over time for posts and comments with key terms and members who used the key terms
Discussions on Ebola, electronic cigarettes, influenza, and marijuana, however, only accumulated to a fraction of the overall discussions on Reddit (Table 2). Although the community as a whole did not frequently talk about these health-related issues, this still amounted to more than 3,000 members for Ebola, the least discussed issue, and more than 137,000 members for influenza, the most discussed issue in a month with a normal level of discussion, May 2015.

**New subreddits to discuss Ebola, Electronic Cigarette, Influenza, and Marijuana**

Members of Reddit created a number of subreddits specifically focusing on Ebola, electronic cigarettes, influenza, and marijuana, although they also discussed the issues in many different subreddits (Table 3). Using the key terms (Table 1), we detected a total of 450 topically dedicated subreddits that were created between October 2007 and May 2015. For example, marijuana was casually discussed in 18,236 subreddits (i.e., subreddits with key terms in posts or comments), while members created at least 244 subreddits (i.e., key terms in names of subreddits) to talk about marijuana.

**Table 3. Newly created communities dedicated to focus on Ebola, Electronic Cigarette, influenza, and Marijuana**

<table>
<thead>
<tr>
<th>Issues</th>
<th>Subreddits, n</th>
<th>Example Subreddits</th>
</tr>
</thead>
<tbody>
<tr>
<td>Electronic cigarette</td>
<td>3</td>
<td>Ecigclassifieds, ecigclassifiedseu, ecigclassifiedsuk</td>
</tr>
<tr>
<td>Influenza</td>
<td>166</td>
<td>Influenza, Birdflu, flu,</td>
</tr>
<tr>
<td>Marijuana</td>
<td>244</td>
<td>Marijuana, ganja, cannabiscultur, ganjaoutlaw, cannabis_marijuana</td>
</tr>
</tbody>
</table>

**RQ 2. What do Reddit members discuss on these health issues (a) in times of elevated activities or (b) in general, if the issues have a steady level of discussions?**

![Figure 2. Word cloud overviews of emerging topics for Ebola (top left), influenza (top right), and marijuana (bottom left) as well as general word cloud overviews for electronic cigarette (bottom right)](image-url)
From RQ1, we learned that discussions focusing concerning Ebola, influenza, and marijuana, each had one sudden increase of activities. Thus, we created word cloud overviews of emerging topics for Ebola, influenza, and marijuana, while creating a general word cloud overviews for electronic cigarette (Figure 2).

According to the word cloud overview generated by the LDA topic modeling algorithm, we can infer that Reddit members are most concerned about ‘risk’ and ‘symptoms’ regarding Ebola. For influenza, members used terms like ‘Mexico’, ‘Obama’, ‘CDC’, and ‘conspiracy’, along with H1N1 influenza related terms (e.g., ‘H1N1’, ‘Swine’) as well as H5N1 related terms like ‘Egypt’ and ‘pig’. Topics regarding ‘legalization’, ‘prohibition’, ‘economy’, and ‘state’ appeared in discussions regarding marijuana. The general word cloud overview for electronic cigarettes has more commercially related terms such as ‘quality’, ‘prices’, ‘shop’, and ‘store’ than the other three discussions, however substantially more terms related to tobacco (e.g., ‘tobacco’, ‘cigarette’, ‘cigar’) are shown in Figure 2. Other notable topics for electronic cigarette that were identified via the LDA were ‘quitting smoking’, ‘fun experience’, and ‘health information’.


To check the validity of the results, we extracted and investigated the URLs to ensure that frequencies are not inflated by marketing content. The types of URLs shared by members were similar in nature for all four issues. Members shared websites that are concerning information (e.g., Wikipedia, CDC), news (e.g., NY Times), personal stories (e.g., blogs), other social media platforms, (e.g., Youtube), different Reddit posts, and commercial resources (e.g., amazon). Although the proportion of each type of URLs is different, members shared a relatively small number of posts and comments with URLs compared to the overall posts and comments focusing on all four issues.

Discussion

Principal Findings

We examined four different infectious disease related or potentially stigmatized health related issues discussed on Reddit. We discovered three periods with higher levels of activities on Reddit. We observed that there were almost twice as many marijuana related discussions in February 2009 compared to the previous month, due – we suspect – to the announcement of a major shift in federal policy. Attorney General Eric Holder confirmed that Drug Enforcement Administration would halt medical marijuana raids and give states the power to regulate medical marijuana usage for pain control in February of 2009. In April of 2009, discussions about influenza almost doubled from the previous month. This is likely due to the fact that a novel strain of H1N1 influenza virus was discovered in North America in the spring of 2009 and the Centers for Disease Control and Prevention (CDC) confirming the first two cases of human infection with H1N1 influenza virus in the United States in April of 2009. On September 30, 2014, the United States had its first diagnosed case of Ebola in Texas, and the first Ebola related death on October 8, 2014. We observed that discussions on Ebola, a potentially fatal infectious disease, surged more than five times from the previous month in October of 2014. The news related to Ebola, influenza, and Marijuana align well with the results from topic model analyses (RQ2). On the basis of these changes of activities, Reddit may be a valuable source of data for tracking the prevalence of public interests on infectious diseases (i.e., Ebola and influenza) and controversial health related issues (i.e., electronic cigarette and marijuana) over time (RQ1).

The result of our analysis on electronic cigarette discussions suggests that Reddit contains more than just commercial content despite the fact there are at least three subreddits focusing on classified content (Table 4). For instance, a subreddit called ‘Ecigclassifieds’ consists mainly of commercial content, thus the content of these subreddits deserves further investigation to better utilize the data. From electronic cigarette discussion, we identified three topics, ‘quitting smoking’, ‘fun experience’, and ‘health information’ that highlighted opportunities for better health education. From their associated terms (see Results), we can infer that Reddit members are seeking

<table>
<thead>
<tr>
<th>Issues</th>
<th>URL, n</th>
<th>Percentages of posts/comments with URLs to the total number of posts/comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ebola</td>
<td>32,863</td>
<td>13.03%</td>
</tr>
<tr>
<td>Electronic cigarette</td>
<td>22,839</td>
<td>6.42%</td>
</tr>
<tr>
<td>Influenza</td>
<td>783,350</td>
<td>11.39%</td>
</tr>
<tr>
<td>Marijuana</td>
<td>390,675</td>
<td>8.12%</td>
</tr>
</tbody>
</table>

1360
information on these three topics. Information seeking behavior on Reddit suggests Reddit’s utility as another social media platform for information distribution and as a data source for understanding user groups (e.g., electronic cigarette smokers) and identifying better health education. Why members are seeking health information on Reddit is an unanswered research question, although a recent study suggests that electronic cigarette related health information from public health agencies may be too difficult for the general public to comprehend.

Reddit members also created at least 450 relevant new subreddits specifically focusing on these four issues. How the content from these subreddits contrast with the content from multiple subreddits on the same issue is an unanswered question. Previous studies\textsuperscript{30,31,33,34,39} analyzed content from a handful of especially dedicated subreddits for their studies. However, our finding suggests that at least for discussions of Ebola, influenza, electronic cigarettes, and marijuana, members mentioned these issues on thousands of subreddits (Table 3). For instance, a common issue like influenza was discussed in over 30,000 subreddits, and even a focused topic like Ebola were discussed in over 4,400 subreddits. Thus, we believe analyzing a wider number of subreddits can improve recall of the relevant content.

Limitation, Future Directions, and User Privacy

Reddit offers substantial potential for understanding the public reactions to health-related topics, however, not without a number of limitations. Although Reddit is a widely-used platform, it is more frequently used by young males\textsuperscript{38,59} and may be subjective to self-selection bias. Reddit members are not necessarily representative of the general public, however, the levels of activity on Reddit aligned with the United States news and deserve a further investigation, especially with respect to location of postings and the overall reactions in Reddit. To better understand the reaction of the general public, studying different platforms and avenues, Facebook and Twitter for example, is warranted. Our analysis suggests that given the increasing popularity and use of Reddit, as well as the increasing frequency of discussions concerning our topic of interests, Reddit provides a productive starting point for investigating infectious disease related or controversial health issues.

Another limitation lies in the methodology. In RQ1, we used a relatively rudimentary lexicon-based approach to extract posts and comments explicit mentioning variations of pre-specified key terms. One major shortcoming of such approach is the selection of key terms. For example, utilizing a large set of key terms will undoubtedly create more false-positives, whereas too limited a set of key terms will surely result in more false-negatives. Moreover, partial matches can produce false-positive matches. We believe the figures for influenza were inflated because ‘flu’ can be a part of a longer word such as ‘fluorine’ or ‘flute’. In future studies, we suggest that precision rather than recall should be emphasized in order to eliminate irrelevant discussions. Other difficulties in mining social media data include the fact that social media text is frequently characterized by extensive of acronyms, abbreviations, and slang terms.\textsuperscript{60} Although we included the most frequently found abbreviations and slang terms, lexicon-based approaches are to omit unknown forms of abbreviations and slang. More sophisticated methods utilizing knowledge-based\textsuperscript{61} or corpus-based\textsuperscript{62} approaches could produce different results. Furthermore, a smaller timeframe can better measure the timeliness of the observed reactions as oppose to the one month timeframe used in RQ1. In RQ2, we relied on a systematic analysis of the news to verify the result of our investigation. However, data driven qualitative analysis\textsuperscript{63} can further bolster our findings and provide the contextual information on the discussions of our interests. Sentiment analysis on the extracted discussion can also provide further clues about general public reactions on various health related topics\textsuperscript{9}.

Research and applications using social media data should be highly sensitive to user privacy, especially for potentially stigmatized topics. Although at least some social media data are publicly available, researchers should consider ethical implications when processing data even for population-level social media research using public data\textsuperscript{64–66}. For this reason, we have refrained from using direct quotations from Reddit users in this paper.

Conclusion

As evident by the frequencies over time of discussions, inflated discussions after major news, as well as newly created subreddits specifically focusing on these health-related issues, Reddit could be a useful platform for understanding the concerns and opinions of the general public, especially for issues focusing on controversial topics, such as abuse and addiction as well as infectious diseases of public health interest. By utilizing the content, we also identified opportunities for better health education that could improve public health outcomes. We created topic models using LDA and generated topically associated words and created word cloud visualizations to show (1) emerging topics by contrasting to the prior topic models or (2) main themes of the discussions. We believe our insights and analyses can be generalized to other similar health related issues in the Reddit platform. Understanding public reactions to these issues has the potential to expand the scope of public-health practice.
Acknowledgments
We restricted our analysis to publicly available discussion content. The study was exempted from review by the University of Utah’s Institutional Review Board (Ethics Committee) [IRB 00076188].

Author AP was funded by National Library of Medicine of the National Institutes of Health under award number T15 LM007124. Author MC’s contribution to this research was supported by National Library of Medicine of the National Institutes of Health under award numbers R00LM011393 & K99LM011393.

The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References


38. Reddit_Member. I have every publicly available Reddit comment for research. ~ 1.7 billion comments @ 250 GB compressed. Any interest in this? [Internet]. 2015. Available from: https://www.reddit.com/r/datasets/comments/3bxlg7/i_have_every_publicly_available_reddit_comment/; Archived at: http://www.webcitation.org/6kgAuNxDE


53. Johnson MA. DEA to halt medical marijuana radis: Holder confirms states to have final say on use of drug for pain control. nbcdnws.com. 2009;
Mining Hierarchies and Similarity Clusters from Value Set Repositories
Kevin J. Peterson, MS\textsuperscript{1}, Guoqian Jiang, MD, PhD\textsuperscript{2}, Scott M. Brue, BS\textsuperscript{1}, Feichen Shen, PhD\textsuperscript{2}, Hongfang Liu, PhD\textsuperscript{2}

\textsuperscript{1} Division of Information Management and Analytics, Mayo Clinic, Rochester, MN
\textsuperscript{2} Department of Health Sciences Research, Mayo Clinic, Rochester, MN

ABSTRACT
A value set is a collection of permissible values used to describe a specific conceptual domain for a given purpose. By helping to establish a shared semantic understanding across use cases, these artifacts are important enablers of interoperability and data standardization. As the size of repositories cataloging these value sets expand, knowledge management challenges become more pronounced. Specifically, discovering value sets applicable to a given use case may be challenging in a large repository. In this study, we describe methods to extract implicit relationships between value sets, and utilize these relationships to overlay organizational structure onto value set repositories. We successfully extract two different structurings, hierarchy and clustering, and show how tooling can leverage these structures to enable more effective value set discovery.

INTRODUCTION
Controlled terminologies are the semantic underpinnings of clinical data and facilitate interoperability\textsuperscript{1}, research\textsuperscript{2}, and quality reporting\textsuperscript{3}. These terminologies are shared knowledge assets, often designed to be used for a variety of purposes and use cases\textsuperscript{4,5}. A value set is a grouping of codes from one or more terminologies used to express some conceptual domain\textsuperscript{6,7}. Value sets narrow the broad semantics of controlled terminologies down to a more targeted domain, and have a variety of practical applications related to data standardization and analysis\textsuperscript{8,9,10}. Like controlled terminologies, value sets are shared assets, generally published in a repository or catalog to promote discoverability and reuse.

As a value set is intended to semantically express some domain of interest, it is important that it adequately convey to the consumer the semantics to which it is bound – or its intension. If this intension is not clearly stated, value set discoverability, or a user’s ability to locate and reuse value sets relevant to their use case, will suffer. As reuse becomes more difficult, users are increasingly likely to simply create a new value set according to their needs. This compounds the problem, resulting in the proliferation of many value sets that are conceptually similar but with slightly different sets of codes. Consequently, it then becomes difficult for a user to infer whether or not these subtle differences were intended and necessary, or simply a result of different authors interpreting the same conceptual space differently. At the extreme, there may also be cases where value sets are inadvertently duplicated\textsuperscript{11}. High redundancy may be a symptom of low reusability – a problem not dissimilar to challenges in software reuse\textsuperscript{12}.

So how, then, do we determine the intension of a value set in order to promote reuse? If left as an exercise for the repository user, the most accessible points of inspection are the value set name and its set of codes. Although generally intended to be descriptive, a text-based value set name places the burden of interpretation on each individual user. Moreover, the name may lack precision, as names have shown to be generally insufficient at expressing complex semantics\textsuperscript{13}. Manual review of the contained code set may provide a better representation of actual intension, but is not without its own challenges. By browsing the code set, a subject matter expert may be able to sufficiently reverse-engineer the value set intension, but this is a manual, potentially laborious process. Furthermore, value set repositories tend to organize content in flat (or nearly flat) structures. This makes it difficult to display and search for similar groups of value sets, or to scope searches to a specific context. More importantly, however, it keeps implicit relationships between value sets hidden, and awareness of these inferred connections is a valuable tool for discovery\textsuperscript{14}.

The purpose of this study is to utilize automated methods of capturing value set intension\textsuperscript{11} to extract additional structure and knowledge from a value set repository. We specifically aim to extract value set clusters based on similarity, and hierarchies based on specialization/generalization. By placing value sets in the context of similar ones, we aim to promote better repository search capability\textsuperscript{15}, and ultimately, better discoverability. Finally, we will show how these methods may be practically implemented and integrated into tooling.
MATERIALS AND METHODS

Extracting additional structure and knowledge from a value set repository is a multi-faceted activity. Specifically, our choice of algorithms and outputs must be user-centered, or oriented towards defined user needs and requests. Also, we cannot assume universal value set structure, as several standardized formats exist along with many non-standard, proprietary representations. Heterogeneously structured value sets must be consolidated via preprocessing and standardization to allow our methods to have the maximum breadth of applicability. Finally, analysis of the results, or the representation of the new knowledge, must be actionable and approachable by the user.

Knowledge Discovery in Databases (KDD) is an effort to place the activity of data mining within the context of a larger process encompassing domain analysis, data preprocessing, and ultimately knowledge presentation and utilization. We structure our methods generally following the guidance of the KDD process.

Data Preprocessing

Selection. The Value Set Authority Center (VSAC) is a National Library of Medicine (NLM) public repository for value sets referenced in Meaningful Use Clinical Quality Measures (CQMs). The VSAC also includes support for value set authoring, expanding its scope beyond Meaningful Use and into other applications and domains. We use this rich value set repository as our primary source of data.

Standardization. The VSAC exposes value sets programmatically via the VSAC v2 REST service, with value sets modeled using Sharing Value Sets (SVS), an Integrating the Healthcare Enterprise (IHE) initiative focused on value sets and their representation. To allow our methods to be portable to repositories beyond the VSAC, we implemented our algorithms to operate on value sets modeled using Common Terminology Services 2 (CTS2), an Object Management Group (OMG) and Health Level Seven (HL7) terminology standard. This necessitates an SVS to CTS2 conversion step in our process for value sets originating from the VSAC (see: https://gist.github.com/cts2/ for CTS2 ↔ SVS implementation details). Our methods are also specifically scoped to extensional value sets, or value sets represented as enumerated lists of codes.

ISO/IEC 11179 is a standard for representing metadata in registries with the goal of promoting standardization and interoperability of data. Part 3 of the specification describes the basic elements of the metadata registry metamodel, and is specifically applicable to our problem space, as it provides structure and formality around our notion of a value set and its contents. For our purposes, we focus on the following elements of the metamodel:

- **Enumerated Concept Domain.** A collection of valid meanings – in this case, a set of explicitly enumerated Value Meanings. This aligns with what we refer to in this study as a value set. Example: VSAC value set BMI Values (2.16.840.1.113883.3.600.1.889)

- **Value Meaning.** A representation of the semantic intension, or meaning of a value. For our purposes, this is a code drawn from a standard terminology. Example: ICD10CM Z68.1, Body mass index (BMI) 19 or less, adult

An Enumerated Concept Domain is a specialization of a Concept, defined as a “unit of knowledge created by a unique combination of characteristics.” This is notable for our study, as it allows us to model Enumerated Concept Domain → Enumerated Concept Domain (or, value set → value set) relationships. It is through ISO/IEC 11179 that we base our assumption that value sets are themselves a unit of knowledge, have intrinsic meaning, and can be meaningfully related to other value sets.

Algorithm Selection

Our aim is to extract meaningful structure from a value set repository, and we focus our efforts on two structural constructs: hierarchy and clustering. Algorithms for extracting these structures from a repository are described below.

**Containment Hierarchy.** Hierarchies, or the arrangement of entities into parent → child relationships, have long been used to structure knowledge artifacts and play important roles in knowledge discovery. Concept hierarchies
have strong mathematical formalizations\textsuperscript{27} rooted in concept maps and are an important part of knowledge management\textsuperscript{28}. These hierarchies, however, are not always explicitly stated (or even envisioned by the content authors). Text document repositories\textsuperscript{29}, Wikipedia entries\textsuperscript{30}, text books\textsuperscript{31}, and on-line dictionaries\textsuperscript{32} are all examples of domains where hierarchies have been extracted via mining the existing data sources.

Containment hierarchy is a general strategy for ordering sets based on strict subsets – for example: \{a, b, c\} \supset \{b, c\} \supset \{c\}. For our purposes, the sets under examination are the code sets for each value set. By recursively nesting subsets, we can begin to build hierarchies. Examples of this hierarchy can be seen in Figure 1.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{containment_hierarchy.png}
\caption{(a) Each graph node represents a set of elements organized with children nodes as strict subsets of the parent.
(b) A practical example highlighting three value sets. Here, \textbf{ProcedureActStatusCode} and \textbf{ProblemActStatusCode} are strict subsets of \textbf{ActStatusCode}.}
\end{figure}

\textbf{Clustering.} Jansen et al. demonstrated that users searching the web for information tend to use short (2.35 terms) queries\textsuperscript{33}. As searching is a fundamental pillar of knowledge management infrastructure\textsuperscript{34}, it is useful for us to take this behavioral model into consideration for our study. Specifically, we recognize that a small number of search terms does not provide enough context to sufficiently pinpoint the desired results\textsuperscript{35}.

Clustering search results around a conceptual topic has been extensively studied as a way to improve search quality\textsuperscript{36,37}. These approaches focus on grouping semantically similar documents together in order to tune the search to the perceived semantic intent of the user. Much like containment hierarchy, clustering aims to extract implicit relationships between value sets. Unlike hierarchy, however, we are not limited to containment relationships. Specifically, we can begin to envision a value set repository as an undirected graph, where the value sets are the nodes and the edges indicate some relationship between them.

Using these relationships, we can begin to analyze groups of closely connected value sets, or clusters, using community discovery approaches\textsuperscript{38,39}. We define a cluster \(C\) focused on a node \(i\) in some graph \(G\) to be \(i\) and all nodes directly connected to it, or its closed neighborhood, denoted as \(C_i = N_G[i]\). One interesting characteristic of these clusters is how tightly connected they are – or, how probable it is that the neighbor nodes of \(i\) are also connected. Densely connected graphs appear in several contexts, ranging from social media connections\textsuperscript{40} to cell biology\textsuperscript{41}. We measure the density of these connections by calculating the clustering coefficient, a measure of how connected the neighbors of a node are to one another\textsuperscript{42}. Given a node \(i\) in graph \(G\), let \(T_i\) represent the number of edges in \(G\) that connect any two neighbors of \(i\), and let \(K_i\) represent the degree, or number of outgoing edges of \(i\). The clustering coefficient \(CC_i\) can then be calculated for any node \(i\):

\begin{equation}
CC_i = \frac{2T_i}{K_i(K_i - 1)}
\end{equation}

Note that the clustering coefficient may only be calculated for a cluster of three or more nodes. As such, only groups of three or more related value sets were considered to be clusters.
Figure 2: Example Clustering Coefficient ($CC_i$) values for three sample graphs.

Figure 2 shows this calculation as applied to three example graphs. Note that as this measurement is focused around a single given node, it is referred to as the *local* clustering coefficient. We can compute the *global* clustering coefficient $CC$ of the entire graph by averaging the clustering coefficients of each node:

$$CC = \frac{1}{n} \sum_{i=1}^{n} CC_i$$

(2)

Thus far, we have considered the edges in our graph to be binary – either two value sets are connected or they are not. To properly reflect our domain, we must account not only for the relationship itself, but the intensity of the connection – in our case, the amount of similarity between two value sets. Assuming we have in place algorithms to compute this similarity, we now are able to treat value set similarity not as a binary condition, but as a degree. We reflect this in our graph by using *weighted* edges. With this change we must also utilize a *weighted* clustering coefficient\(^{43,44}\), which is an adjustment to our original formula to incorporate edge weights. This is similar to the original clustering coefficient equation, but incorporates edge weights ($w$) by calculating the geometric mean of the weights of all triangles originating from the focus node $i$. Note that weights were normalized ($\hat{w}$) to a value in the closed interval $[0, 1]$ by dividing the raw weight by the maximum weight found in the graph: $\hat{w}_{k,i} = \frac{w_{k,i}}{\max(w)}$.

$$\hat{CC}_i = \frac{2}{K_i(K_i - 1)} \sum_{k,j} (\hat{w}_{k,i}\hat{w}_{j,i}\hat{w}_{k,j})^{1/3}$$

(3)

For each cluster, an accompanying Erdős-Rényi random graph\(^{45}\) was created, with $\hat{CC}_i$ calculated similarly for this randomly assembled cluster. If $\hat{CC}_i$ for the actual cluster was higher than what was observed in the random cluster, we considered that cluster to be *dense*\(^{42}\).

**Defining Value Set Similarity**

Although our methods now account for varying degrees of similarity via weighted edges, the remaining challenge is to define what exactly makes two value sets *similar*. We recognize that value set similarity is context-specific and in many cases subjective. For our purposes, we define value set similarity as a measure of shared intension. We use two approaches to extract the intension of a value set: analysis of the name and code set. These methods, informed by the work of Winnenburg and Bodenreider\(^{11}\), are detailed below.

- **Value Set Name.** *Term Identification* is a Natural Language Processing (NLP) technique for extracting concepts from free text and mapping them to controlled vocabularies\(^{46}\). MetaMap\(^{47}\), an NLP tool developed by the National Library of Medicine (NLM), aims to assign Unified Medical Language System (UMLS)\(^{48}\) Concept Unique Identifiers (CUIs) to biomedical free text. Passing in the value set name to the MetaMap tool yields a set of CUIs representing the normalized semantics of the name.

- **Code Set.** A value set’s code set scopes the semantics of the value set by enumerating the permissible values for its domain. Using the code set as a representation of intension has some distinct advantages: (1) it requires
As a result, we represent value set intension as either a set of UMLS CUIs extracted from its name, or its member code set – in either case, a set of discrete values. The similarity of two value sets may then be computed as a measure of overlapping intension. Given two value sets, let $A$ and $B$ denote the derived UMLS CUI set or code set from each. Similarity is then defined as the Jaccard index of these two sets, or their intersection size divided by their union size:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

(4)

The higher the Jaccard index, the more similar we considered two value sets. Conversely, a low Jaccard index indicates dissimilarity, and similarity scores lower than a threshold of 0.25 were not considered in our cluster analysis.

The rationale for using two different similarity strategies is that we cannot necessarily rely on a single measure in isolation. For example, VSAC value sets BMI Values (2.16.840.1.113883.3.600.1.889) and BMI values (2.16.840.1.113883.3.600.1.888) have names that are semantically identical, but share no common codes, as they draw from two different code systems (ICD10CM and ICD9CM, respectively). Conversely, Procedures as Reasons for Admission to ICU Due to Pneumonia (2.16.840.1.113762.1.4.1111.26) and Need for Ventilator (2.16.840.1.113762.1.4.1045.82) share the exact same code set while having semantically dissimilar names. As such, we recognize that using different similarity measures will establish potentially different similarity relationships between value sets.

### Cluster Similarity

If the two different value set similarity measures relate value sets in different ways, we also expect them to produce different clusters of value sets. It is useful then to introduce a method to calculate just how dissimilar these clusters are – or cluster similarity. To determine this, we first compute two value set graphs, one with edges weighted by the code set similarity function and the other using value set name similarity. Next, we extract clusters from both graphs pairwise by a common focus node $i$, yielding clusters $C_i$ and $C_i'$. Finally, we compute similarity at a cluster level by inspecting how many nodes (or value sets) the clusters have in common by calculating the Jaccard index:

$$\text{cluster\_similarity}(i) = J(C_i, C_i')$$

(5)

### RESULTS

We analyzed 3820 total value sets from the VSAC repository, focusing on extracting implicit value set to value set relationships. Our methods sought to mine the overarching structures implied by these relationships – specifically hierarchies and clusterings of similar value sets. The metrics gathered were focused on quantifying the characteristics of these structures in an effort to better understand how they may be ultimately leveraged.

Analysis of the value set clustering characteristics gave insight into how interconnected the value sets were – or, how readily they formed into related groups. A cluster was defined as three or more connected value sets, and two sets of clusters were built using the two different similarity measures. Of the 3820 total value sets analyzed, 3185 clusters were found using the name similarity measurement, while 1190 were found via code set similarity. Table 1 reflects the results of the clustering coefficient calculation, a measure of how densely interconnected the clusters were. Valid clustering coefficients range from 0 – 1, with 0 indicating no neighbor connections and 1 being all neighbors interconnected (or a clique). This analysis was computed twice: once for each similarity algorithm, and the subtables reflect the result of both computations.

The clustering coefficient result for each cluster was also compared to a clustering coefficient for a similar, randomly assembled cluster. Through this calculation we were able to obtain the number of dense clusters, or clusters observed to have a weighted clustering coefficient greater than would be expected to occur randomly. For clusters grouped by value set name, 3170 out of 3185 clusters were considered dense. For code set groupings, it was 964 out of 1190.
Table 1: Clustering summary statistics. The clustering coefficient is a measure of how densely value sets tended to group together based on the two similarity measures.

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Median</th>
<th>Min</th>
<th>Max</th>
<th>Std. Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>$CC_i$</td>
<td>0.4067</td>
<td>0.4257</td>
<td>0.0</td>
<td>1.0</td>
<td>0.2834</td>
</tr>
<tr>
<td>$</td>
<td>C_i</td>
<td>$</td>
<td>4.9067</td>
<td>4.0</td>
<td>3.0</td>
</tr>
</tbody>
</table>

(a) Code Set Similarity

$\bar{CC}_i$: weighted local clustering coefficient

$|C_i|$: number of nodes in a cluster

(b) Name Similarity

$CC_i$: weighted local clustering coefficient

$|C_i|$: number of nodes in a cluster

Relationships between value sets indicated a degree of similarity. Our methods assumed that differences in how we computed value set similarity would result in different relationships, and consequently, different clusters. Table 2 shows in summary how similar the clusters from the two similarity measures were. Cluster similarity was computed as the Jaccard index of a cluster from each similarity measure, compared pairwise by a common focus node. For this measurement, the permissible cluster similarity range was $0 - 1$, with 0 indicating no shared value sets in the clusters (which is impossible, as via pairwise comparison by common focus the clusters will at least share one node), and 1 indicating that the clusters were identical.

Table 2: Similarity comparison of clusters computed from the two different similarity measures. Cluster similarity was calculated by pairwise comparison of clusters produced by both the value set name and code set similarity measures.

<table>
<thead>
<tr>
<th>Cluster Similarity</th>
<th>Mean</th>
<th>Median</th>
<th>Min</th>
<th>Max</th>
<th>Std. Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.2916</td>
<td>0.2143</td>
<td>0.0111</td>
<td>1.0</td>
<td>0.243</td>
</tr>
</tbody>
</table>

Analysis of these relationships was also leveraged to explore implicit hierarchical structures within the repository. Of the 3820 value sets analyzed, 1546 were found to be roots, or value sets with no computed hierarchical parents. To measure the extent of the extracted hierarchy, the longest path to a leaf for each root value set was calculated. Figure 3 summarizes our findings, reflecting the levels of hierarchies found in the repository, from 0 levels (meaning a value set with no children) to 5, the maximum level observed.

Figure 3: Analysis of extracted hierarchy levels. For each root value set (or value set with no computed parents), its height (or longest path to a leaf) was calculated, indicating the level of hierarchy. This figure relates these hierarchy levels to the number of value set roots in which they were observed.

DISCUSSION

We have described methods to extract two different organization structures from VSAC value sets, and we begin our discussion with an examination the structures found. Table 1 demonstrates high clustering coefficient values, indicating that value set clusters tend to be highly interconnected. We also find that the majority of clusters were dense, or exhibited a clustering coefficient value higher than would be expected for a similar random graph.
clustering is especially pronounced in clusters built using the value set name similarity measure, where almost all clusters (99.53%) were dense. This shows that not only are we able to extract clusters of similar value sets from a repository, but that value sets, at least in the VSAC repository, tend to be very closely clustered.

Our ability to infer connections between value sets is predicated on our ability to measure how connected, or similar they are. Utilizing two different similarity measures was assumed to produce different similarity connections, and thus, different clusters of related value sets. Table 2 shows that on average clusters computed using the two different similarity measures contained roughly 29% similar value sets. This indicates that while different similarity measures do indeed produce measurably different clusters, the clusters do show some degree of congruence. As value set similarity itself is inherently subjective and context dependent, we may be able to find underlying trends by looking at where different similarity measures agree.

A hierarchical structure was also successfully extracted from the repository. Figure 3 demonstrates that while the extracted hierarchy is relatively flat on average, some hierarchical structure does exist, and at times, can be as deep as 5 levels (see Figure 4). This is certainly an improvement organizationally over a flat list, especially since it required no manual curation or classification.

Our study purpose challenged us not only to extract structure, but to apply it in useful ways. As such, enhancing tooling with improved affordances for knowledge discovery is our end goal. In our previous work, we outlined the architecture for a value set management tool focused on usability\(^8\). We now leverage this toolset as our primary implementation platform. For hierarchy, Sunburst charts, or visual representations of hierarchy with radiating levels, have shown to be effective representations of hierarchical structures\(^5\). Figure 4 is an implementation of a Sunburst chart displaying an extracted VSAC value set hierarchy.

---

**Figure 4:** A Sunburst chart representing a value set hierarchy. Levels of hierarchy are represented as concentric circles radiating outward from a focus. In this example, the focus value set Mental Disorders is surrounded by radiating levels of hierarchy (shown left). On mouse-over, details of the value set Major Depressive Disorder New or Recurrent are displayed (shown right). The hierarchical path between the two is: Mental Disorders $\rightarrow$ Mental Disorders ICD10CM $\rightarrow$ Mental Health Diagnoses $\rightarrow$ BH Condition involving unipolar depression ICD-10-CM $\rightarrow$ Major Depression $\rightarrow$ Major Depressive Disorder New or Recurrent.

Clustering can be leveraged to suggest alternatives or possibly related value sets to users as they browse. Figure 5 depicts search results from a keyword search. Results found by traditional text-based searching are augmented by using See Also suggestions, where other members of the cluster are displayed as possibilities for further exploration.
In this study we have shown that data mining techniques can be employed to extract implicit relationships between value sets. These relationships in turn can be leveraged to add meaningful structure and organization to a repository. Ultimately, we show that these structures have practical applications in user-facing tooling, enhancing users’ ability to discover the correct value set for their use case, and thus, increasing value set reuse and general repository utility.

Limitations and Future Work

A limitation of our analysis is that hierarchy calculated by set containment algorithms may not always reflect the actual semantic hierarchies. There are certainly instances where a value set may contain a strict subset of codes from another but may not be a logical child, depending on context. Figure 6 illustrates one such example. In this case, the codes of Medication Fill Status are a strict subset of the codes of ProcedureActStatusCode and ProblemActStatusCode. It may be illogical (or at least context-dependent) to state that these value sets share a semantic relationship.

Future directions may include further integration of ISO/IEC 11179. Specifically, we look to consider data elements as a measure of value set similarity, such as grouping value sets by the data they describe (for example, grouping by object class or property). We also expect future work to focus on applying this process to value set repositories other than the VSAC to examine if hierarchies and clusters cross repository boundaries.

Acknowledgment. This study was supported in part by the LHSNet project (CDRN-1501-26638), the caCDE-QA project (U01 CA180940), and the BD2KonFHIR project (U01 HG009450). We also acknowledge Mayo Clinic’s Knowledge Management and Delivery effort, as well as the Rochester Epidemiology Project.
References

A Conceptual Measurement Model for eHealth Readiness: a Team Based Perspective

James Phillips, BIT (Hons) BComm¹, Simon K. Poon, PhD¹, Dan Yu, BIT (Hons) BComm¹, Mary Lam, PhD², Monique Hines, PhD¹, Melissa Brunner, MHlthSc¹,³, Emma Power, PhD¹, Melanie Keep, PhD¹, Tim Shaw, PhD¹, Leanne Togher, PhD¹

¹University of Sydney, Sydney, NSW, Australia; ²University of Technology Sydney, Sydney, NSW, Australia; ³University of Newcastle, Callaghan, NSW, Australia

Abstract

Despite the shift towards collaborative healthcare and the increase in the use of eHealth technologies, there does not currently exist a model for the measurement of eHealth readiness in interdisciplinary healthcare teams. This research aims to address this gap in the literature through the development of a three phase methodology incorporating qualitative and quantitative methods. We propose a conceptual measurement model consisting of operationalized themes affecting readiness across four factors: (i) Organizational Capabilities, (ii) Team Capabilities, (iii) Patient Capabilities, and (iv) Technology Capabilities. The creation of this model will allow for the measurement of the readiness of interdisciplinary healthcare teams to use eHealth technologies to improve patient outcomes.

Introduction

The increase in the use of information and communication technology (ICT) in the delivery and administration of healthcare services has long been implicated to varying degrees as the cause of long term increases in health care expenditure [1–3]. Health technology more broadly however has and continues to have transformative effects on the delivery and administration of healthcare services. These transformations have significant implications for the efficacy and efficiency of the delivery of healthcare services to patients with the aim of improving health outcomes. Continued expenditure on health technology is inevitable. However, the total return on this investment could be improved through the existence of greater efficiency and effectiveness in use of health technologies by clinicians and patients. As such, need emerges for the determination of the level of fit that exists between a healthcare environment, its participants, and its technology.

The focus of this paper is to provide insight into factors that need to be considered for measuring eHealth readiness for healthcare teams, and in the construction of a conceptual model for the measurement of readiness in interdisciplinary healthcare teams. While this research aims to achieve an overarching understanding of the factors, there is a particular focus on eHealth and its application to healthcare within the domain of traumatic brain injury (TBI) rehabilitation. We developed a three phase methodology consisting of a qualitative study and literature review to identify relevant themes, construct and model formation, and model validation. From this methodology emerged a conceptual model for the measurement of the readiness of interdisciplinary healthcare teams to use eHealth technologies in the course of patient care.

Identification of readiness factors

eHealth technology has the capability and potential to transform the delivery of health services to patients everywhere. As it currently stands, there exists no model for the measurement of the ability for interdisciplinary healthcare teams to use eHealth in the context of the whole healthcare environment to improve patient outcomes.

There is a broad base of literature covering specific areas of eHealth such as the effect of public engagement [4], factors affecting clinician acceptance of technology [5], factors affecting team performance [6,7], organizational aspects of change and readiness [8,9] and factors affecting ongoing intervention participation[10].

Physician acceptance and decision to adopt eHealth technology has been identified as significant in explaining physician responses to technology [5,11,12]. As such, the level of technological acceptance by physicians and their decisions to adopt eHealth technology in their practice for their clients will form part of the assessment of the level of eHealth potential. In a study of 408 Hong Kong physicians, seven factors were identified as having significance in explaining physician technology acceptance: attitude, subjective norms, perceived behavioral control, perceived usefulness, perceived ease of use, and behavioral intention [5]. In a study of 519 responses by physicians practicing in the province of Quebec, significant factors influencing perceived responsibility were: perceived consequences,
personal normative belief, and self-identify [11]. Additionally, modelling of 114 Taiwan-based clinicians acceptance of eHealth found three factors with high explanatory fit; technology support and training, compatibility, and intention to use [12]. Factors concerned with physician acceptance of eHealth technology are necessities for diffusion into clinical practice [11].

Where eHealth technology is used by the patient in a setting such as their home, the level of acceptance in its use is likewise a necessity for improved clinical outcomes. Where the patient rejects or is unable to use the technology, the clinical outcome improvement is unable to be achieved. Investigation of such factors governing the engagement of patients with eHealth technology were determined as being: characteristics of users, technological functionality and issues, characteristics of eHealth services, social aspects of use, and eHealth services in use [4,10,13].

Factors have also been identified relating to the organization itself that affect employee engagement and acceptance of change. Culture and incentivization have been identified as contributing to employees having greater acceptance of change [14]. Management support and resource availability are also identified as being factors in organizational adoption of health technologies [12,15,16].

eHealth Literacy Toolkits

There are a number of frameworks which provide sets of measurement items and scales for constructs that have been established [17–20] to assess a participant's level of eHealth literacy. Norman and Skinner developed an eight item measure to ascertain the eHealth literacy of a participant (eHEALS) [17] with items concerned with patient competency with accessing health information found on the internet. The eight items are based on the six factor lily model [18] which establish the factors predicting eHealth literacy as: health literacy; traditional literacy and numeracy; computer literacy; information literacy; science literacy; and media literacy. Additionally there is the eHealth Literacy Assessment Toolkit (eHLA) [20] which similarly attempts to measure eHealth literacy and draws from the Norman and Skinner lily model and from the seven domains model [21].

The literature assessed is from a broad and fragmented body of work that has developed over the past five decades. The identification of relevant factors affecting the acceptance or the use of eHealth technologies is across three clusters: clinician, patient, and organization. Thus, these identified factors and clusters provide the basis for the formation of the conceptual model of eHealth readiness of interdisciplinary healthcare team.

Methodology

The goals of this research were to: (i) gain insight and an in-depth understanding of clinician perspectives of using eHealth technologies as an interdisciplinary team, (ii) develop a measurement model based on those experiences, and (iii) validate the model and propose a measure of readiness of an interdisciplinary healthcare team to use eHealth. The methodology combines qualitative and quantitative methods over three phases to devise a measurement. These phases are outlined below.

Phase 1 – Participant recruitment

Traumatic Brain Injury (TBI) was chosen as the domain due to the interdisciplinary nature of the work and the potential benefits that eHealth can bring to such work [22]. Thus, two focus groups and an interview were conducted to gain an understanding of the attitudes surrounding eHealth and interdisciplinary work within rehabilitation for people after TBI, with a third focus group used to confirm the content analyses of the two prior focus groups and interview [23]. Participants consisted of healthcare professionals working in TBI rehabilitation with two teams based in regional New South Wales (Australia), and the interview participant and the third team based in metropolitan New South Wales (Australia). Recruitment for all four studies was based upon the purposeful sampling technique [24].

Phase 1 – Focus group data collection

Four key questions were used to facilitate discussion between participants based on previously established principles [25]. The questions presented to participants were open-ended and aimed to elicit discussion [26] that uncovered the participants’ experiences with eHealth, the challenges they faced, and their attitudes towards the use of eHealth in
interdisciplinary healthcare. Additional sub-questions and visual prompts were provided as required to ensure that the discussion among the participants was relevant to the research questions.

Four researchers assisted with facilitating the focus groups and collecting data: the primary moderator, assistant moderator, and two assistants. All four researchers took notes on the discussion, and the discussion was digitally recorded and later transcribed verbatim. The transcripts were de-identified to ensure anonymity of participants and organizations. The transcripts were checked for accuracy against the digital recording by the researchers. The final transcripts and a summary of key points were then emailed to participants for verification (member checking). No request for alteration to the transcript or summary was made by any participant.

<table>
<thead>
<tr>
<th>Factor</th>
<th>Focus group 1 [27]</th>
<th>Focus group 2</th>
<th>Individual interview</th>
<th>Focus group 3</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Region</strong></td>
<td>Regional New South Wales (Australia)</td>
<td>Regional New South Wales (Australia)</td>
<td>Metropolitan New South Wales (Australia)</td>
<td>Metropolitan New South Wales (Australia)</td>
</tr>
<tr>
<td><strong>Organization type</strong></td>
<td>Non-government organization</td>
<td>Government and community</td>
<td>Private and community</td>
<td>Government and community</td>
</tr>
<tr>
<td><strong>Services provided</strong></td>
<td>Public inpatient and outpatient rehabilitation</td>
<td>Outpatient rehabilitation services</td>
<td>Acute neurological acute care and rehabilitation</td>
<td>Public inpatient and outpatient rehabilitation</td>
</tr>
</tbody>
</table>
| **Disciplines of the health care professionals** | • Speech pathology  
• Occupational therapy  
• Medical  
• Nursing  
• Care coordinator | • Speech pathology  
• Occupational therapy  
• Care coordinator  
• Social work  
• Administration | • Speech Pathology  
• Speech pathology  
• Social work  
• Case manager |
| **Years working in the team** | • One over 10 years  
• One 5-10 years  
• Three 1-5 years | • Three over 10 years  
• Two 1-5 years  
• One less than 1 year | • Over 10 years | • Two 5-10 years  
• Two 1-5 years  
• One less than 1 year |

Table 1 Summary of participant demographics

Phase 1 – Focus group data analysis
A traditional approach was undertaken for qualitative analysis of the transcripts and is based upon the simple analysis framework by Krueger and Casey [28]. This involved the three main steps of: categorization of raw data, descriptive statements, and interpretation of data. As part of the first step, two researchers independently analyzed and categorized data into categories. The categories of raw data were then cross-analyzed and discussed by a group of researchers to ensure consensus and reliability in the development of descriptive statements for each theme.

Once consensus was achieved on the themes and categories of the data, these descriptive statements were interpreted. As the analysis of focus group data involves a level of subjectivity, care was taken during the analysis to ensure that bias was not introduced in the interpretation of the results. We took into consideration any researcher preconceptions, the specific words spoken by the participants and the context of their statements, and the internal consistency and specificity of statements. To reduce subjectivity, findings were strengthened by having two researchers analyze the data independently, with the results and main findings then given to an independent reviewer for validation.

Phase 1 – Themes from the literature
The primary purpose of the literature review was to identify additional factors which infer the readiness of interdisciplinary healthcare teams to use eHealth technologies in the treatment of patients to improve health outcomes to achieve the development of a model with greater content validity [29].

The themes that were identified within the literature were extracted through a process of interpreting the literary work and summarizing the themes identified within each piece of work. Where such themes were recurring across works in the literature or aligned with the pattern of themes within the focus groups, these themes were flagged for inclusion into the factors that would be used in the initial conceptual model.
Phase 2 – Model factor formation

The Resource-Based View (RBV) [30] was used to provide a framework through which aspects of the organization could be described in the form of a resource giving rise to capabilities where such resources contribute to the performance of the healthcare team specifically and the organization more broadly [31]. Each resource was considered to be a tangible or intangible asset in which the organization has invested in financially, materially, or temporally, and from which the healthcare team can reasonably expect to gain some benefit.

Using this definition of a resource, the factors that were identified within the literature review and the themes that emerged from the content and thematic analyses of the focus groups were transformed into discrete resources. Each conceptual resource was recorded as one or more written sentences describing the principle entity within the resource and the nature of the factor or theme. Hypothetical examples were included with the construction of the given resource if it was thought that such inclusion would clarify the meaning of the resource for the user of the model. In following this process, a link can be made between each focus group theme or literature factor and the set of resources that will make up the model.

Phase 2 – Model structure formation

A pre-existing model structure was used in the initial formation of the conceptual model with the HOT-fit model being used [32] owing to the model’s emphasis on the relationship and fit between the human, organizational, and technological factors in the environment [33].

The establishment of clusters of resources through the construction of three categories was performed with the categories being defined as: (i) User factors (analogous to the human factors) — which contained all resources where the healthcare team or the patient is the principle aspect of the resource, (ii) Organizational and external factors — which contained all resources relating principally to either the organization or factors external to the organization, and (iii) Technology factors — which contained resources principally related to the eHealth technology itself.

Due to the planned use of the Delphi methodology [add ref] with regards to the refinement and finalization of the conceptual model, the risk of information overload [34] was controlled through the use of sub-clusters to reduce the number of resources within each category of the model structure. Domains were established as categories of similar thematic content within the same cluster.

Phase 3 – Model content validation

The Delphi method [35] was used as the method to refine both the content and the structure of the model and to establish the content validity of the model [29] over a series of 3–5 iterations [36] allowing experts to systematically consider the complex problem of readiness measurements [37]. The stopping condition of the method is that of consensus being reached which is a satisfactory proxy for the content validity index’s stated purpose as a measurement of consensus [38].

The recruitment of the experts for the Delphi method occurred through a snowball technique utilizing recommendations of an expert within the information technology field. Experts were selected based upon their expertise domain with the final make-up of the Delphi panel consisting of five experts; four of whom were from The University of Sydney Faculty of Health Sciences, and one from The University of Sydney Faculty of Engineering and Information Technologies. All five experts were present for each of the three rounds required to reach consensus and each provided feedback into the process. At the end of each iteration the feedback was incorporated into the model and the revised model presented for discussion at the next iteration.

Phase 3 – Model construct validation

Q methodology [39] in the form of several q sorts previously developed as part of prior research [40] was used to assess the construct validity of the model through the assessment of the correlation of multiple individuals’ interpretation of the model resources [41]. If there is an appropriate similarity, then the interpretation of the resources and the placement of those resources within the model by individuals will be consistent with the theoretical a priori model, whereby construct validity can be established.

An open card sort and a closed card sort activity [42] were designed to have participants first sort all 59 of the resources of the model into four categories defined and named by each participant. Subsequently, the same participants then sort the same resources into four categories for which the names of the categories were given — the four names being those of the cluster names in the conceptual model. The participants were not aware of the names of the categories for the first sorting task. The card sort activities were performed using an online system and an offline system. Further construct validation is currently being undertaken using an online survey instrument to collect data for statistical validation.
Results and Analysis

The focus groups involved participants from organizations that provide services to the community with the aim of helping their clients transition back to their normal lives. While the team in focus group 1 (FG1) provide rehabilitation services for adults, the second and third focus groups (FG2 and FG3 respectively) provide rehabilitation services for both children and adults. The backgrounds of the three focus groups are similar, however FG2 and FG3 are managed under NSW Health while FG1 is managed externally to this network.

The participant from the individual interview (IP) is the principle of a Sydney metropolitan private practice that specializes in speech pathology services for adults with neurological impairments. Due to the nature of the industry, the participant operates as part of various interdisciplinary teams external to the practice on an ad-hoc basis.

Six main themes emerged from data analysis which are outlined in Table 2.

<table>
<thead>
<tr>
<th>Theme</th>
<th>Overview</th>
</tr>
</thead>
<tbody>
<tr>
<td>Organizational structure</td>
<td>• Positive attitude towards co-location for information sharing and case management.</td>
</tr>
<tr>
<td>Culture and attitudes towards technology</td>
<td>• Non-technology oriented clinicians.</td>
</tr>
<tr>
<td></td>
<td>• FG1, FG3, and IP have positive attitudes towards eHealth.</td>
</tr>
<tr>
<td></td>
<td>• FG2 less positive towards to impact and use of eHealth.</td>
</tr>
<tr>
<td>External organizations</td>
<td>• Lack of access to electronic records held by external organizations.</td>
</tr>
<tr>
<td></td>
<td>• Effective resource location primarily reliant on networks developed by clinicians.</td>
</tr>
<tr>
<td></td>
<td>• Limited information sharing with external organizations.</td>
</tr>
<tr>
<td>IT support</td>
<td>• Level of support variable across groups.</td>
</tr>
<tr>
<td></td>
<td>• No co-located IT support for regional FGs.</td>
</tr>
<tr>
<td></td>
<td>• Difficult for regional FGs to obtain timely support.</td>
</tr>
<tr>
<td>Technology, facilities, and infrastructure</td>
<td>• Wide use of different technologies e.g. tablets, smart phones, email, in-house administrative software.</td>
</tr>
<tr>
<td></td>
<td>• Difficulties with data storage limitations on hospital servers for FG2.</td>
</tr>
<tr>
<td></td>
<td>• Lack of support in most electronic medical records (EMRs) for non-textual data.</td>
</tr>
<tr>
<td></td>
<td>• Lack of available IT facilities for regional FGs e.g. video conferencing equipment.</td>
</tr>
<tr>
<td></td>
<td>• Variable internet quality for regional FGs.</td>
</tr>
<tr>
<td></td>
<td>• Barriers such as cost, training, and client attitude exist with respect to client use of eHealth.</td>
</tr>
<tr>
<td>Policies and Guidelines</td>
<td>• Tight security model for FG1, FG2, and FG3. Some individuals unable to access emails, install apps, or visit certain websites such as YouTube.</td>
</tr>
<tr>
<td></td>
<td>• Restrictions on text messaging and emailing photos.</td>
</tr>
<tr>
<td></td>
<td>• FG1 and FG2 employees feel they are not trusted by the organizations and not empowered to use eHealth.</td>
</tr>
<tr>
<td></td>
<td>• Cumbersome information sharing policies results in non-compliance and ignorance of policies in FG1, FG2, and FG3.</td>
</tr>
</tbody>
</table>

Table 2 Emergent Themes from Qualitative Study

Resource and structure formation

The operationalization of the themes from the qualitative study and the factors identified in the literature review resulted in the creation of an initial set of 44 distinct resources. These resources were categorized into three clusters: (i) the Organizational Capabilities cluster, (ii) the User Capabilities cluster, and (iii) the Technology Assets cluster. These categories were drawn from the HOT-fit model which was used as the initial foundational model for the structure formation. This structure and the contained resources were presented as the initial model for the consideration of the Delphi panel.

Delphi method implementation

The Delphi method was carried out over three rounds to determine the resources which would be used as a measure for each of the clusters of the model. At the completion of the third round consensus was given and it was deemed that the stopping condition, of reaching consensus, was achieved. The final model consisted of four clusters and 59 resources with 15 resources being included based on the feedback and agreement of the Delphi panel. Due to the consensus of the expert panel it was determined that the model possessed sufficient content validity.
Conceptual model construct validation

The q sort was undertaken to establish the construct validity of the model [29] where for a model with construct validity it would be expected that there would be convergence on the categorization of resources. The closed card sort was attempted by 12 participants and was successfully completed by 10 participants — that is they sorted all resources in the activity in a category. The participants of the sort were students from a health technology innovation class with individuals from clinical, health administration, and engineering backgrounds. The results of the sort were analyzed for convergence to identify which resources needing refinement. The evidence from the closed sorting task is that there does appear to be construct validity in at least three of the four clusters.

From the closed sort, there were six items of concern where less than 65% of the participants converged on the placement of the resource into a category — all six items were from the Technology Cluster. Additionally, there were two items — both from the Technology Capabilities cluster — which were categorized (with 70–80% convergence for both items) into a different cluster than that of the conceptual model. Where there is a miscategorization of a resource it must be determined whether the resource composition was the cause of the miscategorization or whether the participants interpreted the theme underlying the resource to be associated with a cluster different to that determined in the content validation phase.

The open card sort was attempted by 12 participants in total with 9 participants completing the activity. Overall 28 categories were created by 9 participants which indicated that some participants created greater or fewer than the four categories they were instructed to create.

The results of the open card sort were sorted into a hierarchical clustering using Ward's method [43] to determine the four clusters with the minimum within-cluster variance. The following themes emerged from the clustering: Organization, Users, electronic medical records (EMR), and Other Technology. However, the inter-cluster distance between EMR and Other Technology is minimal and it would be reasonable to merge them if the number of clusters was unrestricted. The emergent clusters while not significantly dissimilar to the proposed model do not reflect the expected clustering in some cases.

Discussion

The results from the qualitative study confirmed the issues raised in existing literature, including the need for rigorous evaluation of eHealth and the importance of considering the technology-environment fit. The findings also moved beyond the literature and present a conceptual model to measure the eHealth readiness — that is the technology-environment fit — of interdisciplinary healthcare teams. One of the most interesting findings from the study is that although the clinicians faced a number of challenges with regards to the use of eHealth and were unable to achieve the full benefits that it can offer, they presented positive attitudes towards eHealth. The majority of issues highlighted by the study are touched upon in the literature, however the focus is still largely on the technological aspects of evaluation such as functionality and the quality of the system, rather than having an equal focus on the external and environmental impacts on technology. Without understanding the inclusion of organizational factors such as procedures, policies, infrastructure, or even if the technology is effective, it is not likely to be successfully implemented. Thus, factors from both health technology and health informatics evaluation should be drawn to provide the foundation for a rigorous eHealth evaluation model that considers both the effectiveness of technology and the fit of technology within its environment.

Proposal of a health technology evaluation model

This study highlights the gaps within literature regarding the evaluation of eHealth within organizations. In particular, there is no evaluation model that considers the environment of health technology implementation within the context of a healthcare team and which includes factors such as the team or the organizations. The effect of this gap can be seen in the study, where eHealth was implemented with the belief that it can improve processes and client care, however these desired results were not observed. The results indicate that there is a lack of emphasis on particular elements in the evaluation of eHealth in literature such as infrastructure and policy. In the literature, many of these issues are often overlooked as being of lesser importance to factors such as technological and user impacts.

We propose a conceptual model that aims to integrate the two evaluation processes and bridge the gap between the evaluation of the health technology itself and the evaluation of the fit between the technology, its environment, and its users. These two processes need to be integrated to determine whether the technology is effective and safe to use, as well as consider the type of organization or individual that the technology is suited for. The proposed model consists of four clusters each containing domains which give rise to capabilities found to allow healthcare teams to implement and effectively use eHealth technologies for the improvement of patient outcomes. Current literature indicates the
challenges faced in the evaluation of non-traditional health technologies such as eHealth, however there is no coherent solution to bridging the gaps between the existing models. Thus, the proposed framework attempts to integrate health technology evaluation and health informatics evaluation to provide a means for the holistic evaluation of eHealth readiness in healthcare teams. The conceptual model (outlined in Error! Reference source not found.) consists of four clusters: (i) External Factors, (ii) Team Capabilities, (iii) Patient Capabilities, and (iv) Technology Capabilities.

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>External Factors</td>
<td>- Policies Asset</td>
</tr>
<tr>
<td></td>
<td>- Procedures Asset</td>
</tr>
<tr>
<td></td>
<td>- Implementation Asset</td>
</tr>
<tr>
<td></td>
<td>- Cultural Asset</td>
</tr>
<tr>
<td>Team Capabilities</td>
<td>- Team Training Asset</td>
</tr>
<tr>
<td></td>
<td>- Technology Use Asset</td>
</tr>
<tr>
<td></td>
<td>- Leadership Asset</td>
</tr>
<tr>
<td></td>
<td>- Communication Asset</td>
</tr>
<tr>
<td>Patient Capabilities</td>
<td>- Patient Value Asset</td>
</tr>
<tr>
<td></td>
<td>- Patient Ability Asset</td>
</tr>
<tr>
<td>Technology Capabilities</td>
<td>- EMR Asset</td>
</tr>
<tr>
<td></td>
<td>- Technical Asset</td>
</tr>
</tbody>
</table>

**Table 3** Conceptual Measurement Model Clusters and Domains

**External Factors**
The External Factors cluster (Table 4) comprises the set of domains that the clinical team cannot reasonably expect to be able to materially influence, change, or mitigate through intentional action.

<table>
<thead>
<tr>
<th>Domain</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Policies Asset</td>
<td>Quality of documents and resources outlining the vision and strategy of the organization concerning eHealth.</td>
</tr>
<tr>
<td>Procedures Asset</td>
<td>Quality of documents and resources providing concrete guidelines on the use of eHealth by healthcare teams.</td>
</tr>
<tr>
<td>Implementation Asset</td>
<td>Ability of the organization to carry out the operationalization of policy and procedure.</td>
</tr>
<tr>
<td>Cultural Asset</td>
<td>Ability of the organization to affect change in behavior of the healthcare team with respect to eHealth use.</td>
</tr>
</tbody>
</table>

**Table 4** External Factors: Domains and Descriptions

**Team Capabilities**
The Team Capabilities cluster (Table 5) refers to the readiness of a team working in a collaborative manner to deliver healthcare services to patients using eHealth technologies.

<table>
<thead>
<tr>
<th>Domain</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Team Training Asset</td>
<td>Degree of training that a healthcare team has undergone regarding the use of eHealth, its integration into practice, and the effective operation as an interdisciplinary team.</td>
</tr>
<tr>
<td>Technology Use Asset</td>
<td>Measure of the use of eHealth by the healthcare team, the level of technology literacy of the healthcare team, and the extent of difficulties in using eHealth.</td>
</tr>
<tr>
<td>Leadership Asset</td>
<td>Activeness of the healthcare team’s diffusion of eHealth within the organization.</td>
</tr>
<tr>
<td>Communication Asset</td>
<td>Ability of the healthcare team to communicate ideas and concerns effectively and efficiently with external entities and within the team.</td>
</tr>
</tbody>
</table>

**Table 5** Team Capabilities: Domains and Descriptions

**Patient Capabilities**
The Patient Capabilities cluster (Table 6) measures the perceived and measured capabilities of the collective patients treated by the healthcare team to utilize eHealth practices and technologies in improving clinical outcomes. If the patients receiving treatment from the healthcare team are on average unable to use eHealth technologies then the team is as a function of this unable to use eHealth technologies in the patient care and as such the measured readiness of the team to use eHealth technologies will be low or non-existent.
### Technology Capabilities

The Technology Capabilities cluster (Table 7) seeks to provide a measure of the sufficiency of eHealth technologies, the support provided for those technologies, and the physical access afforded to those technologies in allowing for clinical teams to improve the health outcomes of patients. It is anticipated that a high measure in this cluster would correlate with increased efficiency of administrative tasks, increased availability of patient information, the timely provision of information technology support, or the provision of appropriate devices and internet infrastructure to allow for clinical teams to access patient information and other health information when required.

<table>
<thead>
<tr>
<th>Domain</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>EMR Asset</td>
<td>The level of integration of the EMR system used by the clinical team has with other systems in the organization such as in the automatic sharing of data and information between systems.</td>
</tr>
<tr>
<td>Technical Asset</td>
<td>The ability of the healthcare team and patients to access and obtain support for eHealth technologies used in the care of patients.</td>
</tr>
</tbody>
</table>

### Limitations

There are a number of inherent methodology limitations concerned with the qualitative study, the Delphi panel, and the q sort used. The content and thematic analysis conducted on the data from the qualitative study was based on limited non-saturated data which was subsequently supplemented by literature. Owing to this non-saturation it may be that the content validity of the model is limited. The experts used in the Delphi panel were drawn from a pool of individuals from the same research group. This limitation may cause the content validity of the model to be weaker than anticipated. The population used for the q sort had several limitations including that they may not have been representative of the wider healthcare population, that none of the participants were regionally based, and that the population used was small. Additionally, a varying number of categories were created per participant. This will affect the hierarchical cluster analysis as the distances between resources may result in the cluster formation not being representative of the participant’s intent. Further construct validation is as such being currently carried using an online survey instrument being disseminated to healthcare professionals.

### Conclusion

Health technologies and eHealth specifically have the potential to transform the delivery of healthcare services to patients for the purposes of improving health outcomes and quality of life for those patients. Such use of technologies however, is expensive and sustainable expenditure must be established within the framework of existing healthcare systems. Additionally, the value of eHealth technologies used in the delivery of healthcare services is limited by the skills, knowledge, and support available to both healthcare teams and patients — where an ineffective environment for the engagement with such technologies minimizes any beneficial outcomes that may be obtained from the use of the technologies.

The methodology framework has led to the proposal of the four-factor model with content validation and initial construct validation being found. Further work is currently being undertaken to provide construct validity of the measure using an online survey instrument. We believe that this conceptual model provides a foundation for the aggregation of multi-discipline constructs for the purposes of the creation of a measurement model. This model specifically targets interdisciplinary healthcare teams and their readiness to use eHealth and provides at the least a validated set of resources which should be considered for inclusion into future quantitative models.
References


Challenges with Collecting Smoking Status in Electronic Health Records

Fernanda Polubriaginof MD, MA1, Hojjat Salmasian MD, MPH, PhD2,
David A. Albert DDS, MPH1, David K. Vawdrey PhD2,1
1Columbia University, New York, NY; 2NewYork-Presbyterian Hospital, New York, NY

Abstract

Smoking is the leading cause of preventable death in the United States. Obtaining patients’ smoking status is the first step in delivering smoking cessation counseling. In this study, we assessed the quality of smoking status captured in an electronic health record from a large academic medical center. We analyzed data from structured notes, finding that smoking status was documented in 98% of 64,451 hospital encounters in 2016. 32% hospital encounters had discrepant documentation, and 54.5% of patients had implausible changes (e.g., changes from “current smoker” to “never smoker”). Overall, only 2.9% of patients were documented as active smokers, but 36.4% were documented as “unknown” or had discrepancies in their smoking status. These results suggest that patients that smoke are not appropriately identified. Centralized documentation with clinically actionable smoking status categories and implementation of patient-facing tools that allow patients to directly record their information could improve data quality of smoking status.

Introduction

Smoking is an important risk factor for multiple diseases, including cardiovascular diseases and numerous types of cancer. It remains the number one cause of preventable death in the United States.1 The collection of patients’ smoking status during clinical encounters is critical to providing patients with resources to quit smoking. Smoking cessation can be difficult, and clinical visits are opportunities to intervene and recommend smoking cessation programs and therapies. Obtaining a patient’s smoking status is a crucial step in beginning smoking cessation interventions and monitoring progress.2 It may seem that recording updates to smoking status in a timely and accurate manner would be straightforward using modern electronic health records. This may not be the case for several reasons, including lack of standard terminology and granularity for data collection, shifting cultural attitudes regarding tobacco use, and potentially frequent changes in individuals’ smoking behavior.3,4 As the American author, Mark Twain, famously quipped, “Giving up smoking is the easiest thing in the world. I know because I’ve done it thousands of times.”

Given the clinical importance of smoking status, the “Meaningful Use” financial incentive program for electronic health record (EHR) adoption in the U.S. included a requirement for healthcare providers to capture patients’ smoking status electronically in structured fashion.5 Meaningful Use has helped to standardize data collection of smoking status and other information. However, even with improved standards for representing information, data quality issues have persisted in many patient-provided data types, such as race and ethnicity6,7 and family history.8,9 Previous studies on data quality have shown that clinicians describe a need for free-text documentation for expressiveness of documentation; however, these affordances challenge data reuse.10

Appreciating the challenges associated with data quality and the balance between the expressiveness of free-text and the benefits of structured data, we set out to answer a very simple question: how many of our hospital’s patients are known to be active smokers? We undertook a study to analyze how smoking status is currently being collected in a large academic medical center and to evaluate the quality of this data in EHRs.

Methods

We conducted a retrospective analysis of smoking status related data from the Allscripts Sunrise EHR (Allscripts Corp., Chicago IL) used at NewYork-Presbyterian Hospital/Columbia University Medical Center. We included patients that had at least one hospitalization during 2016, specifically patients that were discharged between January 1, 2016 to December 31, 2016.
The EHR system contained thousands of active templates for documentation. Each of these templates contained one to several hundred discrete observations. An observation could be a text box, a Boolean (e.g., a checkbox or radio button), or numeric value. We identified observations in which the description contained the stemmed words “smok”, “cigar” or “tobacco” and queried the EHR database to identify the number of times each one of these observations have been recorded during the study period. The value can be either textual (e.g. when the parameter is “Cigarettes (packs per day)”) with the response typed into a free-text box, or it can be structured (e.g. selected from a picklist). The picklist is often shown as a set of radio buttons that users can pick one. For example, there is a parameter labeled “Tobacco Use / Smoking Status”, and one possible choice in the picklist is “Never smoker.”

Exploratory analysis showed that 94% of patients had at least one structured smoking status observation recorded in a structured field. Based on this finding, we chose to use only the structured data for the remainder of the analyses. For comparison purposes, we classified smoking status into one out of four possible categories: “Current smoker”, “Former smoker”, “Never smoker”, and “Unknown smoking status” as described in Table 1. Observations such as “n/a for age” or “as per pt” were excluded from the analysis. Differences in smoking status recording were classified into two distinct categories: plausible, when the change is feasible to happen, and implausible, when the conflict is not possible to happen or in cases where there is loss of information. The description of this classification is illustrated in Figure 1. We also calculated the time interval (in days) between smoking status documentations to better understand the distribution of the data during the one-year study period. For example, time interval of zero means that both observations were recorded in the same day, time difference of one means that the second observation was recorded one day after the first one was recorded.

Furthermore, we analyzed the discrepancies in smoking status reported for the same patient during the same visit, using data from different clinical notes in the EHR. We also investigated the discrepancies of smoking status recorded by different provider types (e.g., nurses, medical doctors, care coordinators, social workers). To analyze if provider types would have an impact in the number of discrepancies observed, we calculated the number of distinct provider roles recording smoking status for each admission. We then compared the number of distinct provider roles for patients with and without discrepancies in the smoking status recorded.

Overall, data quality of smoking status was assessed based on the percentage of patients with consistent and informative smoking status available (i.e., not classified as “Unknown” in the database, and not conflicting).

Results

Overall, we reviewed 48,909 patients having 64,451 hospital encounters in the one-year study period. We identified 203,048 observations of smoking status for 47,849 unique patients across 62,988 distinct hospital encounters. No smoking status documentation was identified for 1,463 visits from 1,060 distinct patients, representing 2% of the number of hospital encounters and 2% of the overall number of patients. In other words, 98% of patients and 98% of hospital visits had documentation regarding the patient’s smoking status. Of those records with smoking status, 59,663 visits (93%) from 45,822 patients (94%) had this information recorded in structured format. After pre-processing, 45,771 patients (94%), including 59,593 visits (92%) and 129,134 observations were classified into four distinct smoking status categories. The number of observations and the mapping to the simplified smoking categories are described in Table 1. The description of smoking status data during the one-year study period is described in Table 2. Patients had an average of 1.3 visits/patient during the study period, with the maximum number of visits a single patient being 23 visits.

Longitudinal One-year Review

Overall, 15,048 patients (32.9%) had smoking status recorded in a single note, and 30,723 patients (67.1%) had more than one note with documentation regarding smoking status. Among the patients with more than one note with smoking status documented, we identified 83,363 changes in documented smoking status collected longitudinally during the one-year study period.

Among the changes in smoking status documentation, 32,582 (39.1%) had a conflicting smoking status. These discrepancies were observed in records from 15,207 distinct patients, representing 33.2% of our study population. However, because we are working on longitudinal data and smoking status is not a static concept since it can change over time, some of these discrepancies are feasible to happen. For example, someone that never smoked can become a smoker. Others, however, are implausible. For example, a non-smoker cannot become a former smoker, or a current smoker cannot become a never smoker, unless some of the data was recorded incorrectly. Other changes are
plausible, however not good from a data quality standpoint. Having a patient with documentation regarding smoking status and later not having smoking status (smoking status as “unknown”) demonstrates loss of information. Implausible changes as well as changes from a well-defined smoking status to uninformative category were considered discrepancies due to data quality issues. We identified 17,757 discrepancies (implausible changes and loss of information changes), which constituted 54.5% of changes, in 10,836 distinct patients. These discrepancies are represented in Figure 1 as dashed lines, while the other changes are represented in continuous lines.

Table 1. Description of the mapping from smoking status categories as recorded in the EHR to the four clinically actionable categories. Smoking status categories documented in the EHR that utilize the standard criteria defined by the Meaningful Use program are highlighted in bold.

<table>
<thead>
<tr>
<th>Clinically Actionable Smoking Status Categories</th>
<th>EHR Documented Categories</th>
<th>Number of Observations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Never Smoker</td>
<td>Never Smoker</td>
<td>67,052</td>
</tr>
<tr>
<td></td>
<td>Smoker (No)</td>
<td>12,979</td>
</tr>
<tr>
<td></td>
<td>Patient Denies</td>
<td>560</td>
</tr>
<tr>
<td>Current Smoker</td>
<td>Current every day smoker</td>
<td>5,188</td>
</tr>
<tr>
<td></td>
<td>Current some day smoker</td>
<td>1,418</td>
</tr>
<tr>
<td></td>
<td>Light smoker</td>
<td>714</td>
</tr>
<tr>
<td></td>
<td>Heavy Smoker</td>
<td>267</td>
</tr>
<tr>
<td></td>
<td>Smoker, current status unknown</td>
<td>676</td>
</tr>
<tr>
<td></td>
<td>Smoker (Yes)</td>
<td>1,968</td>
</tr>
<tr>
<td>Former Smoker</td>
<td>Former smoker</td>
<td>16,307</td>
</tr>
<tr>
<td></td>
<td>Ex-smoker</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Quit / Stopped</td>
<td>8,275</td>
</tr>
<tr>
<td>Unknown Smoking Status</td>
<td>Unknown if ever smoked</td>
<td>16,514</td>
</tr>
<tr>
<td></td>
<td>Unknown</td>
<td>58</td>
</tr>
<tr>
<td></td>
<td>Unable to assess</td>
<td>63</td>
</tr>
<tr>
<td></td>
<td>N/A / None</td>
<td>569</td>
</tr>
</tbody>
</table>

Table 2. Description of smoking status data during the one-year study period.

<table>
<thead>
<tr>
<th>Descriptor</th>
<th>Sample size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patients</td>
<td>48,909</td>
</tr>
<tr>
<td>Hospital Encounters</td>
<td>64,451</td>
</tr>
<tr>
<td>Encounters with smoking status recorded</td>
<td>62,988 (98%)</td>
</tr>
<tr>
<td>Encounters with smoking status recorded</td>
<td>19,176 (30%)</td>
</tr>
</tbody>
</table>
Figure 1. Changes of smoking status overtime. Dashed changes demonstrate implausible discrepancies and continuous lines represent plausible changes in longitudinal data. The number of changes recorded in our sample is reported in parentheses and the percentage it represents for each category is included in the figure.

On average the time interval between different smoking status documentation was 11 days, with a minimum of 0 days (same day documentation), and a maximum of 362 days. Most patients (80.6%) had a time interval between documentation events of less than or equal to 10 days, with 61.1% of patients having a subsequent documentation event within one day of the previous event.

Figure 2. Number of smoking status changes by time interval documentation. Time interval is measured in days. Number of status changes is represented in logarithmic scale. CDF = cumulative distribution function.
Duplicate Assessments During the Same Hospital Encounter

While it is plausible to observe changes in smoking status over the course of one year, smoking status should not change during the same hospital encounter. Given this rationale, we considered all changes during a hospital encounter to reflect a data quality issue, since smoking status should be consistent throughout a single admission. During the study period, we identified 59,663 distinct encounters from 45,822 patients. Of those, 32.2% of the hospital encounters (19,176 visits) had at least one conflicting smoking status recorded, which includes 14,798 patients (32.3% of our cohort of patients).

Discrepancies Among Various Provider Roles

For patients with a smoking status recorded in a structured field, 70.8% were documented as part of nursing notes, 12.9% came from social work notes, 11.6% from physician notes, and the remaining (4.7%) from notes entered by other health care professionals.

Among hospital encounters with more than one assessment of smoking status, encounters with documentation from a single role of provider (e.g., nurse) had fewer discrepancies compared with encounters containing smoking status assessments from providers with disparate roles (Table 3). For example, if multiple nurses documented smoking status during an admission, the number of distinct provider roles would be equal to one. However, if multiple nurses and multiple physicians documented this information, then the number of distinct provider roles would be two.

Table 3. Description of smoking status data during the one-year study period.

<table>
<thead>
<tr>
<th>Number of distinct provider roles</th>
<th>Encounters without discrepancies</th>
<th>Encounters with discrepancies</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>45.97% (8,791)</td>
<td>26.15% (6,650)</td>
</tr>
<tr>
<td>2</td>
<td>48.08% (9,195)</td>
<td>52.49% (13,345)</td>
</tr>
<tr>
<td>3</td>
<td>5.91% (1,130)</td>
<td>20.94% (5,323)</td>
</tr>
<tr>
<td>&gt; 3</td>
<td>0.05% (9)</td>
<td>0% (1)</td>
</tr>
</tbody>
</table>

How many patients are smokers?

Overall, 54.2% of the patients in our sample were classified as non-smokers, 6.5% as former smokers, 2.9% as current smokers and 3.2% as having unknown smoking status. The remaining 33.2% of patients had at least one discrepant assessment of smoking status documented. We determined that only 63.6% of our study population had a consistent, unchanging smoking status during the one-year study period (Figure 3).

Figure 3. Smoking status of patients seen in 2016.
Discussion

The Centers for Medicare and Medicaid Services (CMS) Meaningful Use program requires participating healthcare providers to record patients’ smoking status in a structured fashion. The program specifies eight distinct categories for collecting smoking status: “Current every day smoker”, “Current some day smoker”, “Former smoker”, “Never smoker”, “Smoker, current status unknown”, “Unknown if ever smoked”, “Heavy tobacco smoker” and “Light tobacco smoker”.

We identified smoking status assessments (either represented in free-text or structured fields) for 98% of patients and 98% of visits. When focusing on structured documentation, we observed that 94% of patients and 92% of visits had at least one structured smoking status observation recorded.

Despite the well-known and well-described importance of collecting smoking status, our institution’s EHR did not have a centralized location to store smoking status information. Smoking status was collected as part of clinical notes, in either structured or free-text format. The fact that disparate healthcare providers recorded this information in several different notes resulted in many inconsistencies across notes. Further complicating the matter, different note templates allowed for different granularities of smoking status data collection. Some templates included a free-text box that allowed clinicians to enter details such as intensity of smoking, number of cigarettes per day, or when the patient stopped smoking. Other templates had only the Meaningful Use-required structured fields embedded.

When analyzing smoking status data in the EHR, we transformed the Meaningful Use categories and other smoking status assessments into four clinically actionable categories: “Current smoker”, “Former smoker”, “Never smoker” and “Unknown smoking status”. We observed that a 33.2% of the patients had inconsistencies in the documented smoking status during the one-year study period and 32.3% of the patients had at least one discrepancy during a single visit. These discrepancies suggest that reliable information on smoking status may not be available for a large number of patients.

Since we used longitudinal data, and smoking status is not a static concept (i.e., it can change over time), we classified smoking status changes into two distinct categories: plausible and implausible. In our study, implausible changes constituted 21.3% of all changes. Previous research has also identified consistency issues regarding tobacco use recorded in different notes in EHR systems. For example, in 2016 a research study used natural language processing to parse clinical notes and extract smoking status from various clinical notes. The authors identified several inconsistencies when comparing smoking status recorded in clinical notes.

Inconsistencies can be attributed to challenges in the data collection process, including clinician-related and patient-related factors. Clinicians may not inquire at all about a patient’s smoking status, or they might ask the question in a manner that leads to bias in the patient’s answer. Depending on how clinicians phrase the question, patients may not feel comfortable answering. On the other hand, patients may have their own motivations to be less than truthful when providing smoking status information to clinicians, or they may inexplicably provide different smoking status responses depending on the person asking.

We conducted an analysis to identify whether hospital encounters with more than one clinical note without discrepancies were more likely to have documentation from a single provider role than encounters with discrepancies. Interestingly, we identified that encounters with multiple notes documented by the same type of provider had less discrepancies than patients with documentation from multiple types of providers. The difference we observed in discrepancies may be explained by the fact that clinicians usually do not read notes from other clinicians’ roles. Previous studies have shown that most clinical notes are not read by the entire clinical team. Instead, clinicians may be more inclined to read clinical notes from their peers (i.e., within the same provider role).

While it is important for multiple providers to assess patients’ smoking status, barriers to accessing previously documented information regarding tobacco use by healthcare providers may increase vulnerabilities that allow discrepancies to propagate.

Going back to our initial question of “how many patients are current smokers?” -- the answer is, we do not know. Based on the analysis conducted in this study, more than half of the patients during the one-year study period were recorded consistently as non-smokers and just 2.9% were recorded consistently as current smokers. In contrast, other population-based studies estimate that 15.1% of adult Americans smoke. One-third of the studied population had inconsistencies in their smoking status, making the determination of tobacco use for these patients difficult. While smoking status was documented in 98% of hospital encounters (and therefore the criteria of Meaningful Use were
satisfied), our one-year sample of hospital encounters did not contain consistent smoking status information for 36.4% of patients.

One limitation of our analysis was the use of data from only a single year and from only a single healthcare system. During a one-year period within our EHR system, we found that 33.2% of patients had discrepancies in documentation of smoking status. Furthermore, 54.5% of those inconsistencies were deemed implausible (Figure 1). Most patients had changes recorded within 10 days of the previous smoking status assessment. Given the short time difference between documentation events, even plausible changes (e.g., converting from “current smoker” to “former smoker”) seem unlikely. These data quality issues demonstrate just some of the considerable challenges healthcare providers and secondary users of EHR data. If we have difficulty in identifying a single meaningful and consistent smoking status using only one-year worth of data, the use and sharing of multiple years of data present even bigger challenges. For example, for encounters with conflicting smoking statuses, which one should be used in a clinical decision support system related to smoking cessation? Or which one should be reported to external organizations? Efforts using smoking status information from EHRs, including future smoking cessation initiatives, should further investigate patients identified as “Unknown smoking status” as well as patients with discrepancies in smoking status.

**Recommendations**

In this study, we observed that smoking status is currently being collected as part of clinical notes by multiple healthcare providers, and for almost all patients. The categories used are not consistent across clinical notes, recording smoking status in different granularities. We propose the use of four distinct clinically actionable categories: “Never smoker”, “Current smoker”, “Former smoker” and “Unknown smoking status”. More detailed information for each one of these could also be collected in a standardized fashion, such as “packs/day” and start and quit date. Currently, this additional information is being captured in free-text format and inconsistent across notes (e.g. some use packs/day while other record this information as cigarettes/day).

In our institution, smoking status is not stored in a centralized location, but is rather being collected as part of disparate clinical notes. The current system of data collection of smoking status presents challenges consistently collecting this information. While it is important for multiple providers to collect patients’ tobacco use information, the fact that this information is collected and stored in various notes without standardization makes it challenging for clinicians to know if the patient already provided their smoking status to other clinicians, and whether this information is longitudinally consistent. In an attempt to solve these challenges, we propose to store patients’ smoking status in a centralized fashion and having clinicians verifying this information in every encounter by asking patients about tobacco use.

One way to improve the consistency and correctness of patient-reported information, such as smoking status, is to allow patients to review and update their own information. This task can be facilitated by health information technology in many ways, including the use of patient portals and tablet computers for this task. Patient-facing tools have been used for collection of multiple patient-provided data types such as race and ethnicity, family history, symptoms, medication reconciliation and adherence. These studies have shown that patients are willing to provide and review their information\(^ {14-16}\), that EHR data is often incomplete or inaccurate\(^ {6-8,17-23}\) and that patients can identify discrepancies, provide useful information and help keeping records up-to-date.\(^ {18,24-26}\) Studies have also shown that there are many benefits of involving patients in their care, including improving patient engagement, patient satisfaction, health behaviors and health status as well as helping to attract and retain patients.\(^ {27,31}\) With patient-facing tools, patients could provide their smoking status based on the four clinically actionable categories, as described above. Patients providing this information to a computer could also mitigate the potential biases introduced by clinicians asking the question.

**Conclusion**

In summary, while 98% of hospital encounters at our institution during 2016 contained information regarding the patients’ smoking status, 32% of the encounters had discrepancies in smoking status information. For encounters with more than one clinical note documenting smoking status information, 54% of the subsequent documentation events had implausible changes. While other sources suggest that approximately 15% of adult Americans smoke, only 2.9% of our patients were consistently documented as current smokers. This finding demonstrates that while
Meaningful Use has improved data collection of smoking status in terms of completeness, we may not be appropriately identifying patients that smoke. Centralized documentation with clinically actionable smoking status categories available for data collection, and implementation of patient-facing tools that allow patients to directly record their information, may help improve data quality of smoking status in EHRs.

References


Design and evaluation of a web-based decision support tool for district-level disease surveillance in a low-resource setting

Meenal Pore, PhD1, David M. Sengeh, PhD1, Purity Mugambi, BSc1, Nuri V. Purswani, PhD2, Tom Sesay, MD3, Anna Lena Arnold, MPhil4, Anh-Minh A. Tran, MPH4, Ralph Myers, MSc4

1IBM Research | Africa, Nairobi, Kenya; 2IBM Singapore, Singapore; 3Ministry of Health, Port Loko District, Sierra Leone; 4GOAL Global, Dun Laoghaire, Ireland

Abstract

During the 2014 West African Ebola Virus outbreak it became apparent that the initial response to the outbreak was hampered by limitations in the collection, aggregation, analysis and use of data for intervention planning. As part of the post-Ebola recovery phase, IBM Research Africa partnered with the Port Loko District Health Management Team (DHMT) in Sierra Leone and GOAL Global, to design, implement and deploy a web-based decision support tool for district-level disease surveillance. This paper discusses the design process and the functionality of the first version of the system. The paper presents evaluation results prior to a pilot deployment and identifies features for future iterations. A qualitative assessment of the tool prior to pilot deployment indicates that it improves the timeliness and ease of using data for making decisions at the DHMT level.

Introduction

The 2014 West African Ebola Virus Disease (EVD) outbreak led to the deaths of over 11,000 people across the three worst affected countries: Guinea, Liberia and Sierra Leone1. A key factor in curbing the outbreak was data management and use in i) case management for patients, ii) operational and logistics planning for the response, and iii) community engagement for outbreak mitigation2. Knowledge of the viral transmission rate, the number and location of suspected cases and their contacts, estimated time of infection and the resources available at health facilities was vital for containing the 2014 EVD outbreak.

The Ministries of Health in the three most affected countries and the partners involved in the EVD response developed a plethora of digital and paper-based data collection solutions to provide real-time data to plan interventions. These tools produced a large number of datasets that were owned and stored by the various partners. A number of digital data collection solutions, such as KoboToolbox3 and CommCare4, as well as custom-built tools, were used to collect data in-field. Several organizations utilized existing data portals during the outbreak, such as Humanitarian Data Exchange, or developed new portals such as Ebola GeoNode and The World Health Organization Data Coordination Platform (WHO DCP), making these digitized datasets available for use across multiple partners.

The process of data collection and sharing was highly fragmented, leading to delays in data availability. For example, delays of several weeks were reported for data entry into the Centers for Disease Control's (CDC) Viral Hemorrhagic Fever (VHF) database2, with up to 50% of confirmed cases in the database lacking outcomes data2. The delays were exacerbated by the lack of a central coordination authority for data and the high turnover of staff, leading to a loss of institutional knowledge of the data available. In addition, digital solutions could not be implemented in all areas due to limited access to power and mobile networks, resulting in duplicate paper-based and digital systems. The Fighting Ebola With Information report produced by USAID extensively discusses the challenges with using data during the EVD outbreak2. While some of these issues were due to the organization of the international response, local infrastructure and skills capacity, there were additional challenges that could potentially have been addressed by technology solutions, as discussed below.

Competing reporting requirements: District-level staff had competing reporting requirements from different agencies and donors for the same data. Manually creating multiple reports was time-consuming, and limited the time available for operational and planning tasks.
Lack of standards for data entry: This could be broken down into three areas: entity resolution, lack of standardized definitions, and lack of standardized collection formats. Entity resolution was difficult due to the large number of entities, particularly facilities and places, with the same or similar names. Lack of standardized definitions led to incompatibilities between datasets e.g. different case definitions were used for reporting EVD cases in different districts. Finally, the lack of standardization of forms and terms between partners made it difficult to verify, collate and use data collected by different organizations.

Lack of structure and security for data sharing: The United Nations Mission for Emergency Ebola Response (UNMEER) was created to support the EVD response across affected countries. However, since it was a new organization, there were no data sharing policies established beforehand, resulting in a lack of clarity over how data could be shared. This led to data sharing delays or in some cases, data not being shared at all. Conversely, data with personally identifiable information were shared through platforms such as Dropbox, Google Drive and emails due to the lack of a secure sharing platform. Even where datasets were shared, information on metadata and version control was difficult to find or resolve.

Siloed data: Digital data collection tools deployed during the outbreak were largely set up as one-off pilots, with different applications built by different organizations to collect the same data in different regions. This resulted in over 300 new, digital solutions being deployed across Sierra Leone, Liberia and Guinea. These data terminated in a multitude of web endpoints, making it difficult to collate similar datasets.

Timeliness and quality of data for making decisions: Digital collection of data in low-resource health settings can improve the timeliness of data collection and reporting, and the quality and completeness of data. However, key information was often shared in non-machine readable formats, such as the CDC and The World Health Organization (WHO) case data that were reported as weekly Situation Reports in PDF format. These had to be manually entered into machine readable formats, delaying decision-making processes.

Skills in the health management workforce: The high volume of digital data collected required analysis by data teams skilled in using software packages, such as QGIS and R. However, the number of personnel trained to conduct the required analyses remained consistently insufficient throughout the outbreak.

In partnership with Port Loko DHMT and GOAL Global, we have developed a system that aims to address the challenges listed above by enabling easier and faster sharing and analyses of health data, for the prevention and detection of large-scale infectious disease outbreaks. The system was designed and developed to improve epidemic preparedness in the district by enabling fast detection and response in the event of an outbreak. Port Loko district, located in the northwest of Sierra Leone, was one of the areas hit hardest by the EVD outbreak. Below, we outline the requirements gathered for the system, an overview of the initial system developed, and the formative evaluation of the system gained through feedback from users during development and a pre-deployment qualitative evaluation.

Methods

User requirements: Initial user requirements were determined in collaboration with DHMT members. Process maps were created in consultation with the DHMT for each of the three datasets used for the system: Integrated Disease Surveillance and Response (IDSR), Infection, Prevention and Control (IPC) and Alerts (district-level reports of deaths or serious disease incidents in the community). During the in-development formative evaluation, the DHMT also helped to identify the key pain points in the current processes, which informed the user requirements. Other requirements for the system were gathered from DHMT team members iteratively using Scrum, a methodology comprising of a loop of ideation, software development, user demos, feedback from end users, and redesign over sprints for a period of six months. User requirements were classified into the following categories: 1) core functionality, 2) backend analytics, 3) usability, 4) security, and 5) network.

Systems requirements: System requirements were determined through an evaluation of the hardware available at Port Loko DHMT, and the needs for such a system beyond pilot scale deployments. These were classified into the following categories:

1. Hardware: What hardware is currently available at DHMTs?
2. Connectivity: What is the current network speed achievable at the DHMT?
3. Scalability: How can the system be scaled nationally and internationally?
4. Extensibility: Can additional functionality be added to the system without needing to redesign the architecture?
5. Data security regulations: Does the system comply with existing data security regulations in Sierra Leone?
**Formative evaluation:** In addition to the iterative feedback captured in the Scrum process, the first version of the system was evaluated through focus group discussions, one-on-one interviews, and a task completion exercise with key stakeholders in the DHMT. The focus group was conducted with DHMT members from the IDS, IPC, and Alerts teams. It was used to identify changes in process maps and data models from those captured at the start of the design process. This ultimately provided the baseline of processes used in disease surveillance and control prior to a pilot deployment of the system. The formative evaluation was also used to ascertain how users had interacted with the system to date, as they had access to the system during the iterative development. The focus group and interviews focused on four questions:

1) What are the key decisions that need to be made and by whom?
2) Which datasets are currently used to make these decisions?
3) What analyses are required to inform and support these decisions?
4) What technologies are currently used or are available to conduct the required analyses?

Individual interviews were conducted with key personnel and managers in the DHMT to understand how they had used, or planned to use, the system as part of their workflow, and to capture detailed feedback on the design of the system. Finally, as part of the one-on-one interviews, a task completion exercise was conducted, in which the interviewees were asked to retrieve and interpret a specific piece of information using the system. The goal was to assess the ease of manipulating data using the system.

The system was piloted from October 2016 to January 2017. The results discussed in this paper are from the iterative design process and the pre-deployment baseline study (conducted in September 2016) described above. The authors plan to publish a separate report on the findings from the pilot.

**Results**

**User and system requirements:** The user requirements gathered through the focus groups and one-on-one interviews could be clustered into five categories: core system, analytics, network, usability and security. Details of specific requirements are given in Table 1. The systems requirements were determined by the existing infrastructure at the District Health office in Port Loko, and are given in Table 2.

**System Architecture:** A schematic diagram of the architecture is given in Figure 1. The system comprises of four modules: i) data and data models, ii) data cleaning and curation services, iii) analytics services, and iv) visualizations. The current version of the system is built on a Python Django framework backend and AngularJS frontend deployed on a Red Hat Enterprise Linux 6.5 server. The backend was developed in Python 2.7 as established Python libraries for data manipulation, machine learning and natural language processing were available. The system is modular, and uses services exposed through Application Program Interface (API) calls. Optimization mechanisms such as minification, caching and background data processing were used to improve the performance of the system for use in areas with very low bandwidth.

The system was deployed on a remote, secure server with HTTP protocols. To access and use the system, users connected to the machine via a VPN connection. Once their log-in was authenticated, they could view pages for which they had authorization.

**Data and data models:** In the current implementation, datasets are uploaded via APIs to a Postgres database. A Postgres database was used due to its ability to ingest relational data.

**Data cleaning, curation, and analytics services:** Disparity in data standards across datasets necessitates that the system can dynamically create data models. Through Data Access Linking and Integration (DALI)\textsuperscript{11}, an IBM Research proprietary software, each file is processed and stored in its own schema. Relationships between schemas are created using semantic closeness between fields. Data cleaning and curation services ensure that relational formats (.xls, .csv, .dbf) can be uploaded into the system. Entities, such as health unit names, are resolved to a single identifier across datasets. These datasets are cleaned and curated using a series of API calls, utilizing new APIs for entity resolution and geocoding using OpenStreetMap place names. The backend analytics services for both health facility and community risk profiling were done using custom-built APIs.

**Visualization:** HTML, CSS and JavaScript were used to build the visualization dashboards.
Table 1: User requirements identified from requirements elicitation during the development of the first version of the system.

<table>
<thead>
<tr>
<th>Category</th>
<th>Component</th>
<th>Functionality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Core system</td>
<td>Datasets</td>
<td>Must include 1) Integrated Disease Surveillance and Response (IDSR, received weekly) 2) Infection, Prevention and Control (IPC, received monthly), and 3) Alerts data (received daily).</td>
</tr>
<tr>
<td></td>
<td>Input</td>
<td>Must accept tabular data formats (e.g. Excel and CSV). Must integrate with common mobile data collection APIs (KoboToolbox and CommCare).</td>
</tr>
<tr>
<td></td>
<td>Output</td>
<td>Cross-browser web interface.</td>
</tr>
<tr>
<td>Analytics</td>
<td>IDSR, IPC</td>
<td>Must follow current MoHS guidelines.</td>
</tr>
<tr>
<td></td>
<td>Alerts</td>
<td>Use baseline mortality rates for Sierra Leone for thresholds.</td>
</tr>
<tr>
<td></td>
<td>Reporting</td>
<td>Must include whether facilities reported on-time for IDSR and IPC.</td>
</tr>
<tr>
<td></td>
<td>Risk score</td>
<td>Provide aggregate risk statistics for all three datasets.</td>
</tr>
<tr>
<td></td>
<td>Smart search</td>
<td>Ability to geographically and semantically search over related datasets that are collected \textit{ad hoc} and are not included in the district health information system (DHIS2).</td>
</tr>
<tr>
<td>Network</td>
<td>Loading time</td>
<td>The initial loading of the page should take no longer than 5 seconds under normal conditions in Port Loko DHMT.</td>
</tr>
<tr>
<td>Usability</td>
<td>Data subsets</td>
<td>System must provide functionality to select data by district, chiefdom and section.</td>
</tr>
<tr>
<td></td>
<td>Data visualization</td>
<td>Data should be represented on graphs and maps, and visualizations should be exported as PDFs.</td>
</tr>
<tr>
<td></td>
<td>Training requirements</td>
<td>The system should require minimal training (less than a day) and should not require specialized software skills (e.g. GIS data skills).</td>
</tr>
<tr>
<td></td>
<td>Security</td>
<td>Access privileges Due to the sensitive nature of the data stored, each user requires secure, individual access to the server.</td>
</tr>
</tbody>
</table>

Table 2: System requirements identified from requirements elicitation during the development of the first version of the system.

<table>
<thead>
<tr>
<th>Category</th>
<th>Functionality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hardware</td>
<td>System will be access via a variety of laptop devices.</td>
</tr>
<tr>
<td>Connectivity</td>
<td>System must be functional on a 6 Mbps internet connection using a 3G modem.</td>
</tr>
<tr>
<td>Scalability</td>
<td>System should be able to store two years' of historical data for a country-level deployment. System must be deployable remotely in an emergency situation.</td>
</tr>
<tr>
<td>Extensibility</td>
<td>System must be capable of incorporating routine datasets other than disease surveillance, and have the ability to build more advanced analytics in the future.</td>
</tr>
<tr>
<td>Data security regulation</td>
<td>Users require privileged access to data. The system must be deployable remotely or in country, depending on countries' regulations on whether data can leave country borders.</td>
</tr>
</tbody>
</table>
Core functionality

The web application consists of four pages: Home, Health Units, Reports and Data Store. The charts on all pages can be exported as PDF files.

**Home Page:** The home page provides the DHMT management with an overview of reporting rates for the number of facilities reporting on time, late or not reporting, along with the aggregate risk levels (a description of how the risk level was calculated is given later). A screenshot of the page (Figure 2) shows the IDSR reporting rates for the district and the risk levels for a given week. The users can choose which administrative level (district, chiefdom and section) they wish to aggregate data over, and the week they wish to see.

**Health Units Page:** The Health Units page provides a list of the peripheral health units (PHU) in the chosen administrative level, prioritized by risk level, as shown in Figure 3. The page provides the DHMT with a view of the most at-risk facilities in a given week. A description of how the risk score is calculated is provided in the Analytics section. The color coding on the page is used to indicate which dataset triggered the alarm, along with a description of the alarm. If the user requires further information about the health facility, the user can click on the menu icon to the right to see a full report, including IDSR, IPC and Alerts data over time for that facility.

**Reports Page:** The Reports page consists of three subpages for each of the datasets: IDSR, IPC and Alerts. Each page provides a map with markers colored by risk level for each of the health facilities. In addition, it includes the reporting rates over time. For IDSR, the page also has an interactive graph of disease notifications, allowing the user to investigate trends over time. Components of the IDSR Reports page are shown in Figure 4. For IPC, the user can investigate specific measures (the assessment measures included were defined by members of the DHMT). For Alerts, the user can select specific case definitions, locations and age ranges. For all pages, the user can choose to view data for a given administrative level and over a range of dates.

**Data Store Page:** The data store accepts non-routine tabular data, such as ad hoc assessments and surveys to allow these data to be easily shared between the DHMT and partners. This page allows users to upload tabulated data and, if the data contains GIS information, to automatically map the data. Users can search across all data sets geographically, using a bounding box to select a geographic area of interest, or semantically. The semantic search utilizes the DALI semantic search API, which is based on ontologies from DBPedia$^{11,12,13}$. A screenshot of the data store page is shown in Figure 5.

**Analytics**

**IDSR, IPC and Alerts:** The risk level for IDSR was determined by whether a disease reported was classified as immediately notifiable$^{14}$, as described in Figure 6. The latest version of the IPC assessment available in November...
2016 was used, which consisted of nine sections on the availability, process and knowledge around protective equipment, decontamination, water, sanitation, screening and isolation. The IPC survey consisted of 68 binary questions, which are in turn converted into a percentage score. The thresholds for the risk scoring shown in Figure 6 are based on the MoHS thresholds used in the national IPC Excel template.

![Figure 2: The Home page provides an overview of the aggregate reporting rate (1) and risk level (2) of facilities.](image)

![Figure 3: The Health Unit page provides a list of health units in a given geographic area selected by the user (1), prioritized by highest risk score. The user can click on the report icon to see a full report for each facility (2).](image)

Alerts are reported by chiefdom through the national health emergency hotline, local district alerts lines or, during the outbreak, through conversations with Ebola responders. All alerts were entered into the eHealth Africa call center system, and the risk score was calculated based on deaths per 1000 population relative to the a baseline mortality value of 17.4 deaths of per month for Sierra Leone, based on data from 2012. The risk level for Alerts was determined by using a moving window to average the deaths reported in the previous four weeks, as given by Equation 1, in which \( N_c \) is the number of deaths per week in a given chiefdom, \( t \) is the current week number, \( P_c \) is the population in a given chiefdom in thousands and \( M.R. \) is mortality rate per 1000 population. The chiefdom was said to be high risk if the monthly mortality rate was below 17.4 deaths per 1000 population, not reporting if no deaths were reported, and low risk if \( 0 < M.R. < 17.4 \) deaths per 1000 population.

\[
M.R. = \frac{N_{ct} + N_{ct-1} + N_{ct-2} + N_{ct-3}}{P_c}
\]

*Equation 1*
Figure 4: IDSR Report page showing (1) health facilities color coded by their current risk level, (2) cumulative reporting rates for the chosen geographic area and (3) disease incidence trends for the chosen geographic area.

Figure 5: A screenshot of the Data Store Page. The user can (1) easily upload local tabulated data files, (2) map files with GIS information and search for files related to the selected file or word using semantic search, (3) view the data in a tabular view and (4) search for files by geographic area using a bounding box.

Performance
Network speeds were measured between clients in Nairobi, Freetown and Port Loko, and the remote server. Due to lower speeds and a less reliable connection, loading times for the first version of the system, which was not optimized, were considerably slower with a stable connection in Sierra Leone (80 seconds) when compared to Kenya (10 seconds). The system was optimized to meet the loading requirement for the entire website of less than 5 seconds under normal conditions in Port Loko.
**Formative evaluation**

The following five key themes emerged from the formative and pre-deployment evaluations.

1) **Visually integrate multiple datasets for decision making:** Prior to the deployment of the system, each DHMT team (IDSR, IPC and Alerts) analyzed only their own datasets for presentation to the rest of the DHMT on a weekly basis. Having access to all three datasets, particularly side-by-side on the Health Units page, allowed users to contextualize information and interact with other data. For example, if a given PHU had an alarm for IDSR, the team could easily check the IPC risk score to see if there were sufficient measures in place to manage the case or cases recorded in IDSR.

![Facility risk score = IDSR score + IPC score (max = 8, min = 2)](image)

**Figure 6:** A schematic diagram of the facility risk scoring algorithm used in the deployed version of the system.

2) **Prioritization of information:** During the pre-deployment evaluation, several suggestions were made for refining the facility risk scoring. Specifically:

- Use of both the absolute number of cases, as well as the population normalized notification rate of an immediately notifiable disease in the scoring algorithm;
- Addition of a weighting factor based on reporting rates for expected death rate per chiefdom, since reporting rates have dropped significantly since the end of the EVD outbreak. The adjusted threshold could then be used to determine under-reporting more accurately. For example, $T_{\text{underreport}} = A r_{\text{reporting}} - M R_{\text{baseline}}$, where $T_{\text{underreport}}$ is the threshold for underreporting for a given chiefdom, $A$ is a constant chosen as the threshold below which an alarm is raised for underreporting, $r_{\text{reporting}}$ is the reporting rate based either from extrapolation of reporting rates over time or a value calculated during an audit, and $M R_{\text{baseline}}$ is the baseline mortality for the country.

3) **Converting decision making into action:** After interacting with the system, several users suggested incorporating a means to track the completion of follow-on actions, such as case investigations by the District Rapid Response Teams (RRTs) for immediately notifiable diseases, into the platform.

4) **Interface Omissions:** From the in-development and pre-deployment evaluations, several user interface omissions were detected. A Comparison page was added as a result of user testing. The page allows IDSR data to be compared between PHUs, and Alerts data to be compared between chiefdoms to enable the DHMT to quickly assess the relationships between PHUs of interest and to visualize trends across chiefdoms.

5) **Improvements to the timeliness of analysis and decision making:** The eight key users interviewed from the DHMT team found that the first version of the system reduced the time taken for the completion of routine tasks. Depending on the specific task, they reported the time taken was reduced to 5 to 15 minutes using the system, compared with 30 minutes to 8 hours using manual analysis. Overall, the system was found to be helpful by members of the DHMT for easy and quick visualization of data for making decisions. The team reported that they could use the system to generate reports and explore data in real-time for routine district health coordination meetings and partners’ meetings held at the DHMT.

**Discussion**

**Sustainability of the platform in the face of data changes:** Data structures changed significantly throughout the development process. During the 6-month scrum process, the IPC structure as provided by the MOHS/WHO changed five times, and the IDSR format was changed twice. While this affected system features, particularly in the
Building smarter epidemic preparedness systems: The initial analytics built into our system were based on the workflows described by users. However, through the formative evaluation, it became apparent that, once data were easily available and presented in such a way that it was easy to consume, users requested further analytics based on their current decision making processes. Some of these would be relatively simple to implement, such as including diseases that are not immediately notifiable in the IDSR risk calculation, and weighting the risk by the number of cases or the population normalized incidence. However, certain aspects are less clear: how should cases at referral facilities be weighted relative to primary facilities given what they are likely to see and be more prepared for more severe cases? In addition, the deployed system could allow more complex analytics, such as epidemiological models, to be run in real time, eventually allowing proactive planning for seasonal outbreaks, such as waterborne diseases during rainy season. Finally, this architecture opens up opportunities for providing further decision support, specifically around planning for potential scenarios.

Conclusions
This paper presents the design and evaluation of an epidemic preparedness system for DHMTs in low-resource settings. The system discussed in this paper was developed in partnership with, and for use by the Port Loko DHMT in Sierra Leone after the 2014 EVD outbreak. It integrates and analyzes multiple data sources and data types in near real-time. It was shown that such a system reduced the time taken for current workflows to five to fifteen minutes, compared eight hours to thirty minutes prior to the system, depending on the specific task. One of the key strengths of the deployed system was the ability for the DHMT to quickly retrieve relevant data for decision-making. For example, from the Home page, they can quickly drill down to see which facilities had not reported for that week, or which were reported as high risk. In addition, the Health Unit page allows users to quickly prioritize which data to look at. The formative evaluation indicated that a system that provides support for day-to-day operations at the DHMT across all programs, not just disease surveillance, would be valuable. Such a system would allow workflows to be tracked, and data to be automatically pushed to and pulled from systems such as DHIS2 and OpenMRS. The development and testing of the system demonstrated the feasibility of running a cloud-based health analytics service.
in a low-resource setting. The formative evaluation highlighted the potential of future versions of the system to enable higher level analytics and decision support to be introduced into the workflow of the DHMT, and for the integration of the system with public health information systems.

Acknowledgments
The authors would like to acknowledge USAID’s Fighting Ebola Grand Challenges (BAA-EBOLA-2014 AID-OAA-A-15-00041) fund for financially supporting this work. They would like to thank Port Loko DHMT and GOAL Global, Sierra Leone for their cooperation and technical expertise that made the project possible, particularly the GOAL Port Loko surveillance data team including Hilton Matthews, Alhassan Dumbaya and Ibrahim Yansaneh.

References
Stage-Specific Survivability Prediction Models across Different Cancer Types

Elham Sagheb Hossein Pour¹, MS, Rohit J. Kate², PhD  
¹Biomedical Informatics Research Center, Marshfield Clinic Research Institute, Marshfield, WI 54449, USA; ²Department of Health Informatics and Administration, University of Wisconsin-Milwaukee, Milwaukee, WI 53211, USA

Abstract

For all cancer types, survivability rates vary widely across different stages of cancer. But survivability prediction models built in past were trained using examples of all stages together and were also evaluated on all stages together. In this work, for ten cancer types and using three machine learning methods, we built survivability prediction models trained on each stage separately and compared their performance with the traditional models trained on all stages together. For both kinds of models, the evaluation was done on each stage separately as well as on all stages together. Our results show that for most cancer types the stages are sufficiently different from each other that it is best to build survivability prediction models separately for each stage. We also found that evaluating survivability prediction models on all stages together, as was done previously, overestimates performance for all the stages on all cancer types.

Introduction

Cancer is one of the leading causes of morbidity and mortality worldwide with approximately 14 million new cases and 8.2 million cancer related deaths recorded in 2012, the latest year for which these numbers are available [1]. In the US, 25.52% of total deaths in 2014 were due to cancer making it the second leading cause of death after heart diseases [2]. Such high rates of morbidity and mortality have prompted researchers to build survivability prediction models for cancer [3, 4, 5, 6]. Accurate prediction of survivability can help physicians make more informed decisions about their patients’ treatment. For example, they may choose new medications or more aggressive therapies for patients with less hope of survival.

Cancer survivability is commonly defined as surviving for five years after diagnosis and its rate greatly depends on the stage of the cancer which is assigned based on tumor size and the extent of spread. Machine learning methods have been extensively used for building predictive models for cancer survivability [7, 8]. Although researchers in past have used a wide variety of machine learning methods and training mechanisms for building cancer survivability prediction models, they did not distinguish between various stages of cancer either during training or during evaluation. Stage was used only as one of the several features for building the models. However, it was recently found that for breast cancer the most suitable model to predict survivability for a stage is the model which was trained specifically for that stage [9]. It was also found that evaluating the models on all stages together, as was done in past, leads to an overestimation of performance. While that work was specific to breast cancer, in this paper we investigated whether those findings also generalize to other cancer types.

We built survivability prediction models using three different machine learning methods for ten different cancer types. To the best of our knowledge, it is the first work to report performance of survivability prediction models across multiple cancer types in a uniform experimental setting. This helps in generalizing the findings across cancer types. For each cancer type, we trained models for each stage separately and compared them with models trained on all stages together. We evaluated both types of models on each stage separately as well as on all stages together. Our results show that the findings regarding breast cancer [9] generalize to most other cancer types. The only few exceptions were for the cancer types with similar survivability rates across various stages along with insufficient number of training examples available for a few of its stages.

Experimental Methodology

Dataset

We used the publicly available SEER cancer dataset [10] for the experiments presented in this paper. This dataset is a collection of cancer incidences that happened in the US since 1973 in certain geographical regions that represent...
around 28% of the US population. The data is collected in an ongoing basis and the latest version used in this study covers years up to 2013 for a total of 9.18 million cancer incidences.

**Table 1:** List of features used to build predictive models for breast cancer survivability.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>Numeric</td>
<td>Age at diagnosis</td>
</tr>
<tr>
<td>Behavior code</td>
<td>Nominal</td>
<td>Code based on aggressiveness of tumor</td>
</tr>
<tr>
<td>Extension</td>
<td>Nominal</td>
<td>Information on extension of tumor</td>
</tr>
<tr>
<td>Grade</td>
<td>Nominal</td>
<td>Category based on the appearance of tumor</td>
</tr>
<tr>
<td>Histologic type</td>
<td>Nominal</td>
<td>Form of tumor</td>
</tr>
<tr>
<td>Lymph nodes</td>
<td>Nominal</td>
<td>The highest specific lymph node chain that is involved by the tumor</td>
</tr>
<tr>
<td>Marital status</td>
<td>Nominal</td>
<td>Marital status at diagnosis</td>
</tr>
<tr>
<td>Metastasis at diagnosis</td>
<td>Nominal</td>
<td>Information on distant metastasis</td>
</tr>
<tr>
<td>Primary site</td>
<td>Nominal</td>
<td>Site in which the primary tumor originated</td>
</tr>
<tr>
<td>Race</td>
<td>Nominal</td>
<td>Method of radiation therapy used in the first course of treatment</td>
</tr>
<tr>
<td>Regional nodes examined</td>
<td>Numeric</td>
<td>Number of regional lymph nodes removed and examined</td>
</tr>
<tr>
<td>Regional nodes positive</td>
<td>Numeric</td>
<td>Number of regional lymph nodes that contained metastases</td>
</tr>
<tr>
<td>Sequence number central</td>
<td>Nominal</td>
<td>Sequential number of tumor during the life time of patient</td>
</tr>
<tr>
<td>Site-specific surgery code</td>
<td>Nominal</td>
<td>Code for surgery of primary site as first course of therapy</td>
</tr>
<tr>
<td>Summary stage</td>
<td>Nominal</td>
<td>Defined according to the spread of cancer</td>
</tr>
<tr>
<td>Tumor size</td>
<td>Numeric</td>
<td>Size in mm</td>
</tr>
</tbody>
</table>

Although the dataset does not include any genomic information or detailed clinical information about patients, it includes several demographic and cancer related attributes for each incidence which can be used as features to predict survivability. Table 1 shows the features that we used in this work. These are largely the same features as have been used in the past work [5, 11, 12, 9] and their detailed descriptions can be found in the documentation of the SEER dataset [10]. Gold standard five-year survivability was determined using survival months, vital status recode and cause of death attributes of SEER dataset using the logic shown in Figure 1. SEER uses the summary stage system for cancer incidences. In this staging system, a cancer incidence can be in-situ (cancer confined to the layer of cells), localized (limited to the organ), regional (has spread to nearby lymph nodes, tissues and organs) or distant (has spread to distant lymph nodes, tissues and organs) summary stage. In past work summary stage was used only as a feature, but in this paper we built prediction models as well as evaluated them separately for each summary stage.

**Figure 1:** Logic used to determine gold standard five year survivability of patients from each cancer type. Survival months, vital status recode and cause of death are attributes in the SEER dataset.

Cancer survivability rates have improved over the years. For example, breast cancer 5-year survivability rate was 75.2% in 1980 in the US and was 90.6% in 2013 [13]. Similarly, it was 48.6% for colorectal cancer in 1975 and was 67.2% in 2013. Hence it is not advisable to use very old data for training and evaluating survivability prediction models. Codes of some of the attributes in the SEER data were redefined in the year 2004 and a few new attributes were also introduced. For the above two reasons, we decided to use only the part of the SEER dataset from the year 2004 onwards. Given that survivability is defined as surviving for five years after diagnosis, we had to also exclude incidences which were diagnosed less than five years before the latest year in the current dataset. The survival rate for in-situ stage is nearly 100% for all the cancer types hence we decided to exclude that stage for survivability prediction.
**Prediction Models**

In this study, we considered ten cancer types – bladder, breast, cervix uteri, colorectal, corpus uteri, esophagus, liver, lung, prostate and stomach. These are among the most common cancer types found in men and women [1]. About 1% of breast cancer incidences occur in men, but for this study we included incidences of only women. We built two kinds of survivability prediction models for each cancer type. For a cancer type, a model of the first kind was built to work for all stages and was hence trained using incidences of all stages together. We call this kind of model a joint model. Most previous work had used only this kind of model in which stage was used only as a feature. Our joint models also use stage as a feature (Table 1). For a cancer type, models of the second kind were built to work for each stage separately. Each model of the second kind was trained using incidences of only its particular stage. We call these stage-specific models. We compared joint models and stage-specific models for predicting survivability on the ten cancer types.

Both joint and stage-specific models can be evaluated on all stages together as well as on each stage separately. When evaluated on all stages together (test data consists of incidences of all stages), a single classification evaluation measure is reported for all the stages. When evaluated on each stage separately (test data consists of incidences of only one stage), a single classification evaluation measure is reported for each stage. Note that stage-specific models can be tested on all stages together by classifying each test incidence using the stage-specific model of that incidence’s stage.

**Machine Learning Methods**

We used naive Bayes [14], logistic regression [15] and decision trees [16] machine learning methods to predict cancer survivability. We did not use support vector machines [17] or neural networks [18] because they would take impractical amount of computational time on cancer types with very large number of incidences and in pilot studies we did not find them to perform better than the methods we used. All these methods have been previously used for cancer survivability prediction [7, 8]. We used the Weka software [19] for all the methods. Among various decision tree methods available in Weka, we used “ADTree” (alternating decision trees [20]) which we found to work best for our dataset through a pilot study. The three machine learning methods we used do not have any major parameter to set in Weka and we used Weka’s default values for all the minor parameters. We also used Weka’s default mechanism for handling missing feature values for each of these methods. However, cancer incidences that had the stage missing or the information to determine patient’s survival missing were excluded from our experiments.

Given the unbalanced nature of our dataset, from very high to very low survival rates among various stages of various cancer types, we used Weka’s cost-sensitive meta-classifier which relatively penalizes misclassifying minority class compared to misclassifying majority class in order to maximize classification accuracy across both the classes. The right weight was determined out of 0.25, 0.5, 1, 2, 4, ..., 18, 20 through internal five-fold cross-validation within the training data.

**Evaluation**

Separate evaluation was done for each cancer type. We used the standard ten-fold cross-validation to evaluate all our models. All folds were stratified to have the same distribution of stages and classes. For each fold, the training data used for the joint model was divided into disjoint subsets according to the stages of the incidences and each subset was used to train the corresponding stage-specific model for that fold. This makes the comparison between joint and stage-specific models fair. The test data was the same for each fold for both kinds of models. The exact same folds were used for each of the machine learning methods.

We used the standard area under ROC curve (AUC) [21] as our evaluation measure. ROC is a curve between a model’s true positive rate (fraction of positive examples correctly classified as positive, also called sensitivity) and false negative rate (fraction of negative examples incorrectly classified as positive, equivalent to 1-specificity). In our experiments, positive class is “survived” and negative class is “not survived”. ROC curve is independent of class distribution [22] which makes it a robust evaluation measure for our task in comparison to other evaluation measures such as precision and recall. Given the large number of evaluations in our experiments, we report only the area under ROC curve (AUC) which summarizes an ROC curve. A perfect classifier has an AUC of 1. When stage-specific models are evaluated on all stages together, their respective ROC curves for each stage needed to be
<table>
<thead>
<tr>
<th>Statistics</th>
<th>Naive Bayes</th>
<th>Logistic Regression</th>
<th>Decision Tree</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Incidences</td>
<td>Survived</td>
<td>Joint</td>
</tr>
<tr>
<td>All Stages</td>
<td>25250 (100%)</td>
<td>53.19%</td>
<td>0.859</td>
</tr>
<tr>
<td>Localized</td>
<td>18598 (73.66%)</td>
<td>65%</td>
<td>0.826</td>
</tr>
<tr>
<td>Regional</td>
<td>4211 (16.76%)</td>
<td>29.05%</td>
<td>0.726</td>
</tr>
<tr>
<td>Distant</td>
<td>2421 (9.59%)</td>
<td>4.67%</td>
<td>0.666</td>
</tr>
<tr>
<td>Breast</td>
<td>208767 (100%)</td>
<td>86.78%</td>
<td>0.872</td>
</tr>
<tr>
<td>Localized</td>
<td>131688 (63.08%)</td>
<td>95.16%</td>
<td>0.77</td>
</tr>
<tr>
<td>Regional</td>
<td>65942 (31.59%)</td>
<td>81.59%</td>
<td>0.786</td>
</tr>
<tr>
<td>Distant</td>
<td>11137 (5.33%)</td>
<td>18.49%</td>
<td>0.662</td>
</tr>
<tr>
<td>Cervix Uteri</td>
<td>12749 (100%)</td>
<td>70.55%</td>
<td>0.882</td>
</tr>
<tr>
<td>Localized</td>
<td>6500 (50.98%)</td>
<td>92.78%</td>
<td>0.805</td>
</tr>
<tr>
<td>Regional</td>
<td>4729 (37.09%)</td>
<td>57.73%</td>
<td>0.709</td>
</tr>
<tr>
<td>Distant</td>
<td>1520 (11.92%)</td>
<td>15.39%</td>
<td>0.683</td>
</tr>
<tr>
<td>Colorectal</td>
<td>128385 (100%)</td>
<td>61.45%</td>
<td>0.89</td>
</tr>
<tr>
<td>Localized</td>
<td>52067 (40.56%)</td>
<td>87.58%</td>
<td>0.786</td>
</tr>
<tr>
<td>Regional</td>
<td>47235 (36.79%)</td>
<td>64.71%</td>
<td>0.748</td>
</tr>
<tr>
<td>Distant</td>
<td>29083 (22.65%)</td>
<td>9.39%</td>
<td>0.766</td>
</tr>
<tr>
<td>Corpus Uteri</td>
<td>36307 (100%)</td>
<td>91.5%</td>
<td>0.91</td>
</tr>
<tr>
<td>Localized</td>
<td>28127 (77.47%)</td>
<td>97.35%</td>
<td>0.817</td>
</tr>
<tr>
<td>Regional</td>
<td>6617 (18.23%)</td>
<td>81.22%</td>
<td>0.806</td>
</tr>
<tr>
<td>Distant</td>
<td>1563 (4.3%)</td>
<td>29.75%</td>
<td>0.718</td>
</tr>
<tr>
<td>Esophagus</td>
<td>12448 (100%)</td>
<td>14.78%</td>
<td>0.862</td>
</tr>
<tr>
<td>Localized</td>
<td>2998 (24.08%)</td>
<td>33.56%</td>
<td>0.84</td>
</tr>
<tr>
<td>Regional</td>
<td>4215 (33.86%)</td>
<td>16.09%</td>
<td>0.749</td>
</tr>
<tr>
<td>Distant</td>
<td>5235 (42.05%)</td>
<td>2.96%</td>
<td>0.758</td>
</tr>
<tr>
<td>Liver</td>
<td>15309 (100%)</td>
<td>18.95%</td>
<td>0.896</td>
</tr>
<tr>
<td>Localized</td>
<td>7551 (49.32%)</td>
<td>30.72%</td>
<td>0.857</td>
</tr>
<tr>
<td>Regional</td>
<td>4771 (31.16%)</td>
<td>10.23%</td>
<td>0.88</td>
</tr>
<tr>
<td>Distant</td>
<td>2987 (19.51%)</td>
<td>3.11%</td>
<td>0.843</td>
</tr>
<tr>
<td>Lung</td>
<td>182958 (100%)</td>
<td>13.3%</td>
<td>0.897</td>
</tr>
<tr>
<td>Localized</td>
<td>27811 (15.2%)</td>
<td>47.43%</td>
<td>0.837</td>
</tr>
<tr>
<td>Regional</td>
<td>42258 (23.1%)</td>
<td>19.91%</td>
<td>0.783</td>
</tr>
<tr>
<td>Distant</td>
<td>112889 (61.7%)</td>
<td>2.42%</td>
<td>0.688</td>
</tr>
<tr>
<td>Prostate</td>
<td>219103 (100%)</td>
<td>94.03%</td>
<td>0.926</td>
</tr>
<tr>
<td>Localized</td>
<td>182954 (83.5%)</td>
<td>97.37%</td>
<td>0.858</td>
</tr>
<tr>
<td>Regional</td>
<td>27561 (12.58%)</td>
<td>95.2%</td>
<td>0.844</td>
</tr>
<tr>
<td>Distant</td>
<td>8588 (3.92%)</td>
<td>19.13%</td>
<td>0.682</td>
</tr>
<tr>
<td>Stomach</td>
<td>18107 (100%)</td>
<td>28.4%</td>
<td>0.924</td>
</tr>
<tr>
<td>Localized</td>
<td>4945 (27.31%)</td>
<td>65.62%</td>
<td>0.891</td>
</tr>
<tr>
<td>Regional</td>
<td>5962 (32.93%)</td>
<td>27.59%</td>
<td>0.813</td>
</tr>
<tr>
<td>Distant</td>
<td>7200 (39.76%)</td>
<td>3.51%</td>
<td>0.775</td>
</tr>
</tbody>
</table>

Table 2: Area under ROC curve (AUC) results comparing joint and stage-specific survivability prediction models built using the three machine learning methods. Separate models were built for each cancer type. A number shown in bold was found to be statistically significant (p < 0.05; two-tailed paired t-test) compared to the number in the same row for the same machine learning method. The numbers in the “All Stages” row of each cancer type can be seen to be misleadingly high when compared to the numbers for the individual stages.
combined into one ROC curve. We achieved this by considering all combinations of individual confidence thresholds to determine combined true positive rate and false positive rate to plot one ROC curve [9].

**Results and Discussion**

Table 2 shows the results obtained by both the joint and stage-specific survivability prediction models (shown column-wise) evaluated on all stages together as well as on each stage separately (shown row-wise). For each cancer type, the models were separately built and evaluated. The first two columns of the table show the number of incidences and survival rates of various stages of the ten cancer types. The rest of the columns show the AUC values obtained by the two kinds of models built using the three machine learning methods. Any AUC value shown in bold was found to be statistically significant (p < 0.05; two-tailed paired t-test) when compared to the corresponding AUC value of the other type of model in the same row for the same machine learning method.

Several observations can be made from Table 2. First, a broad observation is that survival prediction models generally obtain reasonable AUC values on all the cancer types. This indicates that the information present in the SEER dataset is good enough to build reasonably accurate cancer survivability prediction models.

Looking closely at the results obtained by evaluating the models on all stages together (first row of each cancer type) and evaluating them separately on each stage (rest of the rows of that cancer type), one can make a striking observation that the performance obtained on all stages is almost always higher than obtained on each individual stage. Although this appears counter-intuitive, there is an explanation for it. Given the differing survival rates of the three stages, from high for localized to low for distant, it is possible to obtain a seemingly high survivability prediction performance when evaluated on all stages together by simply making predictions strictly based on the stages. For example, calling everyone in the localized stage to survive and others to not survive will give a high accuracy because the predictions will be often correct. But such a predictive model is practically of little use because for a particular cancer incidence it does not tell anything more than what is already known. When such a classifier is evaluated separately on each stage, its performance will be as bad as a majority classifier that predicts every incidence to belong to the majority class. Hence this classifier will do well when evaluated on all stages together but will do very poorly when evaluated on each stage separately.

Learned models also suffer from this tendency because given the cancer survivability data they can easily learn to rank localized stage incidences above regional stage incidences and these in turn above distant stage incidences and thus similarly get a phony boost in performance when evaluated on all stages together. This is exactly what happened in the results shown in Table 2. The results obtained by evaluating on all stages together are thus overestimations of the actual performance. This was earlier observed for breast cancer [9]. Our results show that it is true for other cancer types as well. It is important to note that all past work had only reported performance on all stages together and were thus reporting overestimations. For the rest of the paper, we do not consider the results in the “All Stages” rows of Table 2.

Comparing joint models and stage-specific models for different cancer types and different machine learning methods taking into account statistical significance, one can observe that stage-specific models often do better than joint models. It should be noted that joint models always get examples of all the stages for training hence they get significantly more number of training examples than stage-specific models which get examples that are always subset of the examples joint models get for training in each fold. In spite of that, stage-specific models outperform joint models. For example, for prostate cancer, while the joint model gets 197192 examples for training (9/10th of the total examples in each fold of 10-fold cross-validation), the stage-specific model for the distant stage gets only 7729 examples for training (9/10th of total distant stage examples), but yet the stage-specific model for the distant stage outperforms the joint model on that stage in case of each machine learning method. This shows that various cancer stages are distinct enough from each other that it is best to build separate survivability prediction models for each stage. Adding examples of other stages during training does not benefit in predicting survivability for a particular stage, in fact, it often hurts the performance.

A fundamental requirement for machine learning methods to work well is that training and testing examples come from the same distribution, otherwise the model learned from the training data will not generalize to the test data. The above results indicate that examples of different cancer stages constitute different data distributions and hence mixing them during training, as joint model does, is not beneficial but hurtful. The recent study, [9], had reported

1408
this observation only for breast cancer (our numbers do not exactly match their numbers for breast cancer because they had excluded examples with missing feature values). Our results show that this observation generalizes to other cancer types as well.

**Figure 2:** A graph showing differences between the distant summary stage of the ten cancer types in terms of the number of incidences in that stage and differences in survivability rates from other stages. Stage-specific models were found to perform better than joint models with more number of incidences in the stage and with a higher difference in survivability rate from other stages.

We note, however, that there are a few cases in our results when the joint model statistically significantly outperforms the stage-specific model, namely for the distant stage of bladder (logistic regression), esophagus (logistic regression and decision tree) and liver (all methods), and the regional stage of liver (logistic regression and decision tree). We note that although joint model also does better on the localized stage of prostate cancer for naive Bayes, the actual difference of 0.002 is negligible. Looking further into these cases we find that all of them had relatively few incidences of those stages and in addition they had survivability rates similar to other stages (all on the low side). Hence in these cases when there were insufficient number of training examples to learn from, adding examples of other stages that do not have very different survivability rates was found to benefit. We depict this in Figure 2 in which we show all the cancer types in a plot between their number of incidences in the distant stage and the difference in survivability rate of the distant stage from other stages. The latter was obtained by subtracting survivability rate of incidences in the distant stage from the survivability rate of all incidences not in distant stage. We find that liver and esophagus cancers for which we found the exception on the distant stage appear on the bottom left side in the graph. This is the area representing fewer available training examples for the distant stage and similar survivability rates with other stages. Note that bladder cancer, for which the exception was found for only one method, also appears at the bottom but more in the middle. It is interesting to note that even though distant stage of lung cancer had survivability rate not far from other stages (it is on the left side in the graph), its joint model does not do better. This is clearly because it already had sufficient number of training examples of the distant stage for training the stage-specific model (it is on the top in the graph).
Figure 3: Learning curves for the logistic regression stage-specific models for the distant stage on a few cancer types.

We further look into this in Figure 3 by plotting learning curves for the stage-specific models for logistic regression. It is clear that the learning curve of lung cancer has almost plateaued with around 40% training examples and hence it has little scope of improvement. However, the learning curves of liver and esophagus cancers are yet to plateau indicating that their performance can still be improved with more examples. But on the other hand, for cervix uteri and corpus uteri cancers, although there are not many training examples of the distant stage (they are near the bottom in Figure 2) and their learning curves have not yet plateaued but they still do not benefit by adding examples of other stages (i.e. joint model does worse than stage-specific model). This is because for these cancers the survivability rate, and hence the data distribution, of the distant stage is very different from the other stages (they are on the right side in Figure 2). Please note that besides survivability rates, it is possible that feature values are also differently distributed in different cancer stages.

We also note that compared to other stages, performance was generally better on localized stage for which there were typically more number of examples. The results thus also indicate that performance on other stages could be improved with more training data. Given that our results have shown that including examples of other stages during training is only rarely helpful, a possibility for future work is to leverage examples of other stages through transfer learning [23]. This could be particularly useful for cancer types with fewer number of incidences.

Conclusions

In this paper we built two kinds of survivability prediction models for ten different cancer types. The joint models were trained on all cancer stages together, as in the past work, while the stage-specific models were trained on each cancer stage separately. They were both evaluated on all stages together, and on each stage separately. We showed that evaluating survivability performance on all stages together, as used to be done in the past, is not appropriate because it gives an artificial gain in performance owing to the inherent low to high survivability rates among the cancer stages. Evaluating models separately on each stage gives the realistic measure of survivability prediction performance. We also showed that stage-specific models perform better than joint models indicating that cancer stages are sufficiently different from each other that it is best to train survivability prediction models separately for each stage. With a few exceptions on cancer types with insufficient number of training examples and low survivability across stages, our results generalize to all cancer types and all cancer stages. Future work can
investigate the use of transfer learning between cancer stages to improve performance when faced with insufficient number of training examples.

Acknowledgments

We thank National Cancer Institute’s Surveillance, Epidemiology, and End Results Program for the dataset used in this work.

References


An Open Source Tool for Game Theoretic Health Data De-Identification

Fabian Prasser, PhD¹, James Gaupp, MS², Zhiyu Wan², Weiyi Xia, PhD², Yevgeniy Vorobeychik, PhD², Murat Kantarcioglu, PhD³, Klaus Kuhn, MD¹, and Brad Malin, PhD²
¹Technical University of Munich, Munich, Germany; ²Vanderbilt University, Nashville, Tennessee, USA; ³University of Texas at Dallas, Richardson, Texas, USA

Abstract

Biomedical data continues to grow in quantity and quality, creating new opportunities for research and data-driven applications. To realize these activities at scale, data must be shared beyond its initial point of collection. To maintain privacy, healthcare organizations often de-identify data, but they assume worst-case adversaries, inducing high levels of data corruption. Recently, game theory has been proposed to account for the incentives of data publishers and recipients (who attempt to re-identify patients), but this perspective has been more hypothetical than practical. In this paper, we report on a new game theoretic data publication strategy and its integration into the open source software ARX. We evaluate our implementation with an analysis on the relationship between data transformation, utility, and efficiency for over 30,000 demographic records drawn from the U.S. Census Bureau. The results indicate that our implementation is scalable and can be combined with various data privacy risk and quality measures.

Introduction

The healthcare community is increasingly driven by programs that collect and process large quantities of patient-specific data¹,². At the same time, these programs make use of highly detailed aspects about a patient and their daily activities within, as well as outside of, traditional clinical environments. To ensure that research studies are conducted, and healthcare applications are managed, at scale, it is critical to share data beyond the confines of where it was initially collected³. The push to make data accessible is exacerbated by initiatives that aim to enhance transparency in activities, reproducibility of research findings, and reuse of data for novel investigations. Such initiatives are being driven by federal agencies in the United States (U.S.), such as through the various data sharing policies of the National Institutes of Health⁴ and the National Science Foundation⁵, as well as in Europe, such as through Policy 0070 of the European Medicines Agency (EMA) for the public dissemination of clinical trials data⁶. However, as the quantity and quality of biomedical data grows, so too does its attractiveness to would-be attackers. There is, for instance, ample evidence to show that data breaches for healthcare organizations have grown substantially over the past several years⁷.

Privacy is a complex concept with ethical, legal and societal aspects⁸ and, consequently, it requires a mix of social (e.g., consent and trust) and technical constructs to realize⁹. In this work, we confine ourselves solely in the technical space, where a patient’s privacy is typically supported through a process referred to as de-identification in the U.S. (the convention we use in this paper) and anonymization elsewhere. De-identification is often achieved through the amendment of data that pertains to an individual’s identity. Such amendments may be realized through randomization, generalization to less specific terms, or suppression (i.e., redaction) of factors that can be leveraged to ascertain an individual’s identity (e.g., demographic factors). Specific rules for doing so have been codified in laws and regulations¹⁰,¹¹.

In parallel with these developments, a variety of studies have shown that de-identified health data can be re-identified (i.e., linked to identified individuals)¹². This has sparked a debate in which it has been argued that de-identification fails to provide adequate privacy protection¹³. However, it is important to realize that re-identification studies tend to focus on how attacks can be carried out, not on the likelihood they will be realized. As such, they demonstrate what is possible and not necessarily what is probable¹⁴. It has long been understood that de-identification involves trading privacy risks off against data quality and that privacy risks can never be entirely eliminated⁶.

Most recently, it was suggested that de-identification frameworks should formalize the capabilities of adversaries¹⁵. This shifts the view to an economic perspective, where the publishers gain benefit from sharing data at a certain level of fidelity, while attackers (i.e., the recipients) gain benefit from re-identifying the data. By modeling re-identification as a Stackelberg (or what is also called a leader-follower) game¹⁶ between two players, it was shown that attacks initially
thought to be quite detrimental to the management of biomedical research infrastructure may actually not be likely to occur in practice. Moreover, it was shown that the publisher’s gain can be maximized by de-identifying data under the assumption that the adversary will only attempt re-identification if there is a tangible economic benefit\textsuperscript{16}. While this approach is a notable advance in the field\textsuperscript{17}, it has not been made accessible through an easy-to-use software system.

**Objective and Contributions**

Traditionally, de-identification has been understood to be an optimization problem, in which data is transformed while an increase in privacy protection is traded off against a decrease in data quality. To balance both aspects, models for quantifying them are needed. Typically, a threshold is defined for privacy risks, which reduces the task to a simpler optimization problem, where the objective is to ensure that risk thresholds are met while data quality is maximized. At first glance, the game theoretic model appears to differ from this perspective. This paper reports on an in-depth analysis to identify integration options, as well as how we adapted and transitioned the game theoretic approach into scalable software. We extended ARX, an open source de-identification tool\textsuperscript{18}, with methods for de-identifying demographic and clinical data under the Stackelberg setting.

The specific contributions of this paper are as follows. First, we present an analysis of the game theoretic approach. Our findings show that (1) a specific variant of the game can be implemented with traditional privacy models and (2) the model can be interpreted in a manner that is consistent with the traditional perspective (i.e., as a combination of a privacy model and a data quality model). This interpretation forms the basis of our integration. Second, we further developed a search space pruning strategy that is specific to the game theoretic model and improves the scalability of the implementation significantly. Third, we describe how we integrated the game theoretic model into ARX (version 3.5.0), such that the implementation is fully compatible with all other de-identification methods implemented by the software. This is notable because it implies that the game theoretic approach can be combined, as well as compared, with other models for quantifying health data privacy risks and quality.

Beyond a software implementation, we report on an extensive experimental evaluation with over 30,000 demographic records from the U.S. Census. The results indicate that our highly scalable implementation often provides an overall payout (a measure that combines risk and utility) to the publisher that is comparable to what can be achieved using more sophisticated, though computationally costly, routines. Finally, we conclude this paper by reporting on ways to further extend our implementation towards the handling of high-dimensional (e.g., genomic) data.

**Background and Related Work**

**Data De-identification**

The Privacy Rule of the Health Insurance Portability and Accountability Act (HIPAA)\textsuperscript{10} provides specific guidance for the de-identification of health data via the *Safe Harbor* method. This method specifies 18 rules for modifying explicit (e.g., patient names and Social Security Numbers) and quasi-identifying (e.g., dates, ZIP codes with fidelity above 3 digits) factors that constitute sufficiently high re-identification risk. As an alternative, the HIPAA Privacy Rule permits the use of formal risk assessments under its *Expert Determination* methodology\textsuperscript{10}. This approach designates health data as de-identified when it is shown that the risk an anticipated recipient of the data could uniquely identify an individual is small. The method is notable as it allows for the application of an explicit adversarial model (i.e., anticipated recipient) and can explicitly manage risk through a quantified mechanism. Moreover, the approach is similar to how privacy is maintained in other countries. For example, the upcoming European General Data Protection Regulation (GDPR) states that "account should be taken of all the means reasonably likely to be used [for re-identification]"\textsuperscript{11}.

**Data De-identification Software**

There are various mature software solutions for facilitating data privacy risk assessments and de-identification, which have proven themselves in practice. μ-ARGUS and sdcMicro are open-source tools developed in the context of official statistics\textsuperscript{19}. By contrast, Privacy Analytics Eclipse is a commercial big data platform for risk-based de-identification of structured health records\textsuperscript{20}. In this paper, we focus on ARX, an open source tool that was also specifically designed for de-identifying biomedical data\textsuperscript{18}. ARX has been under constant development since 2011 and has found notable
adoption due to its comprehensive feature set and its easy-to-use graphical user interface. For example, it has been recommended by the EMA for implementing Policy 0070. ARX supports a wide variety of models for quantifying privacy risks and data quality as well as multiple models for transforming data.

### Data Transformation

Formal de-identification is typically applied to data derived from the healthcare setting by reducing the distinguishability of patient-level records. There are varying approaches for doing so, such as the injection of noise to adhere to emerging privacy models like differential privacy. However, the most popular approach in this domain remains reducing the fidelity of attributes that are likely to be exploited in re-identification attacks through linkage to named individuals in some external resource. This process renders records of a dataset to be less distinguishable for an attacker, thus reducing the certainty in executing a re-identification attack.

A generalization hierarchy is often relied upon for managing the valid transformations that can be applied to patient-level values. Two examples are shown in Figure 1. Here, values of an age attribute are transformed to different levels. To prevent overgeneralization, the former approach is often combined with subsequent record-level suppression. In this process, outliers (e.g., records with a high re-identification risk) are dropped from the dataset. The set of all possible combinations of levels with decreasing precision over increasing levels of generalization. Note that assigning a value to level 0 of its hierarchy leaves the value unchanged. In ARX, generalization hierarchies can be composed for categorical and continuous variables. In the latter case, this is accomplished by specifying functions for performing on-the-fly categorization of the space (e.g., creating an arbitrary grouping of heights or weights).

![Figure 1: Example generalization hierarchies for patient demographics that are vulnerable to re-identification attacks.](image)

While generalization can be realized in a variety of ways, in this article, we focus on two in particular. In full-domain generalization, all values of an attribute are transformed to the same generalization level in all records. By contrast, in record-level generalization, the values in each record can be generalized to different levels. To prevent overgeneralization, the former approach is often combined with subsequent record-level suppression. In this process, outliers (e.g., records with a high re-identification risk) are dropped from the dataset. The set of all possible combinations of generalization levels for all attributes forms a generalization lattice, where each element is called a de-identification policy. The generalization lattice for the example hierarchies from Figure 1 is shown in Figure 2. The latter figure also depicts the results of applying two de-identification policies for full-domain generalization, followed by subsequent record suppression. The payout indicated in the figure will be used as an example throughout this paper.

![Figure 2: A generalization lattice, de-identification policies, and output datasets for full-domain generalization.](image)

### Models for Quantifying Data Quality and Data Protection

Data transformation can influence the specificity of the data, so it is critical to balance an increase in protection with a decrease in data quality. To do so, we need to define formal measures for both criteria.

As alluded to, there are many models for measuring data protection, but in this work we focus on re-identification risk, because this is the primary concern addressed by current law and regulation. The re-identification risk of a record is typically estimated by calculating the inverse of the size of the group of indistinguishable records to which it
belongs. The universe of records that needs to be considered when forming groups depends on assumptions about the adversary’s prior knowledge.

First, in the Prosecutor model, it is assumed that the adversary already knows a record for a targeted individual is contained in the dataset. As such, groups must be formed using all records from the output dataset. This is the risk model underlying the $k$-anonymity privacy model. Second, in the Journalist model, it is assumed the adversary has no prior knowledge about membership. Thus, groups can be formed using a population table (i.e., a dataset containing records about all individuals from the population) generalized in the same way as the given record. This is the approach underlying the $k$-map privacy model. If no population table is available, a conservative estimate of the size of a record’s population group can be derived from the given dataset generalized in the same way as the record.

There are also various models for quantifying data quality. In this article we focus on the entropy-based model proposed by Wan et al. For a given record $r$, the information loss function $IL(r)$ returns a number in range $[0, 1]$ where a value of 0 indicates that the record has been preserved in its original state and a value of 1 indicates that all data has been suppressed.

The Game Theoretic Model

The game theoretic approach uses models for privacy protection and data quality to trade both aspects off against each other. For this purpose, it incorporates four intuitive parameters to construct a Stackelberg game in which the players, the data publisher (or defender) and the recipient (or attacker), are motivated by monetary incentives:

- **Adversary Gain**: The benefit that the adversary gains for a successfully re-identified record.
- **Adversary Cost**: The adversary’s cost to launch a re-identification attack against one record.
- **Publisher Benefit**: The benefit that the publisher receives by sharing a record in its original form.
- **Publisher Loss**: The publisher’s loss for one record due to successful re-identification.

By applying a model for quantifying re-identification risks to estimate the attacker’s success probability $SP(r)$ when attacking a record $r$, a monetary cost-benefit analysis can be performed. Our implementation supports all models described in the previous section.

The adversary attacks a record $r$ when the expected payoff is positive. More formally:

$$\text{Attack}(r) = \begin{cases} 1, & \text{if } SP(r) \cdot Gain > Cost \\ 0, & \text{otherwise} \end{cases}$$

As noted earlier, the value of a published record decreases when the amount of information content decreases. Thus, the publisher’s payoff for publishing a record $r$ can therefore be defined as:

$$\text{Benefit}(r) = (1 - IL(r)) \cdot \text{Benefit}$$

Additionally, when a record is successfully attacked, the publisher loses money. The expected loss per record is relative to the success probability of the adversary. Consequently, the publisher’s expected overall payout for a record $r$ is:

$$\text{Payout}(r) = \text{Benefit}(r) - \text{Attack}(r) \cdot SP(r) \cdot \text{Loss}$$

Methods

System Design

The de-identification methods supported by ARX are centered on generic data transformation algorithms that invoke full-domain generalization followed by record suppression. The algorithms rely on a user-specified 1) privacy model to determine which records need to be suppressed and 2) data quality model for optimizing the output.

In essence, ARX iterates over all available de-identification policies. For each policy, it follows a three step process (ignoring the various implemented optimizations): 1) **Transform** all records according to the policy, 2) **Suppress** records as indicated by the privacy model, and 3) **Assess** data quality using the quality model. Once all of the policies have been processed, ARX selects the policy with the best data quality to de-identify the dataset.
Implementation of the Basic Game

Prior work in game theoretic de-identification pursued a record-level perspective by transforming each record in such a way that the resulting payout is maximized. However, the implementation of the model in ARX relies upon full-domain generalization, as this approach has been shown to be more desirable to the biostatistics community. Later on, we will show that this basic implementation can be used to implement the record-level method.

To integrate the game theoretic model with ARX, we defined a privacy model and a quality model. The privacy model, called Profitability, is defined as follows:

\[
\text{Profitable}(r) = \begin{cases} 
1 & \text{if } \text{Payout}(r) >= 0 \\
0 & \text{otherwise}
\end{cases}
\]  

(4)

ARX uses this model to decide which records should be suppressed. We note that for any given dataset \( D \), suppressing all records \( r \in D \) with \( \text{Profitable}(r) = 0 \) is the optimal strategy for record-level suppression. This is because the decision of whether or not a record should be suppressed can be made independently for each record. Specifically, suppressing a record will not affect the information loss measured by the function \( IL \) for any other record. Moreover, when using a population table or the input dataset to calculate the adversary’s success probability \( SP \), removing a record from the output will also not affect the value returned by the function \( SP \) for other records. When using the output dataset to evaluate \( SP \), suppressing a nonprofitable record will only increase the value of \( SP \) for records that are in the same group. However, the group only contains nonprofitable records and increasing the value returned by \( SP \) makes publishing them even less profitable. As a consequence, suppressing a record does not affect whether publishing other records is profitable to the publisher. Obviously, the optimal strategy is to keep all profitable records and to remove all nonprofitable records.

The quality model is an objective function that measures the overall payout of the publisher for the complete dataset \( D \). Suppressed records will not be published, so we neither gain nor lose any money from them:

\[
\text{Payout}(D) = \sum_{r \in D} \text{Payout}(r) \cdot \text{Profitable}(r)
\]  

(5)

By parameterizing ARX with these models, we can solve the game using full-domain attribute generalization followed by record-level suppression. The process is sketched in Figure 3. As can be seen, ARX will suppress all records for which publishing will result in a negative payoff for the publisher. The overall payout for a (potentially generalized) dataset is defined as the sum of the payoffs for each record. The payoff for a suppressed record is defined as zero.

For each policy \( p \)
  - For each record
    - Apply \( p \)
  - For each record
    - Calculate payout
      - Suppress if payout < 0
    - Calculate payout
      (Suppressed: payout = 0)
  - Store optimum

Transformation phase
Payout optimization phase
Payout assessment phase

Figure 3: Basic algorithm executed by ARX for solving the game.

The output of this algorithm is globally optimal (in terms of overall publisher payout) regarding all possible transformations that rely on a combination of full-domain generalization and record suppression. This is because the algorithm implements an exhaustive search; i.e., it assesses every de-identification policy. For each policy, it first generalizes the dataset, then performs record suppression and stores the resulting payout. Since there is only a single optimal record
suppression strategy for any dataset, the algorithm has checked all potentially optimal outputs and the solution must therefore be globally optimal.

We also implemented a pruning strategy that eliminates solution candidates by reducing the search space of potentially optimal policies. To do so, we take advantage of the fact that the entropy-based information loss model used by the game theoretic approach is monotonic over paths in the generalization lattice. This means that, when data is only transformed with full-domain generalization, information loss will increase monotonically on each path from the bottom node of the generalization lattice to the top node. We note that this is not the case when data is transformed with full-domain generalization followed by record suppression\textsuperscript{26}. However, the monotonicity of the quality model can still be invoked to construct a pruning strategy.

For this purpose, we first calculate an upper bound on the maximal payout that can be obtained for a dataset $D$ by assuming that the adversary never attacks (i.e., that $\forall r, SP(r) = 0$):

$$\text{MaximalPayout}(D) = \sum_{r \in D} \text{Benefit}(r) = \sum_{r \in D} (1 - \text{IL}(r)) \cdot \text{Benefit}$$  \hspace{1cm} (6)

Due to the fact that $\text{MaximalPayout}(D)$ only depends on $\text{IL}(r)$, it follows that it also decreases monotonically within the lattice when data is only transformed with full-domain generalization. We can exploit this to prune portions of the search space by implementing the following strategy.

While traversing the solution space, we always keep track of the $\text{OptimalPayout}$, i.e. the payout of the best solution found so far. We note that $\text{OptimalPayout}$ represents the highest payout obtained using full-domain generalization and record suppression. When processing a new policy $p$, we retrieve a dataset $D$ as a result of the transformation phase. Next, we check whether $\text{MaximalPayout}(D) \leq \text{OptimalPayout}$. If this is the case, $p$ and all of its successors can be pruned, because the maximal payout that can be obtained by using them is already lower than the current optimum. Our implementation uses the generic framework provided by ARX for implementing such pruning strategies\textsuperscript{26}.

For context, Figure 2 provides an example. Let us assume that the algorithm first processes the policy $(1, 0)$ by generalizing the data and then suppressing one record. The resulting payout is the current optimum, so $\text{OptimalPayout} = \$1000$. Next, the algorithm processes policy $(0, 1)$. In doing so, it first generalizes to retrieve a dataset $D'$ and finds that $\text{MaximalPayout}(D') = \$800$. Policy $(0, 1)$ and all of its successors (i.e. $(1, 1), (2, 1), (3, 1)$) can now be pruned because none can result in output with a higher payout than the current optimum ($\$800 \leq \$1000$).

**Implementation of Variants of the Game**

There are several variants of the game theoretic model that have been proposed\textsuperscript{16} that we chose to integrate into ARX. The first is the $\text{SH-Friendly}$ variant, which guarantees that the degree of protection from re-identification is no lower than when using the HIPAA Safe Harbor policy. This can be achieved with ARX by using generalization hierarchies that reflect the Safe Harbor policy and by defining appropriate minimal levels of generalization.

The second is the $\text{No-Attack}$ variant of the game, which guarantees that the adversary has no incentives to ever attack. This variant can be implemented in an optimized manner, as can be seen as follows. Let $k$ be the size of the group of indistinguishable records used for calculating $SP(r)$. As described above, the adversary will only attack a record if:

$$SP(r) \cdot \text{Gain} > \text{Cost}. \hspace{1cm} (7)$$

So, it follows that:

$$SP(r) > \frac{\text{Cost}}{\text{Gain}} \Rightarrow \frac{1}{k} > \frac{\text{Cost}}{\text{Gain}} \Rightarrow k > \frac{\text{Gain}}{\text{Cost}}. \hspace{1cm} (8)$$

This indicates that the $\text{No-Attack}$ variant of the game implies a threshold on $k$. Depending on how the game has been configured to quantify re-identification risks (see Section Models for Quantifying Data Quality and Data Protection) this can be implemented with $k$-anonymity\textsuperscript{23} or $k$-map\textsuperscript{24}. As an example, given the realistic parameters derived by Wan et al. (i.e. $\text{Cost} = \$4$ and $\text{Gain} = \$300$\textsuperscript{16}), we obtain $k > \frac{300}{4} = 75$.  

1418
In addition to enforcing a threshold on $k$, we also need to maximize the publisher’s payout. Since we know that the adversary will never attack, Equation 3 can be simplified as follows:

$$Payout(r) = Benefit(r) - \underset{=0}{\text{Attack}}(r) \cdot SP(r) \cdot \text{Loss} = Benefit(r) = (1 - IL(r)) \cdot Benefit$$

(9)

These specifics of the No-Attack variant are important from an implementation perspective. They show that it is sufficient to minimize information loss in order to maximize publisher payout and that payout is monotonic in the lattice when only full-domain generalization is being used\textsuperscript{23,24}. This can be used to speed up record-level generalization.

**Implementation of Record-Level Generalization**

Our implementation of the game theoretic model is fully integrated into ARX. Figure 4 depicts two examples of how the API can be invoked. These illustrate how to de-identify an example dataset $D$ using the game theoretic approach with different models for transforming data and quantifying re-identification risks.

![Algorithm 1](image1)

**Algorithm 1** Full-domain generalization + record suppression

```java
arx ← new ARX();
arx.setPrivacyModel(new Profitability());
arx.setQualityModel(new PublisherPayout());
output ← arx.process(D);
```

![Algorithm 2](image2)

**Algorithm 2** Record-level generalization

```java
for (0 ≤ i < |D|) {
arx ← new ARX();
arx.setPrivacyModel(new Profitability());
arx.setQualityModel(new PublisherPayout());
arx.setSample(i);
output ← output ∪ arx.process(D);
}
```

**Figure 4**: Implementing different transformation models using the ARX API.

The first example shows pseudocode for invoking the API using full-domain generalization followed by record suppression and the prosecutor risk model. It should be noted that ARX can also be configured to use a population table. To do so, the dataset to be de-identified needs to be defined as a sample of the population table. With this configuration, our implementation uses the journalist risk model.

In the second example, we invoke the method without an explicit population table to estimate journalist risks as described previously. For this purpose, we de-identify the dataset $|D|$ times, using a sample of exactly one record of $D$ in each execution. The result for $D$ is simply the union of the individual outputs for each record.

**Experimental Design**

**Materials.** We evaluate our methods on the extract of 32,561 records from the 1994 U.S. Census and the generalization hierarchies used by Wan et al.\textsuperscript{16}. The dataset was enriched with demographics extracted from additional U.S. Census data for the state of Tennessee to enable comparisons with HIPAA Safe Harbor. We used traditional demographics in the form of 1) age, 2) sex, 3) ZIP code, and 4) race as quasi-identifying variables given their known contribution to re-identification attacks.

**Parameterization.** As a base case, we used the setup from Wan et al.\textsuperscript{16}: Adversary cost = $4, based on the costs for obtaining detailed information about individuals online. Publisher benefit = $1,200, based on an analysis of grant funding received divided by the number of records published by the five member institutions of the Electronic Medical Records and Genomics (EMERGE) network, an NIH consortium. Publisher loss = $300, based on an analysis of HIPAA breach violation cases reported by the U.S. Department of Health and Human Services.

In our experiments, we rely on the default parameterization for adversary cost and publisher benefit, but we further defined adversary gain to be equal to publisher loss. By varying this single parameter between $0$ and $2,000$ we were able to investigate a wide variety of scenarios, in which either the data publisher or the adversary is at an advantage.

**Performance Measures.** When reporting results, we express the overall publisher payout relative to the theoretical maximum, which can be obtained when the data will never be attacked. In this case, the publisher payout equals (number of records) * (publisher benefit). This results in a maximal payout of $32,561 \times 1,200 = 39.0732$ million for our setup. We compared our approach to prior work and performed a runtime analysis of our implementation.
Results

Comparison with Prior Work

Wan et al. solved the game with record-level generalization using journalist risks calculated with a population table. As mentioned, we developed methods to solve the game in the context record-level generalization and full-domain generalization with and without supplying a population table. Here, we present results of a comparison.

As a baseline, we used HIPAA Safe Harbor, which we applied by replacing all values of the attribute ZIP code with their initial 3 digits, masking zip codes with populations of less than 20,000 individuals and top-coding ages to 90 and above. The results of the experiment are shown in Figure 5. The first vertical line in each plot denotes the parameterization $GL = Gain = Loss = $300, which has been shown to be a realistic scenario. The second vertical line indicates the point at which the data sharing scenario becomes disadvantageous for the publisher.

![Figure 5: Publisher payout for the output of different implementations calculated using a population table.](image)

The results obtained for the basic game are shown in Figure 5 (a) and (b). There are a several notable findings to highlight. First, all of the methods provided a higher payout than Safe Harbor when the publisher was at an advantage. Second, it can be seen that, up to $GL = $400, full-domain generalization with risks calculated from the dataset resulted in a publisher payout that is close to that achieved using record-level generalization and a population table. This notable because for $GL = $300 the method by Wan et al. achieved only 0.17% more relative payout. Third, it can be seen that, when using a population table, full-domain generalization generally resulted in a payout similar to that obtained using record-level generalization. Even for $GL = $2000, the method by Wan et al. provided only 1.9% more relative payout. Finally, it can be observed that when calculating risks from the dataset, full-domain generalization resulted in a greater payout than record-level generalization up to about $GL = $600. This is because record-level generalization is much more flexible and, therefore, leads to a pronounced overgeneralization when risks are overestimated significantly.

Figures 5 (c) and (d) present the results obtained for the No-Attack variant of the game. It can be seen that this model can not be implemented effectively without supplying a population table. This is an intuitive finding as we have shown that the model implies a significant risk threshold (i.e., up to $\frac{1}{500} = 0.2\%$ for $GL = $2000). Still, it is notable that using full-domain generalization and a population table resulted in higher payout than Safe Harbor for $GL \leq $800.

Analysis of Scalability

We further investigated the scalability of our implementation in ARX. To do so, we (1) measured the impact of the pruning strategy, (2) compared execution times for full-domain generalization with record-level generalization and (3) measured the speed up achieved by implementing the No-Attack variant with $k$-anonymity or $k$-map. The experiments were performed on an Intel Core i5 3.1 GHz machine running a 64-bit Linux 3.2.0 kernel and a 64-bit JVM (1.7.0). Our method for record-level generalization using a population table is not included in the current release of ARX. Using full-domain generalization and a population table took around 2 seconds in the experiments described above. The results shown in Figure 6 focus on methods that calculate risks from the dataset.

Figure 6 (a) depicts the impact of the pruning strategy. It can be seen that the optimization was more effective in scenarios that were advantageous to the publisher. For example, we achieved a 4x speedup for $GL = $300 and only a
Figure 6: Execution times for solving the game without using a population table.

2x speedup for \( GL = $2,000 \). The main reason is that, when the scenario is less advantageous to the publisher, more generalization is required to obtain a good solution to the de-identification problem. This makes the optimization less effective and, consequently, increases execution times.

A comparison of execution times measured for full-domain generalization and for record-level generalization is shown in Figure 6 (b). It can be observed that full-domain generalization is four orders of magnitude more efficient than record-level generalization. This is reasonable, as the latter was implemented by executing full-domain generalization for every record in the dataset (i.e., 32,561 times).

Finally, Figure 6 (c) summarizes a comparison of the execution times measured for a naïve implementation of the No-Attack variant of the game and the optimized implementation proposed previously. It can be seen that implementing the model using traditional privacy models and de-identification algorithms reduced execution times by up to 67%. We measured more pronounced improvements for record-level generalization, which is significantly more computationally complex than full-domain generalization.

Discussion and Conclusions

In this paper, we introduced new variations of the game theoretic approach to health data de-identification and described their integration into the open source software ARX. In contrast to many other methods of data de-identification, the game theoretic approach explicitly accounts for the incentives of data publishers and recipients. The experiments showed that our implementation provides a novel means for balancing publisher payout with execution times by using different models for transforming data. Our approach does not rely upon the existence of a population table, but users are free to specify one when it is available.

There are a wide variety of directions for further improving our implementation. In particular we believe our approach can be extended to support higher-dimensional data, such as genomic summary statistics, akin to the approach recently proposed for the minor allele frequencies of single nucleotide polymorphisms (SNPs). This is a natural extension because ARX is based on a modular design with a focus on scalability. The tool is already able to load very large and high-dimensional data, but two additional modules will need be developed to fully support the processing of such data. First, the existing implementation of ARX’s generalization lattice needs to be complemented with an implementation that represents attribute suppression policies as bit vectors. Second, the search algorithms currently implemented by the tool need to be complemented with a genetic algorithm developed by Wan et al. This algorithm is generic in design (that is, it is not specific unto genomic data) and focuses solely on attribute suppression (i.e., either an attribute is retained or redacted). Since ARX supports attribute-level suppression, integration into the tool should be a feasible endeavour.

Acknowledgements

This work was sponsored, in part, from grants R01HG006844, R01LM009989, RM1HG009034, U01HG006385, and U01HG006378 from the National Institutes of Health.
References


Leveraging Health Information Exchange to Construct a Registry for Traumatic Brain Injury, Spinal Cord Injury and Stroke in Indiana

Saurabh Rahurkar, BDS, DrPH¹, Timothy D. McFarlane, MPH², Jane Wang, PhD¹, Sarah Hoover¹, Flora Hammond, MD³, Jacob Kean, PhD⁴, Brian E. Dixon, MPA, PhD¹,²,⁵
¹ Center for Biomedical Informatics, Regenstrief Institute, Indianapolis, IN, USA; ² Richard M. Fairbanks School of Public Health, Indiana University, Indianapolis, IN, USA; ³ Indiana University School of Medicine, Indianapolis, IN, USA; ⁴ Population Health Sciences, Univ. of Utah School of Medicine, Salt Lake City, UT, USA; ⁵ Center for Health Information and Communication, Department of Veterans Affairs, Veterans Health Administration, Health Services Research and Development Service CIN 13-416, Indianapolis, IN, USA

Abstract

Traumatic brain injury (TBI), spinal cord injury (SCI) and stroke are conditions of interest to public health as they can result in long-term outcomes and disabilities. Specialized registries can facilitate public health surveillance, however only 4% of hospitals in the United States actively engage in electronic reporting to these registries. We leveraged electronic claims and clinical data from a health information exchange to create a statewide TBI/SCI/Stroke registry to facilitate the study of long-term outcomes and health services utilization. The registry contains 109,943 TBI patients, 9,027 SCI patients and 117,084 stroke patients with a mean of 3 years of follow-up data after injury. Additionally, the registry contains data on individual patient encounters, prescriptions and clinical variables. The high-dimensional data with large sample sizes may present a valuable informatics resource for injury research as well as public health surveillance.

Introduction

Following the passage of the Health Information Technology for Economic and Clinical Health (HITECH) act in 2009, the landscape for health IT in the United States changed dramatically. The meaningful use (MU) initiative provisioned under HITECH can be attributed with the widespread adoption of electronic health records (EHRs) among both office-based physicians and hospitals in the US. While 78% of office-based physicians have adopted certified EHRs, 87% physicians in the US use some form of EHR¹. Adoption of EHRs among hospitals is even higher with 97% of all US hospitals using certified EHRs². Requirements in MU prioritize sharing of patient health information collected during the process of routine care between providers and hospitals to provide comprehensive patient health information at all points of care. Additionally, MU incentivizes the creation of specialized disease registries: registries for conditions other than cancer, to improve population and public health.

Traumatic brain injury (TBI) is one of the most prevalent and severely disabling neurological disorders in the US. Nationwide, 1.7 million people experience TBI annually and 3.2–5.3 million live with a TBI-related disability³–⁵. An estimated 53,000 people die every year due to TBI, representing 30.5% of all injury related deaths⁴,⁶. The economic burden of TBI in terms of total lifetime costs due to loss of productivity is estimated at $63.9 billion and an additional $13.1 billion in direct medical costs⁷,⁸. On the other hand, spinal cord injuries (SCIs) while rare, account for about 17,000 cases per year and an estimated range of 243,000 and 347,000 people lived with an SCI in 2016⁹. Incomplete paralysis of all four limbs is the most frequent outcome of SCI with less than 1% of patients experiencing complete neurological recovery by the time they are discharged from the hospital. Additionally, about 13% of all patients experience complete paralysis of all four limbs. Lifetime costs due to SCIs are between $1.5 million and $4.7 million for those with an SCI at 25 years of age. As such, SCIs in combination with TBIs greatly contribute to permanent disability, high medical costs and costs due to loss of productivity.

Although TBIs and SCIs can result in permanent disability, the leading cause of long-term disability as well as the leading preventable cause of disability in the US is stroke¹⁰. Approximately 795,000 people in the US have a stroke each year with more than 75% of these being first or new strokes. Stroke is also the fifth leading cause of death in the US. Racial disparities exist with risk of first stroke in blacks being twice that of whites; compared to whites, blacks are
also more likely to die from stroke. While not a form of TBI, stroke results in injury to the brain that may be permanent and result in impairments in cognition, strength, and swallowing that require specialized rehabilitation training, and/or long-term assistance due to disability. As such, the total cost of stroke due to health care utilization, medication and loss of productivity is estimated at $33 billion each year. Therefore, due to the profound impact these events have on individuals and generally the society; TBI, SCI and stroke are significant from a public health perspective.

TBI, SCI and stroke are chronic diseases based on the World Health Organization (WHO) definition as they have one or more of the following characteristics: permanent, caused by non-reversible pathological alterations; requires special training of the patient for rehabilitation; and/or may require a long period of observation, supervision, or care. Despite understanding the association of TBI and SCI with a host of long-term diseases and disorders, as well as the growing understanding of the causative and accelerative nature of these associations, TBI and SCI are too often seen as isolated events by the health care delivery system rather than as the onset of chronic disease. As a result, TBI and SCI are chronic diseases for which the health care system has only an acute plan of care. Finally, considering that the three conditions: TBI, SCI and stroke, can result in similar outcomes and disabilities following an acute injury; studying them together may be of interest.

To study the long-term outcomes and health services utilization among patients with TBI, SCI and stroke, the Indiana State Department of Health (ISDH) supported the creation of a specialized registry at the Regenstrief Institute. The goal was to create a virtual environment in which public health epidemiologists and researchers could collaboratively examine the TBI, SCI and stroke populations in Indiana. External support was necessary since typical data sources available to state health departments (e.g. Medicaid claims, trauma registries) contain limited details about patients immediately following their injury. In this paper, we describe the creation of the Indiana Registry for TBI, SCI and Stroke, detailing the data housed within the registry while discussing the lessons learned in creating a disease registry to support public health surveillance and research.

Methods

Indiana Network of Patient Care (INPC)

The Indiana Network for Patient Care (INPC) is the longest-tenured and most comprehensive health information exchange (HIE) network in the United States. The goal of creating the INPC was the timely availability of clinical information when and where it is needed. The network includes data from more than 100 hospitals; 12,000 practices; local and state health departments; local and national laboratories; a national pharmacy benefit manager consortium; long term post-acute care facilities; free standing radiology centers; several large-group practices closely tied to hospital systems; as well as payers including the state Medicaid program. The INPC data repository contains over 9 billion pieces of clinical data, including over 195 million text reports, for approximately 12 million unique patients. The INPC employs a robust master person index (MPI) that supports integration of patient data across 1,000 information systems (e.g., lab, radiology) and processes more than 350,000 messages each day.

Data extraction and preparation

In order to create the registry, we first identified patients who had an event of TBI, SCI or stroke between 2005 and 2014. The patients were identified using a set of codes (Table 1) from version 9, Clinical Modification of the International Classification of Diseases (ICD). The inclusion criteria were limited to the requirement that the patients have their first event or the “index” event of TBI/SCI/stroke during the specified time frame (2005–2014). This definition of TBI is broad and based on that which is used by public health agencies to ensure complete capture of all possible cases. Specific analyses by registry users could utilize more precise case definitions based on individual study needs.

For included patients, comprehensive and longitudinal de-identified health record data was extracted at the encounter level. Specifically, we extracted data on demographic (e.g., patient id, age, gender, race, rurality of residence), encounter (e.g., visit type, visit location, diagnosis at visit, length of stay), prescriptions (e.g., medication prescribed, dose, strength) and clinical (e.g., diagnostic test and test results, age at test, procedures conducted) characteristics from the INPC. The complete list of extracted data elements is presented in Table 2.
Table 1: ICD-9-CM Codes used for patient population definitions for TBI, SCI and stroke

<table>
<thead>
<tr>
<th>ICD code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>800.00 - 801.99</td>
<td>Fracture at the vault or base of the skull</td>
</tr>
<tr>
<td>803.00 - 804.99</td>
<td>Other and unqualified or multiple fractures of the skull</td>
</tr>
<tr>
<td>850.0 - 850.9</td>
<td>Concussion</td>
</tr>
<tr>
<td>851.00 - 854.19</td>
<td>Intracranial injury, including contusion, laceration, and hemorrhage</td>
</tr>
<tr>
<td>950.1 - 950.3</td>
<td>Injury to the optic chiasm, optic pathways, or visual cortex</td>
</tr>
<tr>
<td>959.01</td>
<td>Head injury, unspecified</td>
</tr>
<tr>
<td>995.55</td>
<td>Shaken Infant Syndrome</td>
</tr>
</tbody>
</table>

Spinal Cord Injury

<table>
<thead>
<tr>
<th>ICD code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>806.0 - 806.99</td>
<td>Closed fracture of cervical vertebra with spinal cord injury</td>
</tr>
<tr>
<td>952.0 - 952.99</td>
<td>Cervical spinal cord injury without evidence of spinal bone injury</td>
</tr>
</tbody>
</table>

Stroke

<table>
<thead>
<tr>
<th>ICD code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>433.01</td>
<td>Occlusion and stenosis of basilar artery with cerebral infarction</td>
</tr>
<tr>
<td>433.11</td>
<td>Occlusion and stenosis of carotid artery with cerebral infarction</td>
</tr>
<tr>
<td>433.21</td>
<td>Occlusion and stenosis of vertebral artery with cerebral infarction</td>
</tr>
<tr>
<td>433.31</td>
<td>Occlusion and stenosis of multiple and bilateral precerebral arteries with cerebral infarction</td>
</tr>
<tr>
<td>433.81</td>
<td>Occlusion and stenosis of other specified precerebral artery with cerebral infarction</td>
</tr>
<tr>
<td>433.91</td>
<td>Occlusion and stenosis of unspecified precerebral artery with cerebral infarction</td>
</tr>
<tr>
<td>434</td>
<td>Cerebral thrombosis without mention of cerebral infarction</td>
</tr>
<tr>
<td>434.01</td>
<td>Cerebral thrombosis with cerebral infarction</td>
</tr>
<tr>
<td>434.11</td>
<td>Cerebral embolism with cerebral infarction</td>
</tr>
<tr>
<td>434.91</td>
<td>Cerebral artery occlusion, unspecified with cerebral infarction</td>
</tr>
<tr>
<td>435.0 - 435.9</td>
<td>Transient cerebral ischemia</td>
</tr>
<tr>
<td>436</td>
<td>Acute, but ill-defined, cerebrovascular disease</td>
</tr>
</tbody>
</table>

Table 2: List of extracted variables

<table>
<thead>
<tr>
<th>Demographics</th>
<th>Encounter</th>
<th>Medication</th>
<th>Clinical</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient ID (de-identified)</td>
<td>Encounter date</td>
<td>Prescription date</td>
<td>ICD-9-CM procedure code</td>
</tr>
<tr>
<td>Age</td>
<td>Visit type</td>
<td>Drug name</td>
<td>Procedure/test date</td>
</tr>
<tr>
<td>Sex</td>
<td>Clinic name</td>
<td>Drug dose</td>
<td>CPT/HCPCS codes</td>
</tr>
<tr>
<td>Race</td>
<td>Diagnosis at visit</td>
<td>National Drug Code</td>
<td>Test results</td>
</tr>
<tr>
<td>Age at first event</td>
<td>Diagnosis priority</td>
<td>Number of days supply</td>
<td>Age at procedure</td>
</tr>
<tr>
<td>Death year (if deceased)</td>
<td>Age at visit</td>
<td>Dispensed amount</td>
<td>Age at procedure</td>
</tr>
<tr>
<td>Rurality of residence</td>
<td>Length of stay</td>
<td>Age at prescription</td>
<td></td>
</tr>
</tbody>
</table>

Encounter data was derived from admission, discharge, and transfer data as well as medical claims and discharge summary data captured by the INPC. Encounters in the INPC represent visits to a variety of care settings, including acute care hospitals, skilled nursing facilities, and ambulatory providers. Visits are coded by the sending facility and often include admission diagnosis as well as length of stay (for acute care visits).

Medication data was derived from dispensing events with medications identified using National Drug Code (NDC) numbers. The NDC numbers are 11-digit drug identifiers consisting of three segments which identify the labeler (e.g. manufacturer); the product, form and strength (e.g. Prozac, capsule 20 mg); and the packaging (e.g. 100 caps). Consequently, the same active ingredient could have multiple NDC numbers based on differences in the labeler, strength and packaging. Therefore, data from the Anatomical Therapeutic Chemical Classification System (ATC) controlled by the WHO was used to create a crosswalk mapping individual NDC codes to therapeutic drug classes (e.g. Antithrombotic agents, Diuretics). Considerable variation was also seen in the presentation of visit types, which were aggregated into broader categories using the Centers for Medicaid and Medicare Services (CMS) place of service codes and descriptions.

Additional clinical variables, or observations recorded in the EHR, were derived from various types of messages exchanged within the INPC between various clinical information systems. Test results from laboratory information systems, procedure codes from claims messages, problem lists in a variety of messages, and other observations recorded...
for patients who might be of interest for surveillance or research were extracted. These variables only contain discrete values as the registry cannot contain free-text documents that may contain identifiable data.

The Indiana Registry for TBI, SCI and Stroke

Whereas the INPC is an operational system that supports transactional HIE for patient care, the Indiana Registry for TBI, SCI and Stroke is a disease registry designed to support population health surveillance and research. Therefore, the transactional data from INPC were extracted, transformed, and loaded into a virtual server where they could be accessed independently from the transactional system. The transformation of the data included a step whereby medical record numbers were translated into study identifiers and other patient identifiers (e.g., phone number, address) were removed. The registry therefore contains a de-identified dataset with patient age, number of days following the index event, and other data that enable analysis without the possibility of re-identification. As depicted in Figure 1, the data from the INPC is loaded onto a virtual server running within the research environment at Indiana University. Researchers at IU and Regenstrief can access the data via applications (e.g., SQL Server, R, SAS) that run on other virtual machines within the IU environment. Similarly, external collaborators, including the staff at the ISDH Injury Prevention Division, can connect into the virtual environment to analyze the data for surveillance or research aims.

![Figure 1: Diagram depicting data integration from different sources to create the TBI/SCI/Stroke registry](image)

Data Analysis

The analysis focuses on a description of the demographic, encounter, medication and clinical data extracted for the registry along with descriptive statistics for the demographic and encounter data in the registry. First, we present an overview of the registry’s data, detailing the total observations for all type of data including the average follow-up time in years. Next, we determined the incidence rates for TBI, SCI and stroke between 2005 and 2014. Incidence rates were calculated using census estimates on the population of Indiana for each year as the denominator. Unique visits were calculated as each visit that a patient made on a new day. Thus, multiple outpatient visits by a patient on different days would count as separate and unique visits.
Table 3: Overview of the TBI/SCI/Stroke Registry

<table>
<thead>
<tr>
<th>Registry</th>
<th>Distinct patients</th>
<th>Number of observations</th>
<th>Mean years of follow-up</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Clinical variables</td>
<td>Encounter</td>
</tr>
<tr>
<td>TBI</td>
<td>109,943</td>
<td>63,527,763</td>
<td>36,473,161</td>
</tr>
<tr>
<td>SCI</td>
<td>9,927</td>
<td>7,149,726</td>
<td>5,333,407</td>
</tr>
<tr>
<td>Stroke</td>
<td>117,084</td>
<td>126,804,413</td>
<td>53,298,202</td>
</tr>
</tbody>
</table>

* Before or After the event

Results

The TBI/SCI/Stroke registry had 109,493, 9,927 and 117,084 patients who had their first event of TBI, SCI and stroke respectively, between 2005 and 2014 (Table 3). Total follow-up duration for patients was 6.1 years for TBI, 6.39 years for SCI and 5.86 years for stroke patients. Additionally, patients were followed for at least 2.86 years before or after any of the 3 events. The clinical variables table containing information such as procedure codes, test and test results etc. accounted for the majority of the health information in the registry with at least 7 million observations for SCI patients and over 126 million observations for stroke patients. Furthermore, observations ranged between 5 and 53 million in the encounter tables and between 1.3 and 17 million in the medication tables across the TBI/SCI/Stroke registry.

![Incidence of TBI, SCI and stroke in Indiana from 2005 – 2014](image)

The incidence of TBI, SCI and stroke for each year in the registry is shown in Figure 2. While the incidence for TBI and stroke shows a steady increase from 2005 to 2014, that of SCI remains relatively unchanged. Table 4 presents the demographic characteristics of the patients in the registry. Just under half the patients in the TBI (45.11%) and SCI (48.14%) registry were females, with the stroke registry having a higher proportion of female (58.72%) patients. Patients for TBI (75.33%), SCI (52.88%) and stroke (75.04%) were majority white, however there was variation in the availability of race data across the three conditions; whereas about 7% of the TBI and stroke patients lacked information on race, almost 34% SCI patients had no race related information. Majority of the patients in all three registries resided in an urban location.
Table 4: Demographic characteristics

<table>
<thead>
<tr>
<th>Registry</th>
<th>Male</th>
<th>Female</th>
<th>White</th>
<th>Black</th>
<th>Hispanic</th>
<th>Asian</th>
<th>NHPI</th>
<th>AIAN</th>
<th>Other</th>
<th>Unknown</th>
<th>Urban</th>
<th>Rural</th>
</tr>
</thead>
<tbody>
<tr>
<td>TBI (N=109,493)</td>
<td>60,106</td>
<td>49,387</td>
<td>82,485</td>
<td>13,373</td>
<td>3,593</td>
<td>346</td>
<td>252</td>
<td>59</td>
<td>1,927</td>
<td>7,458</td>
<td>79,503</td>
<td>20,135</td>
</tr>
<tr>
<td>SCI (N=9,927)</td>
<td>5,148</td>
<td>4,779</td>
<td>5,249</td>
<td>969</td>
<td>216</td>
<td>25</td>
<td>15</td>
<td>2</td>
<td>90</td>
<td>3,361</td>
<td>6,999</td>
<td>2,189</td>
</tr>
<tr>
<td>Stroke (N=117,084)</td>
<td>48,329</td>
<td>68,753</td>
<td>87,857</td>
<td>16,537</td>
<td>1,852</td>
<td>406</td>
<td>617</td>
<td>39</td>
<td>2,275</td>
<td>7,500</td>
<td>82,739</td>
<td>22,486</td>
</tr>
</tbody>
</table>

*Native Hawaiian or Pacific Islander
American Indian or Alaskan Native

The number of patients with TBI, SCI or stroke varied by age groups as well (Figure 3). TBI was more commonly seen in the younger age-groups with the greatest proportion of patients experiencing the first TBI event between the ages of 15 years and 25 years. On the other hand, stroke patients were generally older with more than half of all patients being 65 years or older at the time of first stroke event. SCI patients were generally younger than stroke patients at the time of first event with most patients having their first event between the ages 45 years and 65 years. Overall, SCI and stroke patients were older or middle-aged whereas patients with TBI were generally younger.

Figure 3: Proportion of total cases by age-group at first event

At the encounter level, the registry had 17,923,338; 2,569,718 and 26,381,982 unique visits in the TBI, SCI and stroke registry for all care settings (Table 5). After aggregating care setting types based on the CMS place of service order, 13 distinct care settings were identified. Of these outpatient care had the greatest number of visits in all three registries. While TBI (13.73%) and stroke (15.04%) had pharmacy as the next most common type of visit, SCI (16.90%) patients had home visits as the second most common type of visit. “Other facility” included facilities which could not be clearly identified or care settings that collectively made up less than 0.5% of the total unique visits.
Table 5: Total unique visits by care setting

<table>
<thead>
<tr>
<th>Care setting</th>
<th>TBI</th>
<th>SCI</th>
<th>Stroke</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ambulance</td>
<td>319,750 (1.78%)</td>
<td>80,399 (3.13%)</td>
<td>799,173 (3.03%)</td>
</tr>
<tr>
<td>Outpatient</td>
<td>6,098,059 (34.02%)</td>
<td>865,397 (33.68%)</td>
<td>7,369,826 (27.94%)</td>
</tr>
<tr>
<td>Inpatient</td>
<td>835,924 (4.66%)</td>
<td>189,775 (7.39%)</td>
<td>1,737,775 (6.59%)</td>
</tr>
<tr>
<td>Emergency Room</td>
<td>956,879 (5.34%)</td>
<td>81,934 (3.19%)</td>
<td>925,296 (3.51%)</td>
</tr>
<tr>
<td>Behavioral and Mental Health</td>
<td>1,364,753 (7.61%)</td>
<td>93,810 (3.65%)</td>
<td>1,136,398 (4.31%)</td>
</tr>
<tr>
<td>Elderly care (NF, SNF etc.)</td>
<td>1,092,879 (6.10%)</td>
<td>159,986 (6.23%)</td>
<td>3,056,923 (11.59%)</td>
</tr>
<tr>
<td>Home</td>
<td>1,516,718 (8.46%)</td>
<td>434,277 (16.90%)</td>
<td>3,036,571 (11.51%)</td>
</tr>
<tr>
<td>Lab</td>
<td>184,294 (1.03%)</td>
<td>21,210 (0.83%)</td>
<td>233,615 (0.89%)</td>
</tr>
<tr>
<td>Pharmacy</td>
<td>2,460,388 (13.73%)</td>
<td>298,184 (11.60%)</td>
<td>3,969,108 (15.04%)</td>
</tr>
<tr>
<td>Public and Community Health</td>
<td>55,062 (0.31%)</td>
<td>3,388 (0.13%)</td>
<td>39,319 (0.15%)</td>
</tr>
<tr>
<td>Inpatient rehab</td>
<td>17,130 (0.10%)</td>
<td>4,969 (0.19%)</td>
<td>25,293 (0.10%)</td>
</tr>
<tr>
<td>Outpatient rehab</td>
<td>8,454 (0.05%)</td>
<td>1,369 (0.05%)</td>
<td>8,453 (0.03%)</td>
</tr>
<tr>
<td>Other facility</td>
<td>3,013,048 (16.81%)</td>
<td>335,020 (13.04%)</td>
<td>4,044,232 (15.33%)</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>17,923,338</strong></td>
<td><strong>2,569,718</strong></td>
<td><strong>26,381,982</strong></td>
</tr>
</tbody>
</table>

Discussion

Using data routinely collected by EHRs during clinical care processes, we created a specialized registry for TBI, SCI and stroke patients. The registry contains large volumes of encounter level data providing details on patients following their index injury with an average of about 3 years worth of data before and after the event. Given the richness and longitudinality of the data, this registry may be a valuable informatics resource for long-term injury surveillance as well as research.

While not the first registry to focus on TBI, SCI or stroke, this registry is unique in that it was created using readily available data part of a HIE network from multiple health systems. In other words, there was no additional effort required on the part of patients or care providers to report cases to this registry. Although there are MU incentives for public health reporting to specialized registries, these criteria are optional in Stage 3 MU and will continue to remain so under the Medicare Access and CHIP Reauthorization Act (MACRA). Consequently, only about 4% of hospitals in the US were actively engaged in public health reporting to specialized registries as of 201516. Since our registry was created by leveraging existing HIE data and does not require providers actively reporting, our methods are likely to capture more cases than passive surveillance systems, therefore allowing for a larger patient population.

The large patient population in our registry presents a novel dataset for studying long-term outcomes and health services utilization among patients with TBI, SCI or stroke. The high-dimensional nature of the data allows for robust study designs to assess research questions focused on outcomes in these patients. A key strength of the registry lies in the availability of visits data, including the type of setting and the length of stay. Additionally, since the data is longitudinal at the encounter level it is possible to track individual patients through the care process. For instance, a 55 year old male patient living in an urban area has a TBI event on day 0 and is air-lifted to the emergency room where the ICD codes provide a diagnosis of “brain hemorrhage with loss of consciousness”. In the same visit, additional ICD codes mention further injuries in the form of multiple fractures and mechanism of injury codes that indicate the cause was a motor vehicle crash in which the patient was the driver. On day 1, the patient is admitted to the inpatient facility where the patient’s progress can be tracked using ICD codes until he is transferred to an inpatient rehab facility by ambulance on day 8. During the patient’s stay, ICD codes provide a snapshot of the aftercare provided to the patient at the inpatient rehab facility. On day 24, the patient is discharged to home. Data of this nature may be of immense value in examining the relationship between care processes and patient outcomes. In addition, it may also be possible to examine outcomes based on specific conditions within any one registry, e.g. comparing post-stroke outcomes among diabetics vs. non-diabetics. Given the availability of data both before and after the event, along with sufficiently large sample sizes; studies using this data could potentially generate high-quality, real-world evidence for policymakers as well as clinical care organizations.

The registry is further a unique informatics resource in that it allows for access by both public health epidemiologists
as well as researchers. Current information resources used by epidemiologists in governmental public health agencies are limited to data feeds around acute events. Trauma registries, which receive data from emergency departments, in-patient stays, and emergency medical services; capture details of an injury when it occurs along with prognosis at time of discharge and from short term rehab. Further, trauma registries depend on primary data collection and require dedicated staff for abstraction of patient record and other activities. Syndromic surveillance systems, which also receive data from emergency departments, capture data on patient-reported chief complaints and, in some states, admitting diagnoses\textsuperscript{17}. These resources, however, report de-identified data in a way that does not allow the health department to examine repeat visits for the same individual longitudinally. Nor do these resources provide information on follow up care after discharge. By linking individuals’ medical records and hosting the linked data in a de-identified yet protected environment, the registry facilitates collaboration between health services researchers and epidemiologists. This method may become a model for other diseases as well as replication in other states.

**Role of Health Information Exchange in Creation of the TBI, SCI, and Stroke Registry**

A commitment from Indiana’s community HIE, the INPC, to support public health and research played an instrumental role in the creation of the TBI, SCI, and stroke registry. The robust and highly connected nature of the INPC made it possible to gather data from discrete sources from across Indiana. As such widespread HIE adoption may be viewed as a precursor to creating a registry of this nature. In its current state, HIE adoption within individual states as well as nations is heterogeneous. While the business case for HIE adoption largely focuses on the benefits to clinical care, the literature evaluating the effect of HIE on health outcomes finds only mixed evidence for these benefits\textsuperscript{18}. Our work demonstrates the feasibility in creating a public health resource using HIE data and as such the potential public health benefits of HIEs should be explored further.

To our knowledge, this is the first specialized registry created for surveillance as well as research that uses data from a mature HIE. While other registries for TBI and stroke exist, these were created as a result of initiatives specifically aimed at creating epidemiological datasets and not as a positive externality of an existing resource. Similarly, while research datasets that focus on these conditions exist, enormous resources were expended for their conception. Whereas our registry was built using secondary data from existing clinical processes versus collection of primary data which involves significant resource expenditure and time. Although, it is true that the establishment of an HIE also presents a significant investment of resources, the ability to create registries and research datasets presents a desirable unintended consequence. As such considering these benefits in cost-benefit analysis may present a better understanding of “true” cost of HIE adoption.

**Limitations of EHR-derived Registries for Population Health**

We observed three important limitations of available EHR data from the perspective of public (or population) health in the process of creating the Indiana Registry for TBI, SCI and stroke. First, data from EHRs, while voluminous, are also “noisy” and heterogeneous, which presents analytical challenges for rapid, just-in-time scenarios. Many of the planned analyses for the data in this registry require significant effort from researchers and epidemiologists to “clean” and prepare analytical datasets. Thus, the hope of a turnkey solution for extracting EHR data for rapid ascertainment of population health — a tenet of the Learning Health System\textsuperscript{19}— that is also of interest to agencies that respond to emerging disease threats, is unlikely given the reality of current EHR systems.

Second, while the breadth of data from EHR systems holds many opportunities for public health their depth remains a challenge. Public health departments are interested in examining not only utilization, mortality and clinical outcomes but also the social determinants of health. In particular, public health agencies desire robust data on race and ethnicity as they are chief among organizations who analyze health disparities. Data on race and ethnicity was missing for one-third of SCI patients, a rate not unlike other studies involving EHR data. Furthermore, data on the social determinants of health are not routinely captured by EHR systems\textsuperscript{20}. Many epidemiological cohort studies capture these data directly from patients. While direct capture may not be possible, deriving such data from other sources (e.g. clinical notes) could help improve the value of specialized registries.

Finally, the registry was unable to tap into free-text data available within the INPC due to concerns around identifiable
data that may be embedded within discharge summaries, history and physical notes, admission notes, etc. While methods exist for automating the de-identification of a clinical note\textsuperscript{21,22}, these methods are not yet routinely used by health systems or HIE networks. If such methods could be implemented, concepts could be automatically extracted from free-text for use by registries.

**Future Directions**
The TBI, SCI and stroke registry presents a wealth of data which may be used to evaluate the outcomes for these conditions. The following analyses are currently being conducted by our team for dissemination in the coming years:

- **Opioid usage and secondary stroke events:** Prolonged opioid usage is associated with adverse cardiovascular events. In our analysis of the medication data from the stroke registry we found that opioid and opioid combination medications were the second most commonly prescribed drug class among stroke patients both before and after the stroke event. We will compare secondary stroke outcomes in patients that had prolonged opioid usage with those that did not.

- **Antidepressant usage and stroke outcomes:** Mixed evidence suggests that second generation antidepressants may be associated with adverse cardiovascular events. Depression is a frequently observed co-morbidity among stroke patients and understanding whether the use of second generation antidepressants is related to increased risk of adverse events may inform best practices in the care of these patients. We will examine adverse cardiovascular events and all-cause mortality in stroke patients who were diagnosed with depression and were also medicated with second generation antidepressants.

**Conclusion**
We used EHR data from multiple health systems and providers to build a specialized registry of TBI, SCI and stroke patients. In addition to providing injury surveillance, data from this registry may also be used for research purposes. Further, this data may provide a deeper understanding of the care process as well as long-term outcomes in TBI, SCI and stroke patients. Our registry demonstrates that it is feasible to create a specialized registry by leveraging EHR data from an HIE.

**Acknowledgements**
This work was supported by a contract from the Indiana State Department of Health (No. 16862). The authors acknowledge Katie Hokanson, Director, Trauma and Injury Prevention Division at ISDH and James F. Malec, PhD, Senior Research Professor of Physical Medicine & Rehabilitation, IU School of Medicine, for their support and guidance on the conduct of our work.

**References**
2. Office of the National Coordinator for Health Information Technology. Non-federal Acute Care Hospital Electronic Health Record Adoption. Health IT Quick-Stat #47 2016.
Stakeholder Use and Feedback on Vaccination History and Clinical Decision Support for Immunizations Offered by Public Health

Sripriya Rajamani, MBBS, PhD, MPH1, 2, Aaron Bieringer, BS3, Similoluwa Sowunmi, MPH3, Miriam Muscoplat, MPH3
1Public Health Informatics Program, 2Institute for Health Informatics, University of Minnesota, Minneapolis, MN; 3Minnesota Department of Health, St. Paul, MN

Abstract
National initiatives on Electronic Health Records (EHRs) recognize the vital role of public health and recommend reporting to Immunization Information Systems (IIS) and access of its clinical decision support for immunizations (CDSi). The objective of this study was to collect stakeholder feedback on access and utilization of CDSi from the Minnesota Immunization Information Connection (MIIC), Minnesota’s IIS. Input was solicited using a semi-structured questionnaire developed by experts, and from a sample of 17 key informants from February 2015 through May 2016. Analysis highlighted the appreciation of MIIC services, comprehensive vaccination history across providers and CDSi functionality, with public health users relying on MIIC. It also identified issues such as data entry due to read-only view, data quality and communications for improvement. These findings underscore the critical role of IIS, need to engage stakeholders, ensure CDSi updates, maintain good data quality, and promote bi-directional data exchanges across EHRs-IIS.

Introduction
Immunization Information Systems (IIS) are population-based, secure computerized systems present in most US states and territories1. IIS serve as a hub for immunization data given across providers over time. As delivery of certain preventive services, including immunizations, spreads beyond the confines of traditional healthcare organizations, IIS play a critical role in collating immunizations obtained through a spectrum of stakeholders. It aims to present a comprehensive vaccination history and addresses the potential record scatter across providers. IIS contains clinical decision support for immunizations (CDSi), incorporating recommendations from the Advisory Committee on Immunization Practices (ACIP)2. These guidelines are complex and include a range of factors like age, number of doses, dose intervals, medical conditions and vaccine contraindications. IIS are recognized as one of the strategies to utilize to meet Healthy People 2020 goals of decreasing vaccine-preventable infectious diseases by improving vaccination rates.

National initiatives that promoted adoption of electronic health records (EHRs)3,4 have also advocated for data exchange across settings using nationally recommended standards5. These regulations have recognized the vital role of public health and recommended reporting to Immunization Information Systems (IIS) as one of the priority areas. Recent regulations have built upon this to include bi-directional communications across EHRs-IIS6. The Advancing Care Information (ACI) measures for Merit-Based Incentive Payment System (MIPS) includes submission of immunization data and receipt of immunization forecasts and histories from the public health immunization registry/immunization information system7. Given these regulatory drivers, the Preventive Taskforce recommendations8, studies on IIS effectiveness9,10 and the increasing use of CDS with embedded immunization guidelines in EHRs, there is a need to understand the stakeholder access and utilization of IIS CDSi.

EHR-IIS research to-date comprises of single clinical setting reports11,12 and assessing automated reporting from EHR to IIS13,14. The overlap of functionalities between EHRs and IIS have been outlined with recommendations for guiding their integration15. Bi-directionality across IIS and EHRs has been advocated wherein benefits of integration supersede the challenges posed by variability in EHR technologies and various IIS16. Given the increasing importance of CDSi, there have been efforts to examine EHRs for functionality17 and creation of computable CDSi logic18. Study on access to IIS decision support within EHR19 concluded that visual integration of external registries into an EHR was feasible with improvement in provider satisfaction and registry reporting. Recent studies have examined EHR-IIS data exchange and access of IIS CDSi, but are focused in single system/setting19,20. A collaborative strategy emphasizing shareable solutions has been proposed with goals of making CDS sustainable for private and public sectors21. Given this landscape, there is a growing need for research to understand the access and use of CDSi at point of care, specifically through EHRs.

The Minnesota Immunization Information Connection (MIIC)22, the IIS in Minnesota, has been operational since 2002 and currently has 84 million immunizations and 7.8 million individuals across the life span. Overall, there are approximately 5,300 active organizations in MIIC ranging from primary care clinics, specialty providers, hospitals, pharmacy, schools and public health. MIIC currently offers an option to access MIIC from the provider EHR and is branded as ‘Alternate Access’23. This solution generates a query to MIIC for vaccination history and
forecasting and is based on demographics of the record in the EHR. It does not require separate MIIC log-in and addresses the issue of data entry for the query. With this functionality, the history and forecast are presented to the user within their EHR. This data can either be displayed as a ‘read-only’ view or can be ‘integrated’. Many have implemented a ‘read-only’ option, but moving towards ‘integrated’ with recent EHR system upgrades.

Prior research by authors has examined the technological and organizational context around immunization reporting to MIIC to portray increasing electronic reporting from EHRs using standards and real-time methods. Recent MIIC CDSi studies have focused on understanding the queries and its variability across providers and EHR implementations and analyzing the CDSi presentation through direct access by IIS interface and by access through electronic health records (EHRs) to outline similarities and differences. Building upon prior work and on national recommendations which promote IIS CDSi use, the goal of this research is to understand various aspects of MIIC CDSi from end-user/stakeholder perspectives. The main study objective was to collect stakeholder feedback on access and use of MIIC CDSi with the intent of utilizing results for prioritizing system enhancements and for program improvement. The overarching goal is to enhance the utility of IIS and its CDSi functionality. The study will have potential implications for using IIS CDSi as a strategy to increase immunization rates and improve population health.

Methods

Subject and Site Selection

Data was collected from a purposive sample of subject matter experts selected from healthcare systems and public health clinics. A total of 17 professionals were interviewed from 12 organizations representing 324 individual care sites. Their expertise spanned across public health and clinical care settings, and across clinical and administrative roles. These experts were selected based on their prior collaborations with MIIC and for their familiarity with the tool. The interviewee selection was also based on their ability to share information at a health system level vs. a site level, as approximately nine healthcare delivery systems hold 90% of the health services market in Minnesota due to the high presence of integrated delivery networks. Likewise, the interviewees of local public health were chosen from relatively high volume clinics to get granular information on MIIC CDSi access and some had an EHR in place for the last few years. Table 1 presents their various roles which range from vaccine coordinator, clinical analyst, nurse administrator, quality improvement manager and EHR project manager.

Table 1: Participant Roles and Settings

| Roles          | Clinical Care Setting                              | Public Health Setting                        |
|               | Health System / clinic                             | Local health department / public health clinic |
| Clinical       | Clinical Analyst (1)                               | Clinical Director (1)                        |
|               | Vaccine Coordinator (2)                            | Vaccine/Immunization Coordinator (2)         |
|               | Licensed Practical Nurse (1)                       | Immunization Manager (1)                     |
|               | Vaccine and Infection Prevention Specialist (1)    | Nurse Epidemiologist (1)                     |
|               | Health Outreach Coordinator (1)                    | Public Health Nurse (2)                      |
| Administrative | Nurse Administrator (1)                            | EHR Project Manager (1)                      |
|               | Quality Improvement Manager (1)                    | Support Staff (1)                            |

Development of the Interview Questionnaire

The study utilized qualitative methodology to collect information from main stakeholders and users of MIIC CDSi. These are subject matter experts from healthcare systems and local health departments / public health clinics. A semi-structured interview questionnaire was developed by study authors (SR, AB) and refined through input from MIIC leadership and experts in CDSi and IIS. The tool aimed to collect data across four domains and consisted of both structured and open-ended questions to facilitate the gathering of information (Refer to Table 2 for the Interview Questionnaire). The tool was pilot tested with a clinic first, minor edits incorporated and was then utilized as the main data collection instrument.

The background section aimed to understand the number of sites in the organization that had access to MIIC CDSi functionality (both embedded EHR access and also access by MIIC interface); time and reason for installation of EHR access feature and if it was included in EHR training or other user education efforts. This was followed by
a section on the awareness and use of MIIC CDSi collecting data on number of users, type of users (e.g. provider, nurse, clinic manager) and if it was accessed for all visits/select visits. This section also focused on immunization data across EHRs and MIIC and if data missing across either systems was shared/reported, including highlighting of errors. It also aimed to solicit awareness of MIIC CDSi. The third section addressed the value of MIIC CDSi by gathering data on perceived value of the tool, information that is sought (e.g. vaccination history, immunization forecast) and integration with workflow. Tracking of missed opportunities by the system/clinic (if any) and ideas for MIIC CDSi improvement were some essential elements of this interview process. The closing section aimed to collect information on organizational initiatives under which MIIC CDSi is promoted (if any) and other organizational subject matter experts.

Table 2: Interview Questionnaire

<table>
<thead>
<tr>
<th>A. BACKGROUND</th>
</tr>
</thead>
<tbody>
<tr>
<td>1) What is your role at your organization?</td>
</tr>
<tr>
<td>2) We understand that MIIC CDSi data is sent to your EHR. What is the number of sites that have access to this data and what number of these sites use the CDSi data in their practice?</td>
</tr>
<tr>
<td>3) When was this feature installed in your system? Why? What led to this decision?</td>
</tr>
<tr>
<td>4) What was the process for assessing client immunization history and vaccine recommendations (decision support) prior to that?</td>
</tr>
<tr>
<td>5) Is this functionality included as part of EHR training or other user education efforts?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>B. AWARENESS AND USE OF MIIC CDSi</th>
</tr>
</thead>
<tbody>
<tr>
<td>6) How many users are there in your clinic/health system that use the CDSi data from MIIC?</td>
</tr>
<tr>
<td>7) Who typically accesses this functionality? Nurse, provider, clinic manager?</td>
</tr>
<tr>
<td>8) How many client visits does your system have annually?</td>
</tr>
<tr>
<td>9) Is MIIC accessed during all visits or only for select visits / client group?</td>
</tr>
<tr>
<td>10) How is this integrated with workflow? Is it similar across various clinic types (pediatric vs. other settings)? When is the CDSi data reviewed?</td>
</tr>
<tr>
<td>11) When viewing CDSi information from MIIC, does data not present in EHR get entered into your system? If not, why not?</td>
</tr>
<tr>
<td>12) If there is data in EHR and not in MIIC, do they get reported back to MIIC?</td>
</tr>
<tr>
<td>13) If errors noticed in the MIIC data (CDSi or immunization data), are they noted and reported back to MIIC?</td>
</tr>
<tr>
<td>14) Are users aware that the source of the immunization data (history and forecasting) is MIIC?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>C. VALUE OF MIIC CDSi</th>
</tr>
</thead>
<tbody>
<tr>
<td>15) What do users typically look for/what are users most interested in? New immunization data or validate current ones and/or look for vaccine recommendations?</td>
</tr>
<tr>
<td>16) Please speak to the value of the MIIC CDSi information. Does your organization find this information valuable? If not, why not?</td>
</tr>
<tr>
<td>17) Does your system use CDSi offered by any sources other than MIIC? How is this integrated into workflow?</td>
</tr>
<tr>
<td>18) If you didn’t have access to MIIC for CDSi what would your organization use? Has your organization looked into alternate CDSi solutions?</td>
</tr>
<tr>
<td>19) Do you track immunization rates by clinic and try to find “missed opportunities”? If so does the MIIC CDSi data help you determine a missed opportunity, or do you do something else to calculate that?</td>
</tr>
<tr>
<td>20) What would like to see happen to improve use and value of this feature? (e.g. software improvement, ideas to promote use)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>D. CLOSING</th>
</tr>
</thead>
<tbody>
<tr>
<td>21) What are the initiatives under which this functionality is promoted/monitored? (E.g. quality improvement, public health, clinical decision support); if not, why not?</td>
</tr>
<tr>
<td>22) Who else in your organization/clinic would be involved with this and could be subject matter expert (SME) for this purpose?</td>
</tr>
<tr>
<td>23) Any additional information you would like to share or relevant topic that needs to be addressed?</td>
</tr>
</tbody>
</table>

Data Collection

Subject matter experts were contacted over email and interview questions were shared a couple of weeks in advance to help them prepare and gather additional information as needed. An opportunity was given to extend the interview to their colleagues as needed. These interviews were conducted over the time period of February 2015 through May 2016. WebEx was used for set-up and recording of the meetings. Three study authors (SR, AB and SS) were present during all the interviews and two authors (SR and SS) took detailed notes during the interview process which was utilized for the study analysis. Verbal consent was obtained prior to the interview and each lasted for approximately 45 minutes. Notes were exchanged immediately after the interview and cross referenced so that any discrepancy can be resolved.
**Data Analysis**

Preliminary analysis and exchange of notes (SR, SS) occurred between the interviews. Analysis of qualitative open-ended answers was supported by NVivo 11 software (QSR International). One of the study authors (SR) participated in all the interviews, was one the main note takers, has worked in public health and in immunization program for many years and was the main coder. These codes were synthesized through discussion (SR, AB and MM) which lead to consolidating categories that overlapped and identifying higher level themes. The responses to structured questions were analyzed using descriptive statistics and tabulation was utilized to present results. The responses considered important to highlight were selected through group consensus (SR, AB, MM).

**Results**

Table 3 presents the responses by the 17 subject matter experts from 12 organizations, and highlight the key areas with MIIC access. Themes and categories identified in the analysis along with sample quotes are in Table 4.

**Table 3: Access to MIIC across Clinical Care and Public Health Settings**

<table>
<thead>
<tr>
<th>Health Care Settings</th>
<th>Access to MIIC</th>
<th>Part of EHR Training</th>
<th>Type of MIIC Access</th>
<th>Integration with Workflow</th>
<th>MIIC Data / CDSi in Care Process</th>
<th>Main User Base</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Clinical Care (Health System / Clinic)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Organization A</td>
<td>✓</td>
<td>From EHR; read-only view</td>
<td>Part of patient preparation &amp; before visit</td>
<td>MIIC lookup only when vaccine due/overdue in EHR</td>
<td>RN; Pediatric providers</td>
<td></td>
</tr>
<tr>
<td>Organization B</td>
<td>✓</td>
<td>From EHR; read-only view</td>
<td>Previsit access for routine &amp; in visit for same day appts</td>
<td>First look at EHR &amp; then access MIIC as needed</td>
<td>Nurses</td>
<td></td>
</tr>
<tr>
<td>Organization C</td>
<td>✓</td>
<td>From EHR; read-only view</td>
<td>Pre-visit access for routine / physicals</td>
<td>All pediatric visits include MIIC lookup</td>
<td>Nurses</td>
<td></td>
</tr>
<tr>
<td>Organization D</td>
<td>✓</td>
<td>From EHR; bidirectional data exchange</td>
<td>Currently check in visit; moving to pre-visit planning</td>
<td>Depends on clinic &amp; visit type</td>
<td>Nurses; occasionally providers</td>
<td></td>
</tr>
<tr>
<td>Organization E</td>
<td>?</td>
<td>From EHR; read-only view</td>
<td>No set protocol; varied access</td>
<td>Only if associated with infectious disease</td>
<td>Medical Assistant</td>
<td></td>
</tr>
<tr>
<td>Organization F</td>
<td>×</td>
<td>No integrated EHR access; separate MIIC log-in</td>
<td>No set protocol; varied access</td>
<td>Only for well child visits &amp; in cases of uncertain history in EHR</td>
<td>Front desk staff</td>
<td></td>
</tr>
<tr>
<td>Organization G</td>
<td>×</td>
<td>From EHR; read-only view</td>
<td>Chart scrub 3 days advance; same day for sick visits</td>
<td>For family practice, general medical &amp; sick visits; not for specialty visits</td>
<td>Mainly clinic assistants</td>
<td></td>
</tr>
<tr>
<td>Organization H</td>
<td>✓</td>
<td>From EHR; bidirectional data exchange</td>
<td>Pre-visit, but some variation in practice</td>
<td>Most visits except for urgent care</td>
<td>Nurses; medical records team</td>
<td></td>
</tr>
<tr>
<td>Organization I</td>
<td>×</td>
<td>No integrated EHR access; separate MIIC log-in</td>
<td>Part of patient preparation &amp; before visit</td>
<td>All pediatric &amp; prenatal diabetic visits; inconsistent for other visits</td>
<td>Nurses; providers; billing</td>
<td></td>
</tr>
<tr>
<td>Organization J</td>
<td>✓</td>
<td>From EHR; bidirectional data exchange</td>
<td>MIIC look-up during vaccine appointment scheduling</td>
<td>All visits to public health clinic</td>
<td>Public health nurses</td>
<td></td>
</tr>
<tr>
<td>Organization K</td>
<td>✓</td>
<td>From EHR; bidirectional data exchange</td>
<td>MIIC look-up during vaccine appointment scheduling</td>
<td>All visits to public health clinic</td>
<td>Public health nurses; registration staff</td>
<td></td>
</tr>
<tr>
<td>Organization L</td>
<td>✓</td>
<td>From EHR; bidirectional data exchange</td>
<td>Part of patient visit</td>
<td>Look up during immunization walk-in clinics &amp; home visits</td>
<td>Public health nurses</td>
<td></td>
</tr>
</tbody>
</table>

✓ - included; × - not included; ? - not sure; EHR - Electronic Health Record; RN - Registered Nurse; appts - appointments
Most of the organizations interviewed (10 of 12, 80%) had MIIC access within their EHRs. Of the 10 organizations with embedded EHR access to MIIC, almost all had included this functionality as part of their EHR training (8 answered yes and 2 were uncertain). All the public health clinics interviewed (with the exception of 1) had bi-directional data exchange across MIIC and EHRs which indicated that data was being transferred from MIIC to EHRs and integrated into the EHR record. Of the organizations which had pre-visit planning in place (10 of 12), almost all had access to MIIC as part of patient preparation/arrival. Scenarios which prompted access to MIIC varied across the organizations, and predominantly part of pre-visit and MIIC look-up during a clinical encounter was done only for urgent care visits. Two organizations (A and B) which are the dominant integrated health systems in the state utilized the CDSi present in their EHR and looked up MIIC only for vaccination history. MIIC was accessed by nurses, clinic assistants and care support professionals (e.g. front desk, scheduling, billing) in all the organizations (12 of 12, 100%) and had limited access by providers (e.g. physicians).

Table 4: Themes, Categories and Quotes regarding MIIC

<table>
<thead>
<tr>
<th>Themes</th>
<th>Categories</th>
<th>Sample Quotes</th>
<th>Setting and Role of Interviewees</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recognition of the Value of MIIC</td>
<td>Vaccination history</td>
<td>…&quot;Valuable in public health – see clients which are mobile population – so using MIIC which is shared system is important (as clients drop on and off insurance)&quot;</td>
<td>Public health; Public health nurse</td>
</tr>
<tr>
<td></td>
<td>Immunization forecast</td>
<td>…&quot;MIIC forecasting is nice as it gives earliest date – if kids come in early, it is easy to make a judgement call&quot;</td>
<td>Clinical and public health setting; Clinical analyst, Vaccine coordinator</td>
</tr>
<tr>
<td></td>
<td></td>
<td>…&quot;no, we don’t use any other – rely on MIIC&quot;</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>…&quot;like the forecasting functionality&quot;</td>
<td></td>
</tr>
<tr>
<td>Appreciation of MIIC services</td>
<td></td>
<td>…&quot;MIIC is valuable to end-users; to have all that information handy&quot;</td>
<td>Clinical care and public health setting; Licensed practical nurse, Vaccine / immunization coordinator</td>
</tr>
<tr>
<td></td>
<td></td>
<td>…&quot;very easy to access and use….print reports&quot;</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>…&quot;honestly, MIIC is amazing ….have used for 16 years….seen it only get better&quot;</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>… “Absolutely of value – I am in MIIC constantly – at the click of the button – rely on MIIC to make sure immunizations are given correctly”</td>
<td></td>
</tr>
<tr>
<td>Benefits of System and Workflow Integration</td>
<td>EHR Integration</td>
<td>…&quot;what’s helpful now – within Epic ambulatory module – with click of a button, can get into MIIC”</td>
<td>Clinical care setting; Nurse administrator</td>
</tr>
<tr>
<td></td>
<td>Fit with Workflow</td>
<td>…&quot;I do the day before (check MIIC) to see if there are any catch-up opportunities”</td>
<td>Public health; Public health nurse</td>
</tr>
<tr>
<td></td>
<td></td>
<td>…&quot;part of our routine pre-visit planning”</td>
<td></td>
</tr>
<tr>
<td>Functionality and System Requirements to Address</td>
<td>Vaccine Details</td>
<td>…&quot;more details on vaccine received, specifically the brand. Who is the manufacturer? Who gave the vaccine?&quot;</td>
<td>Clinical care setting; Nurse administrator</td>
</tr>
<tr>
<td></td>
<td>Forecaster Improvement</td>
<td>…&quot;foreign born don’t fit into usual schedule routine”</td>
<td>Public health; Nurse epidemiologist</td>
</tr>
<tr>
<td></td>
<td></td>
<td>…&quot;love for the forecaster to be improved”</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Data Quality</td>
<td>…&quot;MMR rates in MIIC is really low – not due to missed opportunities, but due to missing shots in MIIC”</td>
<td>Public health; Immunization manager</td>
</tr>
<tr>
<td></td>
<td>Data entry</td>
<td>…&quot;need to manually enter MIIC data”</td>
<td>Clinical care setting; EHR proj. manager</td>
</tr>
<tr>
<td></td>
<td>System Capacity</td>
<td>…&quot;Sometimes MIIC is extremely slow – doesn’t query quick enough as I have hundreds of people (but slows during the day) – sometimes 4-5 min”</td>
<td>Clinical care setting; Outreach coordinator</td>
</tr>
<tr>
<td>Need for Better Communications</td>
<td>Information Sharing</td>
<td>…&quot;communications from MIIC to sites for error corrections will be good”</td>
<td>Public health; Immunization coordinator</td>
</tr>
<tr>
<td></td>
<td>Outreach</td>
<td>…&quot;when MIIC is down, we are not informed – so, an issue when kids are in the clinic”</td>
<td>Clinical care setting; Clinical analyst</td>
</tr>
</tbody>
</table>
Four main themes were identified which included feedback on MIIC CDSi tool and its utility and items which
needed action: Recognition of the Value of MIIC; Benefits of System and Workflow Integration; Functionality
and System Requirements to Address and Need for Better Communications. MIIC value proposition was due to
the comprehensive vaccination history provided, the immunization forecaster (CDSi) and other
services/functionality offered (e.g. Reminder/Recall, Improbable shots report, Vaccine Management). The access
which was integrated within the EHR eliminating the need for a separate user interface and data entry was highly
appreciated by the users. Overall, MIIC access seemed well integrated within the workflow of various
organizations. Five functional/system requirements were identified for improvement ranging from need for
additional details on the vaccine given; improvement of the forecaster by inclusion of guidelines for those who
don’t fit into regular schedule; improving the quality of data in MIIC; need for bi-directional data
exchange/dynamic data flow to eliminate dual data entry and finally enhancing the capacity of the system to
facilitate quick response time to queries. Need for information sharing and outreach was mapped into the final
theme on the need for better communications (from MIIC to its users/stakeholders) which underscored the need
to inform users ahead of time on planned system upgrades/down time, to facilitate planning of clinic vaccinations.

Discussion

Results point to strong support of MIIC and recognition of its value and the services/tools offered. Almost all
participating organizations (12 of 12), which include 324 individual sites had access to MIIC in varying formats
embedded within their care delivery. These represent approximately 25% of the sites/clinics in the state and so a
reasonable indicator of the overall pattern. Majority of the organizations interviewed (10 of 12, 80%) had MIIC
access within their EHRs, but some of those did not have dynamic data exchange which supported data entry from
MIIC into EHRs. But, this issue should be resolved in the near future, as current version of the prominent EHR
product used in Minnesota (Epic) supports this functionality and organizations are upgrading to this version.

One of the findings was the systematic difference between private and public health clinics in their access and use
of MIIC and its perceived utility, with public health clinics/local health departments much more appreciative of a
public health information system such as MIIC. This is likely due to the patient population they serve, many of
whom have sporadic encounters with the healthcare system and have providers who are spread across locations
and systems resulting in an immunization record scatter. IIS serves as a hub to collate this vaccination history
across providers and locations and presents immense value to these settings. The appreciation of MIIC is also due
to the fact that low resource settings likely have EHRs with fewer capabilities such as clinical decision support
and hence rely more on MIIC services such as CDSi and various tools/reports (client assessment, vaccine
management).

As initiatives are being explored to increase the utility of MIIC CDSi, this study yielded important findings on the
main user base, who are primarily nurses in both clinical and public health settings along with support
professionals (e.g. front desk, scheduling, billing), with limited access by providers (e.g. physicians). The
literature on clinical decision support and national recommendations advocate for access during clinical
encounters/point of care. But the majority of settings in this study pointed to access of MIIC CDSi as part of pre-
visit planning. As per study participants, this timing of access seems to integrate better with their workflow as
vaccinations are typically administered during scheduled well-child visits for pediatric population. As the
recommendations are to access CDSi during the point of care, additional research needs to be done to understand
the obstacles (if any) to access the CDS tools during a clinical encounter. This is an important issue which will
affect the design and implementation of EHRs and their integration with IIS.

This study pointed to some system/functionality needs which required actions by MIIC program management.
The need for better communications was acted upon by MIIC leadership with regular email updates sent to
stakeholders on scheduled MIIC downtimes and also immediate information shared on unexpected data exchange
malfunctions so that an organization can re-submit their data. Though MIIC had communication mechanisms on
error corrections, the interviews pointed to the need for selecting the right professional at the provider end to receive
the messages and to address those errors.

The study emphasized the need for good quality of data in the IIS (MIIC in this study) as the forecasting is
dependent on IIS data and becomes irrelevant with incomplete data in the IIS. The issue of missing shots
impacting completeness of data identified by stakeholders is due to non-submission of data by some organizations
and efforts are being made to increase provider participation in MIIC. Likewise, the need for better data quality
in MIIC was addressed by creation of a data quality coordinator position whose responsibility is to monitor quality
of data and identify systematic errors with incoming data and also implement protocols for regular data quality
assessment. The increasing bidirectional movement of data across MIIC and EHRs due to capacity of EHRs to
support this data interchange highlights the increasing need to monitor quality of data to ensure data is correctly
being attributed to organizations which have the shot and those that incorporate and report them to IIS. These findings also supported the need to use a data quality monitoring tool on a regular basis for data validity checks.

One of the limitations of this study are that it did not cover all the sites that accessed MIIC CDSi. The participating organizations represent approximately one fourths of user base and so findings can be applicable to other sites that are similar in size and patient population. Participants were chosen based on their expertise and their position within the organization to share practice information from a system level, but it’s possible that there is variation across sites. The other shortcoming is that, after completion of this research, three of the main healthcare systems with hundreds of clinic sites have either upgraded their EHR product or switched to a new EHR platform and so the functionalities (specifically the capacity for dynamic data interchange across MIIC and EHRs) may have since improved.

Another limitation of this study is that it focuses on a single state IIS. MIIC is part of WIR consortium28 which is a collaboration of group of states on the same vendor platform as MIIC and is likely that some of these findings are applicable to them, but need additional studies for validation. Additional aspects which are unique to this study setting and can potentially limit generalizability is the predominance of integrated health care delivery network structure in the state and the market dominance by a single EHR vendor. Yet another limitation is that this research did not examine the role of third-party immunization clinical decision support29 and its role. This study focused on understanding the overall access and utilization of MIIC CDSi and additional studies are needed to understand the variability in CDSi across MIIC and third-party solution. The American Immunization Registry Association (AIRA)30, a membership and advocacy organization for IIS across the various states should share findings from this and other IIS studies to share lessons learned.

Conclusion

With the recognition of the importance of public health reporting and the recommendation to look up and use CDSi offered by IIS, there is a need to understand the current status of utilization. This study collected stakeholder feedback on access and use of vaccination history and MIIC CDSi with the intent of utilizing results for prioritizing system enhancements and program improvement. Research insights can be applied to other scenarios of public health decision support (e.g. case reporting) and other situations of HIT implementation and evaluation. The overarching goal is to enhance utility of IIS and its CDSi functionality and study will have potential implications for using IIS CDSi as a strategy to increase immunization rates and improve population health.

Acknowledgements

The authors would like to thank the various stakeholders from private health systems and from public health clinics who participated in this study for their time and valuable input.

References


Formative Evaluation of CareNexus: a Tool for the Visualization and Management of Care Teams of Complex Pediatric Patients

Pallavi Ranade-Kharkar, PhDc, MS1,2, Chuck Norlin, MD1,3, Guilherme Del Fiol, MD, PhD1
1Department of Biomedical Informatics, University of Utah, Salt Lake City, UT; 2Intermountain Healthcare, Murray, UT; 3Department of Pediatrics, University of Utah Health Sciences Center, Salt Lake City, UT.

Abstract
Complex and chronic conditions in pediatric patients with special needs often result in large and diverse patient care teams. Having a comprehensive view of the care teams is crucial to achieving effective and efficient care coordination for these vulnerable patients. In this study, we iteratively design and develop two alternative user interfaces (graphical and tabular) of a prototype of a tool for visualizing and managing care teams and conduct a formative assessment of the usability, usefulness, and efficiency of the tool. The median time to task completion for the 21 study participants was less than 7 seconds for 19 out of the 22 usability tasks. While both the prototype formats were well-liked in terms of usability and usefulness, the tabular format was rated higher for usefulness (p=0.02). Inclusion of CareNexus-like tools in electronic and personal health records has the potential to facilitate care coordination in complex pediatric patients.

Introduction
Children and youth with special health care needs (CYSHCN) generally have or are at a risk of developing chronic medical and mental health conditions, resulting in an increased need for specialized medical, therapeutic, equipment, family support, and other services. 15.6% (approximately 11 million) of the pediatric population has special needs and accounts for about a third of the total healthcare spending associated with children. Coordinating care for these medically complex patients is crucial for enabling efficient use of resources, reducing costs, enhancing communication between patient/family and provider, and improving patient/family and provider satisfaction.

Caring for CYSHCN often involves a large number of participants from disparate settings, working independently, and serving in various roles that may wax and wane in importance or need over the continuum of patient care. The clinical care teams often include the primary care physician and a number of specialists, care coordinators, therapists, and social workers. However, the care teams of CYSHCN also extend far beyond the hospital/clinic walls and may include schools; community resources such as support groups and family advocacy groups; cultural and charitable organizations; multiple payers and funding agencies; and family members, neighbors, and friends. In a previous study, we identified and described information needs of physicians, care coordinators, and parents to support care coordination of CYSHCN and categorized them into information goal types. One of the goal types we identified was care networking and we defined it as “building a patient’s care team or network, knowing team member identities and roles, and sharing pertinent information to enable activities/actions as a team”. Keeping track of who is involved in a patient’s care at any given time, what their roles are, their goals and feedback, and preferred contact information; and finding new care team members to fit patient and family needs, was described as extremely challenging by the interviewees. Because they lacked appropriate tools, physicians, care coordinators, and parents of CYSHCN resorted to workarounds such as spreadsheets, hand-written sticky notes, refrigerator magnets, and memorization to store and track care team-related critical information.

In spite of the great potential for electronic health records (EHRs) to help providers coordinate care, current EHR systems do not adequately support the needs of care coordination. Vawdrey et al. note that we need better tools to support care team-related information in commercial EHR systems. Usability of EHRs also falls short in supporting the unique needs of representing information about complex patients. A recent systematic review reported that lack of appropriate software functionality and poor user interfaces were linked to patient safety concerns. Electronic personal health records (PHRs) have been proposed as a strategy to support care coordination. However, few PHRs have been developed for the domain of pediatrics due to lack of standards for pediatric content and customizations needed for chronic conditions. With lack of PHRs to support care networking, parents of CYSHCN bear the burden of maintaining information about their child’s care team and repeating their “story” while coordinating care among a large and diverse care network.
In this study, we partner with care team members in key and complementary roles: physicians, parents, and care coordinators of CYSHCN, to gain insights into designing an application for clinicians and patients/families for the purpose of care networking. We iteratively design two alternative user interfaces to view, understand, share, and manage patient care team information. We then implement the designs and conduct a formative evaluation of the usability, usefulness, and efficiency of the user interfaces.

Methods

This study used a within-subject design comparing the interactions of physicians, parents, and care coordinators of CYSHCN with two user interface designs (graphical and tabular) of a prototype software we named “CareNexus” to accomplish goals and tasks related to creating and managing patients’ complex care networks. The study involved both, granular tasks designed to assess usability, and high-level tasks focused on solving a care coordination problem described in vignettes. The study addressed the following research questions: 1) to what degree are features offered by CareNexus to create, understand, and manage care networks of CYSHCN easy to use and efficient?; 2) how useful are the features offered by CareNexus?; 3) how do the graphical and tabular displays of the care networks compare in terms of usability and usefulness? The study was approved by the University of Utah Institutional Review Board under protocol #IRB_00096357.

CareNexus tool design

The design of CareNexus was guided by Information Foraging theory18, Shneiderman’s principles for information visualizations19, and Jakob Nielsen’s heuristics for user interface design20. The Information Foraging theory draws an analogy between a bird foraging for food and humans foraging for information. The optimal foraging effort seeks maximum “benefit” from minimal “cost” of information seeking by identifying rich information patches. We enabled users to optimize their information seeking effort by providing information patch enrichment (i.e. providing ways to get to the relevant content quickly and easily). We also applied Shneiderman’s visualization principles by offering an overview of information at the first level, then implementing zoom-in/zoom-out functions for the information, and finally providing information details on demand. Additionally, we have incorporated Nielsen’s principles for user interaction design to the CareNexus prototype.

The design of CareNexus followed an iterative design methodology based on rapid prototyping, analyzing, and refining cycles guided by feedback from representative users from each of the target user roles: physician, parent, and care coordinator. Tabular representation of medical data in the form of charts are common in current EHRs (e.g. Cerner™ uses tables to display patient care teams). This made the tabular design an obvious choice. A recent systematic review on innovative visualization of EHR data reported that color, lines, shapes, and visual diagrams have been effectively used to render patient data21. Thus, we opted to design a graphical interface that depicts the care team as a visual diagram as an alternative to the tabular format. We started off with “low-fidelity” prototypes in the form of whiteboard diagrams and software mockups. As the design matured, we transitioned to web-based “high-fidelity” prototypes using the AngularJS™ framework, Java™ RESTful Web Services, and MySQL™ database. We further incorporated the critique from human factors and usability experts into the mature designs.

The functional specification for CareNexus is derived from the results of our previous work focused on eliciting information needs and associated goals that are raised by physicians, care coordinators, and family members while coordinating the care for CYSHCN7. CareNexus is designed with the objective of supporting the information goal of care networking, specifically creating, understanding, and managing patient care networks or care teams. CareNexus does this by supporting previously identified goal sub-types of care networking: 1) care team building, 2) care team member identities, 3) contact information, 4) shared team knowledge, and 5) shared team action.

Participants and setting

Participants in the iterative design phase and the study were recruited from primary care sites that participate in pediatric Patient-Centered Medical Home (PCMH) Demonstration projects in Utah22. Each practice has a designated care coordinator and has one or more “family partners” who are actively-engaged parents of CYSHCN. The iterative user design phase included one of the co-authors (CN), one parent, and one care coordinator. For the formative evaluation, we recruited a purposive sample of 21 subjects (7 each of physicians, parents, and care coordinators) with the following criteria: 1) a minimum of 2 years of current experience caring for CYSHCN; 2) experience across a wide range of clinical and patient conditions; and 3) no previous exposure to the CareNexus tool. The participants were invited by email to join the study by the project director of the PCMH Demonstration and co-author CN. We determined the sample size by following recommendations from the literature23,24.
Case vignettes
The two case vignettes used in the study were adapted from the “Essential Information for Children with Special Healthcare Needs” project headed by the HL7 Child Health work group. The primary author contributed to this project by suggesting use cases and writing storyboards for the selected use cases. The case vignettes were representative of the clinical conditions, information needs, and challenges that often face this cohort of patients and are comparable in complexity. Each case vignette consisted of a narrative about the patient’s clinical and social context and a care coordination problem related to a current event or episode in her/his life that needed to be resolved using CareNexus. The two case vignettes were further customized to the role of the study participant (physician, parent, or care coordinator). The case vignettes and the associated data were synthetic and were approved by the users in each of the three roles who participated in the user interface design phase.

Procedure
The study was conducted either in an office setting at the work sites of the participants or at their homes. The two case vignettes and two interface designs resulted in four possible case vignette/display format combinations: 1) case vignette 1 + graphical format; 2) case vignette 1 + tabular format; 3) case vignette 2 + graphical format; and 4) case vignette 2 + tabular format. Each of the participants interacted with two of the four combinations in random order such that all participants interacted with both the displays and both case vignettes.

The study session began with a brief introduction of the study. In study part 1, the participants were asked to complete the following steps for their first case vignette/CareNexus display format combination: 1) usability tasks: perform 22 tasks (Table 1) distributed over the 5 goal subtypes of care networking; and 2) problem-solving: identify care team members to communicate with using CareNexus to resolve a problem related to care coordination for a current episode of care as posed by the case vignette. This was followed by a questionnaire that assessed the usability and usefulness of one display format of CareNexus. These steps were repeated for their second case vignette/CareNexus display format in study part 2. Finally, the study subjects were asked to rate the usefulness of a set of CareNexus features and provide open-ended comments and suggestions. The participants were not provided with a tutorial of CareNexus. The goal was to assess the intuitiveness, usability, usefulness, and efficiency of CareNexus without any prior exposure to the user interface of CareNexus. The user sessions were recorded using Hypercam, a screen capture software.

Data analysis
Using the video recordings from Hypercam, each of the usability tasks were coded for: 1) ability to carry out the usability task to successful completion, and 2) time to completion. Given that the data represented repeated measurements of ratings, where the user rated two different interface designs, a paired sample data analysis was used. Comparison between the ratings of the two interface designs was performed using mixed-effects linear regression, with repeated measurements nested within user, controlling for the covariates of perceived vignette complexity, experience of the study subject with the patient conditions in the vignette, and the sequence in which the interface designs were evaluated. We developed a 14-item questionnaire with Likert-scale response options (1=strongly disagree; 5=strongly agree) to assess the usability and usefulness of CareNexus. The questionnaire included five questions from the System Usability Scale and nine questions that measured self-perceived ability to understand the gist, create, and manage care networks and related information. The individual questions were aggregated into two composite scales: usability (questions 1, 3, 8, 12, and 13) and usefulness (questions 2, 4, 5, 6, 7, 9, 10, 11, 14) to maximize reliability and generalizability. Reliability analysis was performed using Cronbach’s alpha by aggregating ratings for the two user interface designs for the composite scales.

Results
CareNexus user experience
The design of the user interface of CareNexus required ten iterations that were performed before the formative evaluation. The resulting user interface of CareNexus with the graphical care network format for case vignette 1 is shown in Figure 1. Figure 2 shows the tabular care network display for case vignette 2.

The patient banner (section #1, Figures 1 & 2) gives a quick overview of the patient and indicates additional needs if applicable (e.g. need for a language interpreter). Section #2 below shows the event timeline with boxes intuitively labeled to indicate clinical events (e.g. outpatient, inpatient, and emergency room visits) and other significant events in the patient’s life providing information-patch enrichment. The start and end of timeline defaults to six months before and three months after the current date respectively. The user can view a wider or narrower timeline range by using...
Figure 1. CareNexus user interface with a graphical view of the care network for case vignette 1.

Figure 2. CareNexus user interface with a tabular view of the care network for case vignette 2.
the navigation menu buttons or the mouse wheel. Clicking on individual events, displays detailed information about the primary care team member involved in the event in the bottom left-hand side of the screen (section #4). Shared team knowledge of the patient demographics, clinical conditions, a quick reference to the contact information of the person most involved in the patient’s care, and a quick summary of the patient’s encounters aims to provide patient context efficiently and accurately.

Users can visualize the patient’s care network in section #3 (Figures 1 & 2). Knowing who is currently involved in the patient’s care, in what role, and how to contact them is vital to care networking. The care network is displayed either in a graphical (Figure 1) or tabular format (Figure 2). In the graphical format, the patient is shown in the center of the network and is surrounded by either the family (yellow), medical (green), or non-medical (blue) care network member nodes. The tabular format displays the care team members in a table sorted by status (“active” at the top, followed by “less active”, and finally “inactive”). The family network includes immediate and extended family members involved in the patient’s care. Care team members belonging to a clinic setting are categorized as medical network members (e.g. primary care physician, care coordinator, neurologist, and social worker). The non-medical team members include the extended care network such as the school, community support groups, and durable medical equipment providers. The color gradient of the nodes of the care network and the thickness and style of the connecting lines indicate closeness of the care team to the patient’s care. There are three levels of closeness or importance: 1) active (color: darkest, connecting line: bold and solid); 2) less active (color: lighter, connecting line: medium and solid), and inactive (color: grey, connecting line: medium and dashed). The closeness or importance of a care team member to the patient’s care network can be manually assigned or can be inferred based on whether the team member has been involved in the patient’s care within a certain time frame: 1) “active” indicates activity within the last 3 months; 2) “less active” indicates activity within the last 6 months; and 3) “inactive” indicates no activity for over 6 months. The label across the top gives a quick summary of the number of active, less active, and inactive care team members. Our design goal was to provide optimal cues to users to help them understand the gist of the patient’s care network and the ability to get more information on demand, per Shneiderman’s visualization principles.

The bottom left of the screen (section #4) shows details of the care team member with the team member identity and contact information of the selected care team entity (primary care team member involved in a timeline event or a member from the care network). This information display resembles a “business card” following Nielsen’s design principle of matching real world and software system representations. Next to it we display the status, recent appointments, patient care goals, and feedback (if applicable and available) of the care team member. Building and managing a patient’s care team (section #5) can be accomplished by using the search, add, update, and delete functionality conveniently co-located in bottom center of the screen. Users can update and/or delete only those care team members who have been added by them. Finally, shared care team actions are supported by providing information about who is responsible for which task, due date (if applicable), and the status of the task. The users can view all team members’ tasks and add/delete/update their own task list. Following Nielsen’s design principles, we have maintained simplicity and color/font consistency throughout the design of CareNexus and minimized the need for user recall by providing convenient tool tips.

Ease of use, efficiency and usefulness ratings
The study participants were able to successfully complete all of the 22 usability tasks (Table 1). The median time to completion was less than 7 seconds for all but 3 tasks (searching for and manually adding care team members, and identifying events took > 7 seconds). All users also successfully identified all the care team members needed to resolve the care coordination problems posed by the case vignettes. Users highly rated all CareNexus features with highest ratings for the timeline, team member business/contact cards, and most recent and next appointments (Table 2).

Table 1. Usability tasks and time to completion (average, median, range; in seconds).

<table>
<thead>
<tr>
<th>Usability task</th>
<th>Average time (s)</th>
<th>Median time [min-max]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identify the “inactive members” of the patient’s medical care team.</td>
<td>1 ± 0.2</td>
<td>1 [1-1]</td>
</tr>
<tr>
<td>Identify the “less active” member(s) of the patient’s family network.</td>
<td>1 ± 0.3</td>
<td>1 [1-1]</td>
</tr>
<tr>
<td>Identify the “less active” member(s) in the patient’s medical network</td>
<td>1 ± 0.3</td>
<td>1 [1-1]</td>
</tr>
<tr>
<td>Identify the patient’s event timeline.</td>
<td>1.1 ± 0.3</td>
<td>1 [1-2]</td>
</tr>
<tr>
<td>Find the patient’s conditions.</td>
<td>1.2 ± 0.5</td>
<td>1 [1-3]</td>
</tr>
</tbody>
</table>

1445
Table 1 (continued). Usability tasks and time to completion (average, median, range; in seconds).

<table>
<thead>
<tr>
<th>Task</th>
<th>Average (SD)</th>
<th>Median (Min-Max)</th>
</tr>
</thead>
<tbody>
<tr>
<td>How many total members are in the patient’s family network?</td>
<td>2 ± 0.2</td>
<td>2 [2-2]</td>
</tr>
<tr>
<td>Identify the “active” member(s) of the patient’s family network.</td>
<td>2 ± 0.2</td>
<td>2 [2-2]</td>
</tr>
<tr>
<td>Who are the medical specialists the patient is actively/currently seeing?</td>
<td>2 ± 0.2</td>
<td>2 [2-2]</td>
</tr>
<tr>
<td>What is the preferred contact number for the patient’s main contact?</td>
<td>2 ± 0.2</td>
<td>2 [2-2]</td>
</tr>
<tr>
<td>Find the patient’s name on the screen.</td>
<td>2.1 ± 0.4</td>
<td>2 [2-3]</td>
</tr>
<tr>
<td>Identify the care team action(s) of the currently logged in user.</td>
<td>2.4 ± 0.8</td>
<td>2 [2-5]</td>
</tr>
<tr>
<td>How many of the care team actions are “Done”?</td>
<td>2.8 ± 1.0</td>
<td>2 [2-5]</td>
</tr>
<tr>
<td>Find the patient’s main contact person (by name or role).</td>
<td>2.8 ± 1.2</td>
<td>2 [2-6]</td>
</tr>
<tr>
<td>What is the contact information for an “active” care team member of the patient’s non-medical care team?</td>
<td>3 ± 0.2</td>
<td>3 [3-3]</td>
</tr>
<tr>
<td>Zoom in and zoom out on the event timeline.</td>
<td>4.2 ± 1.0</td>
<td>4 [3-6]</td>
</tr>
<tr>
<td>Identify the events related to the current patient episode on the timeline.</td>
<td>5.5 ± 1.6</td>
<td>5 [3-10]</td>
</tr>
<tr>
<td>Delete the care team member you added.</td>
<td>5.5 ± 0.8</td>
<td>6 [5-8]</td>
</tr>
<tr>
<td>What are the goals and/or feedback of any one of the active specialists in the patient’s care team?</td>
<td>6.6 ± 2.7</td>
<td>6 [4-17]</td>
</tr>
<tr>
<td>How many care team actions are currently displayed?</td>
<td>6.7 ± 1.7</td>
<td>6 [5-17]</td>
</tr>
<tr>
<td>Search for a new care team member.</td>
<td>13.4 ± 2.3</td>
<td>14 [8-17]</td>
</tr>
<tr>
<td>Identify events on the timeline ± 6 months from today.</td>
<td>14.9 ± 5.1</td>
<td>15 [8-29]</td>
</tr>
<tr>
<td>Manually add a new care team member.</td>
<td>32.8 ± 5.2</td>
<td>31 [26-43]</td>
</tr>
</tbody>
</table>

Table 2. Usefulness ratings of CareNexus features.

<table>
<thead>
<tr>
<th>Feature description</th>
<th>Rating (1=not at all useful; 5=very useful)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
</tr>
<tr>
<td>Patient's primary contact on the patient banner</td>
<td>4.85</td>
</tr>
<tr>
<td>Timeline of events</td>
<td>4.95</td>
</tr>
<tr>
<td>Zoom-in/Zoom-out for event timeline</td>
<td>4.33</td>
</tr>
<tr>
<td>Display of three separate care networks</td>
<td>4.81</td>
</tr>
<tr>
<td>Display of number of care team members per network</td>
<td>4.67</td>
</tr>
<tr>
<td>Color-scheme supported display of “active”, “less active”, and “inactive” care team members</td>
<td>4.81</td>
</tr>
<tr>
<td>“Business cards” for the selected care team member</td>
<td>4.90</td>
</tr>
<tr>
<td>Most recent and next appointments for the selected care team member</td>
<td>4.90</td>
</tr>
<tr>
<td>Goals for the selected care team member</td>
<td>4.86</td>
</tr>
<tr>
<td>Feedback from the selected care team member</td>
<td>4.71</td>
</tr>
<tr>
<td>Search for new care team members</td>
<td>4.62</td>
</tr>
<tr>
<td>Add to (search and manual), delete, and update the care network</td>
<td>4.86</td>
</tr>
<tr>
<td>Care team actions</td>
<td>4.71</td>
</tr>
</tbody>
</table>

Comparison between CareNexus user interface designs

Cronbach’s alpha for the composite variables of *usability* and *usefulness* are given in Table 3. There was a significant lower mean usefulness rating for the graphical interface compared to the tabular interface, after controlling for complexity, experience, and the sequence in which the interfaces were evaluated (adjusted mean difference=−0.12; 95% CI: -0.22,-0.01; p=0.02). Although statistically significant, the differences are very small and may not indicate clinical significance. There was a non-significant lower mean usability rating for the graphical interface compared to the tabular interface, after controlling for the same criteria (adjusted mean difference=−0.12; 95% CI: -0.26,0.03; p=0.12). Given our sample size of 21 we had 80% power using a two-sided alpha 0.05 comparison to detect a paired sample standardized mean difference of 0.64, which represents a moderate to large effect size by Cohen's criteria.

Table 3. Cronbach’s alpha for the composite variables.

<table>
<thead>
<tr>
<th>User interface design format</th>
<th>Usability</th>
<th>Usefulness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Graphical</td>
<td>0.79</td>
<td>0.87</td>
</tr>
<tr>
<td>Tabular</td>
<td>0.83</td>
<td>0.94</td>
</tr>
</tbody>
</table>
Table 4. Ratings of the CareNexus interface designs.

<table>
<thead>
<tr>
<th>Measurement</th>
<th>Tabular design</th>
<th>Graphical design</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1. I thought the system was easy to use.</td>
<td>4.72 0.10</td>
<td>4.55 0.10</td>
<td>0.21</td>
</tr>
<tr>
<td>Q2. I was able to grasp the gist of the patient’s care network.</td>
<td>4.81 0.11</td>
<td>4.51 0.11</td>
<td>0.06</td>
</tr>
<tr>
<td>Q3. I found the various functions in this system were well integrated.</td>
<td>4.62 0.11</td>
<td>4.56 0.11</td>
<td>0.50</td>
</tr>
<tr>
<td>Q4. I was able to find the care team members relevant to the case vignette.</td>
<td>4.82 0.09</td>
<td>4.65 0.09</td>
<td>0.124</td>
</tr>
<tr>
<td>Q5. I was able to find the pieces of information I needed to accomplish the tasks in the case vignette.</td>
<td>4.85 0.08</td>
<td>4.76 0.08</td>
<td>0.18</td>
</tr>
<tr>
<td>Q6. It was easy to understand the meaning of the information presented.</td>
<td>4.86 0.10</td>
<td>4.61 0.10</td>
<td>0.05</td>
</tr>
<tr>
<td>Q7. I was able to find the contact information for the care team member(s) I need to communicate with.</td>
<td>4.85 0.08</td>
<td>4.77 0.08</td>
<td>0.20</td>
</tr>
<tr>
<td>Q8. I would imagine that most people would learn to use this system very quickly.</td>
<td>4.82 0.08</td>
<td>4.69 0.08</td>
<td>0.15</td>
</tr>
<tr>
<td>Q9. It was easy to search for new care team members.</td>
<td>4.82 0.10</td>
<td>4.69 0.10</td>
<td>0.21</td>
</tr>
<tr>
<td>Q10. I was able to find goals of the specialists working with the patient.</td>
<td>4.86 0.08</td>
<td>4.80 0.08</td>
<td>0.24</td>
</tr>
<tr>
<td>Q11. I was able to find feedback of specialists working with the patient.</td>
<td>4.86 0.08</td>
<td>4.75 0.08</td>
<td>0.14</td>
</tr>
<tr>
<td>Q12. I think that I would like to use this system frequently.</td>
<td>4.82 0.08</td>
<td>4.79 0.08</td>
<td>0.64</td>
</tr>
<tr>
<td>Q13. I found the system very cumbersome to use. (reversed criteria)</td>
<td>1.08 0.09</td>
<td>1.33 0.09</td>
<td>0.04</td>
</tr>
<tr>
<td>Q14. Compared to the tools/workflow I currently use for care networking, I thought that CareNexus made it easier to accomplish care networking.</td>
<td>4.86 0.08</td>
<td>4.79 0.08</td>
<td>0.46</td>
</tr>
<tr>
<td>Usability (composite scale)</td>
<td>4.77 0.07</td>
<td>4.66 0.07</td>
<td>0.12</td>
</tr>
<tr>
<td>Usefulness (composite scale)</td>
<td>4.84 0.06</td>
<td>4.72 0.06</td>
<td>0.02</td>
</tr>
</tbody>
</table>

*adjusted for perceived vignette complexity, experience with patient conditions in the case vignette, and sequence in which the interface designs were evaluated.

Open-ended comments
The study participants echoed their appreciation for CareNexus in their comments and offered suggestions for improvements (Tables 5 and 6).

Table 5. Open-ended comments by participant roles.

<table>
<thead>
<tr>
<th>Role of participant</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Physician</td>
<td>“The care network was easy to use. It was extremely helpful to have the timeline to assess where the patient has been and where she is heading with her case.”</td>
</tr>
<tr>
<td></td>
<td>“This was quite easy to navigate.”</td>
</tr>
<tr>
<td></td>
<td>“Excellent tool.”</td>
</tr>
<tr>
<td></td>
<td>“Nicely divided into family, medical, and non-medical.”</td>
</tr>
<tr>
<td></td>
<td>“Timeline is fantastic.”</td>
</tr>
<tr>
<td></td>
<td>“It is actually helpful to know the missed appointments. I have to go to two different screens to see the missed ones in my current EMR.”</td>
</tr>
<tr>
<td>Care coordinator</td>
<td>“The visual timeline is a great way to help patients with appointments.”</td>
</tr>
<tr>
<td></td>
<td>“I like that it is all on one screen and there aren’t a lot of tabs to navigate through.”</td>
</tr>
<tr>
<td></td>
<td>“I like the color coding.”</td>
</tr>
</tbody>
</table>
Table 5 (continued). Open-ended comments by participant roles.

<table>
<thead>
<tr>
<th>Role of participant</th>
<th>Suggestions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Care coordinator</td>
<td>“I like that it is all in one place.”</td>
</tr>
<tr>
<td></td>
<td>“Nice interface compared to the current EMR.”</td>
</tr>
<tr>
<td></td>
<td>“I found the graphical visually more over stimulating or busy.”</td>
</tr>
<tr>
<td></td>
<td>“Well developed and user friendly. This app would replace our Excel registry. I can’t say enough positive about the app – love it.”</td>
</tr>
<tr>
<td>Parent</td>
<td>“I could use this on a regular basis finding what I need.”</td>
</tr>
<tr>
<td></td>
<td>“I currently do not have a tool for care networking other than a notebook.”</td>
</tr>
<tr>
<td></td>
<td>“Looks to be very exciting and useful.”</td>
</tr>
<tr>
<td></td>
<td>“CareNexus is very user friendly.”</td>
</tr>
<tr>
<td></td>
<td>“I love the display of goals and feedback.”</td>
</tr>
</tbody>
</table>

Table 6. Suggestions by participant roles.

<table>
<thead>
<tr>
<th>Role of participant</th>
<th>Suggestions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Physician</td>
<td>“It would be nice to connect the timeline events to notes.”</td>
</tr>
<tr>
<td></td>
<td>“Let the user choose the interface format: tabular vs. graphical.”</td>
</tr>
<tr>
<td>Care coordinator</td>
<td>None.</td>
</tr>
<tr>
<td>Parent</td>
<td>“It would be nice to have this tool available in languages besides English.”</td>
</tr>
<tr>
<td></td>
<td>“Being able to scan in documents, prescriptions, IEP documents would be helpful.”</td>
</tr>
<tr>
<td></td>
<td>“Reminders to schedule specialist appointments would be great.”</td>
</tr>
</tbody>
</table>

Discussion

Pediatric patients with special needs have large care teams with members in various roles corresponding to different aspects of patients’ lives, such as treatment and management of health conditions, developmental challenges, educational needs, and financial support. Previous research in the domain of patient care teams has addressed availability of care team-related information in the inpatient setting, supporting team work within the same care setting, and tailored applications for the needs of patients with certain conditions. However, less has been done for designing clinician- and patient-facing applications to address the needs of medically complex patients that cross the boundaries of specific settings and conditions. The goal of our research is to address this gap by designing, developing, and evaluating two alternative user interface designs for a prototype of CareNexus, a tool to visualize, understand, share, and manage care team related information for complex pediatric patients.

The study participants highly rated the feature set and the overall user interface of CareNexus in terms of usability and usefulness. These findings are important based on the technology acceptance model (TAM) which stipulates that perceived usefulness and ease-of-use are predictors of actual use. Several factors may have contributed to these findings, including deriving the requirements from a systematic information needs analysis, the early involvement of representative users, an iterative design approach based on the information foraging theory, Shneiderman’s visualization principles, and Nielsen’s usability principles. All users completed 100% of the usability tasks (time to completion for 19 out of the 22 tasks was less than 7 seconds). The problem-solving segment of our evaluation approach encouraged the users to quickly grasp the gist of the care network, identify the care team member(s) relevant to the problem, and access their contact information. The patient event timeline was the highest ranked component of the application. Users found it to be a very intuitive and quick way to gain shared team knowledge about recent visits, issues, and current status of the patient’s care. The tabular format of the care network was preferred over the graphical format, however the differences in the ratings were small. Participants liked both the formats and users should be allowed to choose between the two formats, per individual preference, as suggested by one of our physician participants. Further research is needed to investigate if our findings regarding tabular versus graphical displays generalize to other applications as well as the different factors that may influence display format preferences. Display of three separate networks and the associated color coding made it easier for the users to identify care team members. Having the contact information along with the preferred contact readily available can be very beneficial, especially in emergent situations which was described as a need for this cohort of patients. Sharing the goals and feedback provides a way to create a shared sense of common ground between the team members. Finally, care team actions enable processes that require shared responsibility between care team members ensuring that members of the care team (including the parents) are “on the same page”. Viewing the care team actions and their status is valuable for all team members and may reduce the need for time-consuming and sometimes unreliable person-to-person communication.
The results of our formative evaluation are promising and warrant future work on: 1) analyzing the information sources for the care team-related information displayed in CareNexus, 2) developing algorithms to automatically populate applications like CareNexus, and 3) integrating CareNexus into EHR and PHR workflows. Future studies should also focus on extending the findings in this study to other patient populations needing chronic care management.

Limitations
The case vignettes used in this study were adapted from the use cases identified by the HL7 Child Health work group. Although they are representative of clinical and patient conditions of children with special needs, it is possible that different conditions may require other design features. The formative evaluation assumes availability of accurate and up-to-date care team information. High-quality care team data may not be readily available in real world systems and that may influence users’ perception of usefulness. Also, CareNexus is designed to be used in tandem with EHR and PHR systems. Further studies should investigate the usability, usefulness, and efficiency of CareNexus integrated in the user workflow.

Conclusion
We describe the design and formative evaluation of two alternative user interfaces of CareNexus, a prototype of a tool to view, understand, share, and manage patient care team information. We followed an iterative design approach guided by the information foraging theory, information visualization principles, and user interface design heuristics. In addition, feedback from representative users was incorporated early into the design. Twenty-one users participated in the formative evaluation of the resulting graphical and tabular user interfaces. Users highly rated the usability, usefulness, and feature set of CareNexus, and were able to complete the usability tasks in a short amount of time. The tabular format was rated higher for usefulness but the difference was small indicating that the users liked both the formats. Tools that enable understanding the gist of a patient’s care network across organizational boundaries, the temporal nature of care team relationships, details of contact information, goals and feedback of those involved in the patient’s care, and ability to find providers to match patient/family needs have the potential to facilitate care coordination and team collaboration.

Acknowledgements
This investigation was supported by the University of Utah PHR and Population Health Research Foundation, with funding in part from the National Center for Research Resources and the National Center for Advancing Translational Sciences, National Institutes of Health, through Grant 5UL1TR001067-02 (formerly 8UL1TR000105 and UL1RR025764). We would like to acknowledge the HL7 Child Health work group for allowing us to adapt their use cases for the case vignettes used in this study. Finally, we would like to thank the socio-technical team at the Department of Biomedical Informatics, University of Utah, for their valuable feedback on the user interface designs.

References
8. O’Malley AS, Grossman JM, Cohen GR, Kemper NM, Pham HH. Are electronic medical records helpful


24. Nielsen J. Why you only need to test with 5 users. *Nielsen Norman Gr.*


Understanding Patient Questions about their Medical Records in an Online Health Forum: Opportunity for Patient Portal Design

Tera L. Reynolds, MPH, MA;1 Nida Ali, BA;2 Emma McGregor, BSE;2 Trish O’Brien;2 Christopher Longhurst, MD, MS;3 Andrew L. Rosenberg, MD;4 Scott E. Rudkin, MD, MBA;5 Kai Zheng, PhD1
1University of California, Irvine, CA; 2University of Michigan, Ann Arbor, MI; 3University of California San Diego Health System, San Diego, CA; 4University of Michigan Health System, Ann Arbor, MI; 5University of California Irvine Health, Irvine, CA

Abstract

There are many benefits of online patient access to their medical records through technologies such as patient portals. However, patients often have difficulties understanding the clinical data presented in portals. In response, increasingly, patients go online to make sense of this data. One commonly used online resource is health forums. In this pilot study, we focus on one type of clinical data, laboratory results, and one popular forum, MedHelp. We examined patient question posts that contain laboratory results to gain insights into the nature of these questions and of the answers. Our analyses revealed a typology of confusion (i.e., topics of their questions) and potential gaps in traditional healthcare supports (i.e., patients’ requests and situational factors), as well as the supports patients may gain through the forum (i.e., what the community provides). These results offer preliminary evidence of opportunities to redesign patient portals, and will inform our future work.

Introduction

The clinical data (e.g., laboratory results, clinician notes) contained in medical records have long been a product of, and central to, the interactions that take place between patient and physician in a clinical encounter. These data were traditionally created, updated, and used almost exclusively by physicians. However, research suggests that providing ready access to medical records facilitates patient engagement,1 which can lead to improved health outcomes and reduced costs.2 While patients have had the right to access these records since 1996,3 the process of requesting a paper copy has tended to be prohibitive and, once obtained, can soon become difficult to manage and are easily lost.4 To address these barriers, U.S. health IT policies such as Meaningful Use,5 and initiatives such as OpenNotes,6 have encouraged that patients be given direct, electronic access to their medical records. In response, many healthcare organizations have deployed patient-facing technology, commonly referred to as patient portals, connected to their electronic health record systems. The literature suggests that patient use of these portals may have numerous benefits, including improved quality of patient records,7 home monitoring infrastructure,8 satisfaction with patient-provider communication,9 and health status.10 However, despite these potential benefits, patients have not accessed the portals11 at the rates predicted based on the high level of interest.12 In fact, according to the 2015 U.S. Health Information National Trends Survey, only about 27% of Americans had accessed their health data online, including through patient portals, in the past year.13 Even more concerning is the significant disparities in use found among certain groups, including some minorities,11 adolescents,14 and incarcerated patients.15

So, why has ready access not resulted in more access? According to the literature, a key barrier to realizing the full potential of patient portals is the lack of useful information despite the abundance of data they make available. For example, a recent study found that many patients’ needs for online portals were not just for accessing the data, but for obtaining personalized and actionable knowledge.16 These needs, however, have not been adequately met in the pervasive patient portal design. Relatedly, low health literacy and numeracy, meaning a lower “capacity to obtain, process, and understand basic health information and services needed to make appropriate health decisions,”17 can be a barrier to patient portal use.18 Many portals present data to patients in the same or similar way that it is presented to healthcare providers; however, studies have shown that patients may have difficulty understanding medical jargons used in free-text document, such as clinician notes,18 and interpreting quantitative data, such as laboratory results.19 These findings suggest that current patient portals may have been designed at too high of a level of health literacy and numeracy, or do not provide enough, or the right, support to aid patient understanding.

At the same time, there is often not enough time during a clinical encounter to fully meet patients’ needs for knowledge. While it varies by healthcare setting (e.g., primary care, specialty care), there is evidence suggesting that, on average, physicians are spending less time with each patient – from over 20 minutes per patient in 1998, to
around 17, or even lower, in 2016.\textsuperscript{10,21} A related issue is that studies have found that some physicians frequently interrupt patients,\textsuperscript{22} which may make it difficult for patients to get their questions answered. These factors can also negatively affect the patient-physician relationship,\textsuperscript{23} which may reduce the likelihood that patients will trust their physician. In sum, while healthcare provider institutions are expanding patients' access to clinical data through patient portals, this technology often does not adequately meet patients' knowledge needs, and access to healthcare's traditional social supports,\textsuperscript{24} including informational, is shrinking.

Therefore, in order to actually use the available data for personalized decision-making, patients must often tap into supplemental sources of knowledge. Increasingly, this means searching online for general and/or individualized health information.\textsuperscript{25} In fact, according to one study, 72\% of U.S. internet users reported doing this.\textsuperscript{26} There are, of course, a variety of online resources of varying levels of interactivity and quality from which patients may choose, but health forums have been growing in popularity (e.g., MedHelp.org, CrowdMed\textsuperscript{26}). These forums provide users with a platform to ask their health-related questions, so that others (typically peer patients) can provide assistance.

Among the myriad of patient questions, we were particularly interested in questions accompanied by data that appeared to be directly copied from the individual's medical record, and pasted into the forum post (or in some instances transcribed; referred to just as copy/paste below). Examples of such data include laboratory results, vital signs, and excerpts from radiology reports and discharge summaries. This content is usually posted publicly, which provides an opportunity for researchers to understand the nature of patients' questions about their clinical data. This offers direct evidence of comprehension issues, as well as additional forms of support that may be needed. Similarly, the nature of the answers to these questions, especially those that patients find useful, point towards the supports those patients may gain through the forum discussions. While the insights from these analyses may have broader implications for patient-provider communication channels, in this study, we focused on their opportunities to improve patient portal design to better meet patients' needs.

In this study, we preliminarily assessed the feasibility of this approach by focusing on a particular type of clinical data, laboratory results, which are widely available and frequently viewed in patient portals,\textsuperscript{18} a common source of questions,\textsuperscript{27} and tend to be easier to identify in question posts. Similarly, we focused on MedHelp (medhelp.org), which is one of the most popular online health forums.

**Methods**

**Dataset:** All questions and threaded replies from MedHelp were downloaded in September 2016. This dataset contains over 2 million questions and over 8 million answers posted by over 2 million unique users on numerous health conditions (e.g., diabetes, asthma) and health-related topics (e.g., healthy recipes). Question posts on MedHelp are very diverse; for example, patients may present symptoms and/or laboratory results in order to determine whether a clinic visit is necessary, and/or request emotional support during a stressful time (see Figure 1).

**Identifying Potentially Relevant Posts:** Before we could address our research objective – to understand the nature of question posts containing laboratory results that seem to be copied/pasted from medical records, and their threaded replies – we first needed to identify relevant posts. We did this in four steps.

The goal of Step 1 was to iteratively develop a list of keywords, and determine the number of posts containing copied/pasted laboratory results (relevant posts) returned by each. To do this, we utilized the live MedHelp website. Specifically, for each keyword (e.g., “lab result help”), we reviewed the first 90 posts (3 pages) of the results to (a) count the number of relevant posts; and (b) examine the “Related Questions” section for additional keywords to test.

Since it is difficult to develop a comprehensive list of keywords, and missing keywords could result in missing relevant posts, in the second step, we sought to identify features (patterns) that distinguish relevant posts from other posts. To do this, we translated the three keywords with the highest number of relevant posts (“Lab,” “Blood work,” and “Profile”), and their common alternative spellings, into three queries in the Indri Query language.\textsuperscript{28} The final queries were run against an inverted index built over all MedHelp question posts. We then manually reviewed the retrieved results to identify the patterns indicative of copying/pasting (e.g., specific test name plus numeric result, such as TSH 0.11). During the review of the last 100 results (a total of 600 were reviewed), no additional general patterns were found, suggesting theoretical saturation had been reached.\textsuperscript{29}

In Step 3, we encoded the identified patterns as regular expressions in Python v2.7. These regular expressions were tested against the annotated corpus created in the second step, and iteratively refined until at least 80\% recall was achieved – i.e., if there are 100 relevant posts in the annotated corpus, at least 80 are retrieved.
In the fourth step, we applied the final regular expressions to the entire dataset, which returned 64,922 potentially relevant question posts – matched a pattern indicative of laboratory results copied/pasted from medical records.

Figure 1. Screenshot of a MedHelp forum post illustrating the copied/pasted laboratory results. Response to question superimposed and outlined in blue.

Analyses: We generated an initial random sample of 1,000 of the potentially relevant question posts, and then retrieved all their peer replies. At least two research assistants independently (1) reviewed posts for relevance, and (2) analyzed relevant posts, using a qualitative content analysis approach, until no new information was emerging (theoretical saturation). In addition, where possible, a “satisfactory” response(s) was identified based on the feedback of the person posting the original question (e.g., “Thanks for info explains a lot.”), and the codes assigned to these posts were eventually compared to the codes assigned to posts not identified as “satisfactory” responses. The content analysis team met weekly to discuss and merge category/theme lists; all disagreements were resolved through discussion. Importantly, if we had not reached theoretical saturation after reviewing all posts in the initial sample, we would have repeated this procedure until we reached theoretical saturation.

For (1), question posts were determined to be irrelevant, and therefore excluded, if they did not contain any laboratory results, or contained results that were not from the individual’s medical records (e.g., from a scientific study). Of the 400 question posts reviewed before theoretical saturation, 146 were relevant and, therefore, further analyzed (in 2). Likewise, when reviewing the threaded replies to relevant question posts, any duplicate posts or posts not attempting to answer the original question (e.g., using the thread to ask their own related questions) were excluded. Of the 500 reply posts reviewed before saturation was reached, 417 were eligible for further analysis (in 2), including 289 answers to questions and 128 replies from the original question poster.

Results

Question Posts

The reviewed sample of posts contained an average of 2.1 questions per post, with a range of zero questions (e.g., just providing an update) to six. Questions covered laboratory results not pertaining to any specified conditions, as well as to a wide range of identifiable conditions or concerns, including sexually transmitted infections (e.g., herpes); heart-related concerns (e.g., hypercholesterolemia); liver conditions (e.g., hepatitis); kidney and pulmonary function; hormone imbalances (e.g., testosterone); cancers (e.g., breast); and autoimmune (e.g. lupus) and thyroid disorders (e.g., hyperthyroidism). Furthermore, while some posts included questions exclusively about the laboratory results, many asked about laboratory results in the context of their medical history and/or symptoms. We present the results of this analysis in three sub-sections: Topics, Requesting, and Situational factors.

Topics: While some posts included questions pertaining to more than one topic, most (~71%) focused on one of the following topics: diagnosis/cause, management/treatment, laboratory report, test/diagnostic, risk, and prognosis (see Table 1). In addition, sub-topic categories emerged as more specific descriptors of patients’ questions. In this sample, questions were most commonly about medications, symptoms, and next steps. For example, the subject line of one post reads, “Please tell me what to do next.” In the body of this post, it becomes evident that she is requesting help “…pulling this [relevant medical history, laboratory test results, symptoms, etc.] info together…” in order to get a step closer to identifying the cause of her symptoms. Less common sub-topics include effect of treatment (e.g., potential adverse effect), comorbidities (i.e., existing diagnoses), timing (e.g., how far apart treatments should be), risk behavior (e.g., alcohol use with certain conditions), and life-style (i.e., diet and exercise).
Table 1. Topics of copying/pasting question posts, and representative quotations.

<table>
<thead>
<tr>
<th>Topics</th>
<th>Representative quotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diagnosis/Cause</td>
<td>“I am VERY afraid that hypercal. is being caused by malignancy, but the Endo. is not worried…”</td>
</tr>
<tr>
<td>Management/Treatment</td>
<td>“What should my Armour thyroid dose be? Do I need a T3 level? Should I also be on Cytomel...dose to start?”</td>
</tr>
<tr>
<td>Laboratory report</td>
<td>“What language would you expect to find in a report for a patient who has not been previously diagnosed with cancer via biopsy?”</td>
</tr>
<tr>
<td>Test/Diagnostic</td>
<td>“… To my knowledge it [Immunophenotype test] would get a stain of the cells in the sample and look for antibodies for HPV…and if it found some it would try to find the type or strain?”</td>
</tr>
<tr>
<td>Risk</td>
<td>“I have been advised to do RAI therapy, and told that if the uptake is the same, that the nodule will be the only area affected, but that there is a 10% chance of the whole thyroid being affected and the possibility of my going hypothyroid over the following year(s).”</td>
</tr>
<tr>
<td>Prognosis</td>
<td>“… Just received my week 6 results. HCV RNA PCR Taqman 2.0&lt;25 IU/ml detected. This was after 4 weeks with SOC and an additional 2 with BOC added… Is it looking good that I will beat this monster?”</td>
</tr>
</tbody>
</table>

Requesting: Several categories emerged describing what patients were requesting with their post to the MedHelp community; specifically, patients requested opinion, advice, generic information, emotional support, and personal experience (see Table 2). Requests for opinions and advice were by far the most common in this sample. Those requesting opinions tended to provide their laboratory results and, often, relevant medical history and symptoms, and ask the community to interpret it in some way (e.g., likely diagnosis). Those requesting advice were asking for actionable opinions, and a sub-set of these were asking for the community’s assistance in deciding between two or more, often treatment-related, options. In addition, less frequently, posters requested information that was not necessarily personalized to them; for instance, one patient wrote, “Looking for information about chest aches/pain…” Others included language indicative of distress, fear (e.g., ‘I am terrified”), or other emotions, and were categorized as requesting emotional support. Finally, some posters explicitly invited other patients to share personal experiences so that they could learn from them.

Table 2. Categories describing what patients are requesting, and representative quotations.

<table>
<thead>
<tr>
<th>Requesting</th>
<th>Representative quotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Opinion</td>
<td>“Now after showing these results to the doctor he simply increased my dosage of thyroxine sodium to 100mcg once a day. Can u please give me second opinion…”</td>
</tr>
<tr>
<td>Advice</td>
<td>“Would you recommend more testing?”</td>
</tr>
<tr>
<td>Decision support</td>
<td>“I’m vacillating between having the rt breast re excised &amp; await the biopsy results - Or just having a mastectomy w/o awaiting any further test results…Your advice is welcomed!!!”</td>
</tr>
<tr>
<td>Information (generic)</td>
<td>“what treatments are available?”</td>
</tr>
<tr>
<td>Emotional support</td>
<td>“I am really upset now to think I have something really wrong with me”</td>
</tr>
<tr>
<td>Personal experience</td>
<td>“I’d like to hear others experiences so I can better understand all this.”</td>
</tr>
</tbody>
</table>

Situational factors: Two main situational factors emerged; specifically, (1) whether patients are posting prior to discussing their results with a healthcare provider or after, and (2) whether they have a diagnosis or not. These situational factors, as well as sub-categories and representative quotations, are summarized in Table 3.

When it was possible to identify the motivation, patients posting questions prior to a healthcare consultation (also referred to as pre-consultation) did so for several reasons, namely waiting, preparing, and determining the need for an appointment. Many patients were waiting for an appointment with their physician, which often was not for several months, but wanted answers or to take action now. These posts tended to exhibit language indicative of distress (as described above). Others were preparing for their upcoming appointment – trying to obtain information that they could utilize in their consultation. In one such post, the patient wrote, “My doctor is very easy to work with but endocrinology is not his specialty (Family Practice). Please provide what information I would need to share
with my doctor for any changes/revisions in my current medication.” Still others were trying to understand their results in an effort to determine if they needed to make an appointment with a physician at all.

Similarly, patients posted questions about their laboratory results to Medhelp after discussing them with their physician (also referred to as post-consultation) for variable reasons, including seeking a second opinion, clarification/explanation, answers or a way forward when their physician is at a loss (referred to as ‘Doctor does not know’ in Table 3), and assistance in making a decision when they have been offered options. In this sample, seeking a second opinion was by far the most common reason for post-consultation posts – patients wanted another opinion on their physician’s interpretations and/or recommendations. In many cases, patients seem to doubt, disagree with, dislike, feel uncomfortable with, or mistrust their physician or their physician’s conclusions. In other cases, they received conflicting opinions (e.g., from different specialists). Other patients just wanted clarification or explanation of information that they discussed with their physician. For example, in one post the patient stated, “I asked questions, but I did not understand what deciding factor determines continuing treatment.” On the other hand, some patients reported that their physician did not know what their results meant or what to do next. Often these patients were still experiencing bothersome symptoms, and they seemed to be seeking fresh ideas. Finally, some patients’ physicians had given them options, and they were struggling to make a decision.

Table 3. Situational factor categories and sub-categories with representative quotations.

<table>
<thead>
<tr>
<th>Situation Factor</th>
<th>Representative Quotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Healthcare – Pre-consultation</td>
<td></td>
</tr>
<tr>
<td>Waiting</td>
<td>“… Can't get doctors appointment for 2 months. Any of your doctors have an ideas??????????”</td>
</tr>
<tr>
<td>Preparing for appointment</td>
<td>“… We r going to the Endo on Monday - what do I ask?”</td>
</tr>
<tr>
<td>Determining if need medical follow-up</td>
<td>“… She not nor am I a deliver in follow up visits unless something is wrong.”</td>
</tr>
<tr>
<td>Healthcare – Post-consultation</td>
<td></td>
</tr>
<tr>
<td>Second opinion</td>
<td>“I have been talking to another doctor… Can you tell me what you think?”</td>
</tr>
<tr>
<td>Clarification/ explanation</td>
<td>“… My MD stopped the Armour Thyroid… He also said the elevated T3 can have the same sxs as hypothyroidism. I've read alot, but I still don't understand how?…”</td>
</tr>
<tr>
<td>Doctor does not know</td>
<td>“… Although my numbers look OK I am still very tired... Doctors unsure of why or what to do…”</td>
</tr>
<tr>
<td>Options</td>
<td>“…I have a choice whether I want to go back on PegIntron... or Pegasys…”</td>
</tr>
<tr>
<td>Diagnosis status</td>
<td></td>
</tr>
<tr>
<td>Diagnosed: Not questioned</td>
<td>“new labs,Im still so frustrated… This is 2.5 grains of erfa and 6.25 mg of levo. A little better but still not good. I met with the pharmacist and my doctor last week…”</td>
</tr>
<tr>
<td>Diagnosed: Questioned</td>
<td>“The drs office said my results were great, but I dont think thats right…. Does this not suggest I am borderline in hyper or hypo?... My doctor is insisting I am depressed.”</td>
</tr>
<tr>
<td>Not diagnosed</td>
<td>“My husband has an abnormal bloodn test… could you let me know what could be his diagnosis?”</td>
</tr>
<tr>
<td></td>
<td>“For quite some time, I have had a number of “issues” and have not gotten to the bottom of it.”</td>
</tr>
</tbody>
</table>

Finally, three different diagnosis statuses were identified in the reviewed posts – diagnosed: not questioned, diagnosed: questioned, and not diagnosed (See examples in Table 3). Some patients posting questions had a diagnosis that they did not question (at least in the post), and were instead often questioning management/treatment aspects, prognosis, or another aspect. On the other hand, some patients had been diagnosed by their physician, but had doubts as to whether that was indeed the issue, and as such often asked diagnosis questions in their post. Finally, there were patients that did not have a diagnosis, and then this tended to be the focus of their post as well. Some were receiving abnormal test results for the first time, and others had been searching for a diagnosis for a while.
Answers

In this sample, there were an average of 2.0 replies per question post, with a range of 0 to 16. The replies that addressed the original question poster provided information, advice/suggestions, opinion, emotional support, and personal experience (See Table 4). They also requested more information from the question poster.

Information: The MedHelp community provided information on a number of topics, including diagnosis/cause, tests/testing, potential seriousness/risk, prognosis, management/treatment, and resources. More specifically, answers provided information on how a diagnosis is made, common symptoms associated with particular diagnoses, and common causes of abnormal test results and/or symptoms. In addition, information on tests/testing was offered, such as the reliability and accuracy of testing methods, reliability of different laboratory values (e.g., TSH), different options for tests, whether tests are invasive and any required preparations (e.g., fasting), and generally what laboratory results mean. Some answers also included information on potential seriousness/risk and prognosis. Furthermore, the community provided information on management/treatment, including general treatment approach for a given condition (e.g., “...doctors typically just monitor it, no real treatment needed.”), goals of treatment (e.g., management, cure), explaining or comparing options, treatment safety/risk, and new treatment options. Finally, answers offered social support and informational resources (e.g., pointing to websites or other MedHelp posts).

Advice: Answers also provided advice/suggestions related to physician/medical professional, addressing/treating health concern, further testing, and more information. Many reply posts suggested that question posters discuss their questions, and information provided by the community, with their physician. They also sometimes recommended that the question poster see a specialist, get a second opinion, or get a new doctor all together. In addition, the community provided advice/suggestions on addressing/treating health concerns, including whether or not treatment is needed, specific types of medication and dosages to try, changes in medication and/or dosage, suggestions for how to proceed with treatment (e.g., timing), and diet and supplementation recommendations. Some answers also recommended additional tests that they feel the question poster should consider. Finally, many answers suggested that the question poster get more information, including by using search engines, going to sites/resources that they felt were particularly helpful, and asking lots of questions of their physician.

Opinion: In addition, community members provided first and second opinions on topics, including primary concern, interpreting test results, possible diagnosis/cause, seriousness/risk, prognosis, and management/treatment. In cases where people had comorbidities or multiple pieces of information that conflicted, some answers provided opinions on the priority; for example, one community member said, “You're focusing too much on the test results and not enough on how you are feeling.” The community also provided interpretations of laboratory results, often including their opinion on possible or likely diagnosis/cause or prognosis given their interpretation of the results (and symptoms and history, when provided). Finally, answers provided opinions on management/treatment such as whether they think a proposed treatment is appropriate, treatment (often, medication) and/or dosage they feel would be better, and the best way to make treatment decisions for a given issue (e.g., certain laboratory value).

Emotional Support: Replies from the community also provided emotional support through empathizing, encouraging, offering well wishes and congratulations, reassuring, showing concern or caring, and by indicating that the question poster is not alone – others have felt or experienced the same. Many of these were responding to posts where there was language indicative of distress (as discussed above) or another emotion (e.g., cautious excitement); therefore, whether it was bad news (e.g., tests are positive for a certain condition) or good (e.g., tests indicate the poster is pregnant), most responded in an emotionally supportive way (e.g., “Don't give up!”). In addition, even if the question poster did not appear to explicitly request emotional support, answers would often still provide it.

Personal Experience: Answers provided personal experiences in order to offer reassurance, or another type of emotional support, as well as to provide information, opinions, or advice. They tended to use their experience to support a claim or recommendation. For example, one community member wrote, “I didn't hear a heartbeat until i was 10 weeks. So... try not to worry too much.” While another answer provided the following support for her suggestion that the problem may be pulmonary, “I just had 4 days of intense fatigue and being short of breath. It wasn't until the chest ache and coughing kicked in that I realized it was my asthma being a brat.”

More Information: Finally, community members often asked follow-up questions and requested that the question poster provide more or updated information. Such replies suggest that, given the information provided, they could not help or would be able to help more with additional information.
<table>
<thead>
<tr>
<th>Providing Information</th>
<th>Representative quotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diagnosis/cause</td>
<td>“There are many possible causes. Fatty liver, alcohol related liver disease, problems with the bile ducts are among the common causes for these elevations.”</td>
</tr>
<tr>
<td>Tests/Testing</td>
<td>“In addition to fluctuations in actual VL... the actual method of testing being done can produce different actual counts. Hence the sensitivity threshold at which they can measure.”</td>
</tr>
<tr>
<td>Potential seriousness/risk</td>
<td>“This condition does not have any serious sequelae.”</td>
</tr>
<tr>
<td>Prognosis</td>
<td>“The disease at this point is generally incurable.”</td>
</tr>
<tr>
<td>Management/ Treatment</td>
<td>“Cortisone creams are often prescribed for the patches of red, dry skin but will also leave you at risk for UV exposure in those areas for the rest of your life.”</td>
</tr>
<tr>
<td>Resources</td>
<td>“… see parathyroid.com for more info on this...”</td>
</tr>
<tr>
<td>Advice/suggestion</td>
<td>“Followup with your personal physician is essential.”</td>
</tr>
<tr>
<td></td>
<td>“Get a new, decent doc who understands the difference between IBS and post-gallbladder problems.”</td>
</tr>
<tr>
<td>Addressing/treating</td>
<td>“Take the Statin. It reduces the chance of having a heart attack even in people with normal cholesterol. And your LDL is high…”</td>
</tr>
<tr>
<td>health concern</td>
<td>“I'd push for hormonal testing - cortisol and deficiencies.”</td>
</tr>
<tr>
<td>Further testing</td>
<td>“Arm yourself with as much information as possible and make an informed decision.”</td>
</tr>
</tbody>
</table>

**Opinion (First or Second)**

| Primary concern | “I would think your experience of dealing with Narcolepsy, which is awful by the way, should be the primary concern here.” |
| Interpreting test results | “The low PTH suggests that the high Ca++ initially discovered was "non-parathyroid" origin but if it were a malignancy the Ca++ would typically be on the further rise.” |
| Possible/likely diagnosis/cause | “You have positive antibodies (autoimmune thyroid disease) - most likely this is early Hashimoto's.” |
| Prognosis | “I think you're going to beat it, yes I do and I haven't seen anyone relapse yet who has taken the triple therapy.” |
| Risk/seriousness | “Personally I would feel better if your CRP were also in the first quartile of risk (low risk <1.0)... facts are you have a high LDL cholesterol and a level of CRP that is above the lowest risk group.” |
| Management/ Treatment | “I think Vit D treatment is fine as long as Ca++ is monitored closely.” |
| Emotional support | “it just makes sooooo angry when docs act like that.I know if it were them feeling bad they would take the med!!” |
| Personal experience | “I have Geno 1A and also 1B and two doctors insisted on Intron.But again it's probably personal perspective.” |

**Sufficient answers:** The replies determined to be “sufficient answers,” based on the reply from the original question poster, suggested that the question poster felt somewhat reassured or relieved, perceived themselves to have an improved understanding, and/or that they had or were going to take a recommended action. The largest number of “sufficient answers” were in reply to thyroid- and liver-related questions; however, “sufficient answers” were identified in threads related to infectious diseases, pregnancy, autoimmune disorders, cancer, and unknown condition. “Sufficient answers” more commonly provided opinion and emotional support, and more specifically tests/testing information, as well as concern/caring and encouragement emotional support.
Discussion

The results of this study provide evidence that it is feasible to identify and characterize the nature of patients’ questions related to laboratory results, and to characterize the nature of the answers to these questions. Specifically, they revealed (1) a typology of patient confusion when viewing laboratory results, (2) patients’ social support needs, (3) the contexts of questions, as well as (4) the type of support that patients may gain, and find ‘satisfactory,’ from peers in online health forums. Similar categories have emerged from studies in the human-computer interaction and patient portal literature, beyond laboratory results and outside of peer-to-peer settings, suggesting that our findings are robust. However, our study had several unique aspects – including, a “target-rich” source of actual patient information and knowledge needs across multiple types of health conditions, and the classification of answers, which few studies have previously done – that resulted in a more thorough typology and novel findings. While our results are preliminary, they suggest potential opportunities to improve patient portals to better meet patients needs.

First, we found a typology of patient confusion about their laboratory results, which includes topics such as management/treatment. Although others have identified similar categories, their typologies were not as comprehensive. For example, when analyzing secure messages sent via a portal, Sun et al., found similar categories of questions aimed at patients’ healthcare teams, but we identified categories, such as risk, that they did not. We were likely able to more comprehensively identify types of patient questions because of our rich dataset, which uniquely situates this study to add to the current understanding of the types of patient questions related to their clinical data. For example, compared to Sun et al.’s study, we may have identified more clinical topics, because many patients seek information online first. In fact, Wright et al. reported that patients were much more likely to search the internet after viewing their problem list through a patient portal, rather than contacting their healthcare provider. As our results suggest, after searching online, some of these individuals may believe that there is no need to, or may choose not to, follow-up with their physician.

Second, we observed that patients had social support needs, such as informational, that they tried to meet through the forum. Other studies have reported similar categories; for example, Silence et al. analyzed advice solicitation in an online breast cancer support group, and identified five patterns, (1) requesting advice, (2) requesting opinion or information, (3) disclosing a problem(s), (4) announcing a plan of action, and (5) asking “anyone in the same boat?” Despite small differences, the overlap, as well as the fact that we identified our categories across a number of conditions, indicates robustness, and that our results may be relevant beyond this study setting and population.

Third, while other studies have peripherally reported situational factors, it emerged as a central theme in our study. For example, Powell et al. reported motivations for online information seeking, including perceived barriers to traditional information sources, as well as to seek reassurance, a second opinion, and a greater understanding. In addition, this study also found that most online health information seeking was associated with consultations, either in preparation for or to find additional information afterwards. In the context of our study, there appears to be some differences in the types of questions asked based on situational factors. We are in the process of extending this work by developing a model of the different health information needs in different situations.

Finally, while there is literature characterizing “best answers” on question and answer websites, to the best of our knowledge, our attempt to examine peer answers that patients perceive to be satisfactory in an online health forum is novel. The results of this pilot suggest that answers perceived to be “sufficient” by patients more often included opinion and emotional support, and they usually provided more specific information on tests and testing, as well as concern/caring and encouragement emotional support. This may be because these elements were better aligned with what question posters were requesting. In terms of the emotional support, it may also be that when answers showed concern/caring or offered encouragement, it elicited more positive responses from the question poster, regardless of whether it actually helped them gain a better understanding of their situations. Our future work will seek a more complete picture of the types of answers that patients find helpful and the underlying reasons.

Implications: These results provide preliminary evidence of opportunities to improve patient portal design. In terms of content, as discussed above, patients often search online first and may not follow-up with their healthcare provider; therefore, it is crucial to provide more of the information and knowledge that patients need at the point of viewing their laboratory results in patient portals. This could reduce the time and effort burdens often inherent in seeking medical information in the vast internet and, at the same time, ensure that they are consuming, and making decisions based on, accurate information carefully curated by their healthcare professionals. One way to provide more of the information patients needs may be better understanding differences in information needs based situational factors, which could offer an opportunity to provide more personalized content in patient portals, based on elements that are often available (e.g., appointment date). In terms of features, since patients tend to seek these
social supports in online communities, it suggests that they may be lacking in patient portals and their healthcare encounters, which could present opportunities for additional patient portal features. For example, supporting patients desire for personal stories from patients “in the same boat,” by allowing patients to opt-in to a social feature that enables finding similar patients and/or more fluid groupings based on similar information needs, could be beneficial.

**Limitations:** There are several limitations to this work. First, the emphasis was to comprehensively identify relevant themes and categories emerging from this rich dataset; therefore, posts were not double coded. Future work will include more quantitative analysis, and validation of qualitative findings. Second, while it is possible to identify posts with indicators that the original question poster perceived the answer to be helpful, there were a relatively large number of answer posts to which the original question poster did not reply. In addition, there may be other factors, independent of perceived usefulness of the answer, that affect how someone responds, including personality, culture, and level of emotional distress. Therefore, the results related to “sufficient answer” should be interpreted with caution, and warrant further investigation. Finally, in this pilot we focused on one type of clinical information and one online health forum. While there is support in the existing literature that these are crosscutting categories; however, the extent to which these results extend beyond this study setting is unclear. Based on the results of this pilot, in our future work, we intend to include the other, harder to identify and less frequently available types of clinical information (e.g., discharge summaries and clinical notes), as well as other online health forums.

Despite these limitations, this work provides direct evidence of patients’ comprehension issues related to their laboratory data. The results also highlight the types of supports they need, and the type that they are able to gain, and deem ‘satisfactory,’ through online health forums. This knowledge directly informs how patient portals can better provide social supports for patients. Finally, this study provides a foundation for our future efforts, including collaborating with patient portal stakeholders (e.g., patients, developers, administrators) to assess how this study, and follow-up studies, can be incorporated into the design and redesign processes to improve patient portals.

**Conclusion**

This pilot study provides evidence for the feasibility of (1) identifying online forum posts containing patients’ laboratory results copied/pasted from their medical record; (2) identifying question topics, patient support needs, and situational factors; and (3) characterizing what patient peers are providing and requesting in response. An extension of this work is currently underway, and will characterize the similarities and differences in information needs based on context. In addition, the results presented in this paper provide a foundation for future quantitative work involving computational methods. Ultimately, this line of work may lead to improving the design of patient portals, and the way that we present clinical data to patients. Such improvements may, eventually, lead to more patients being engaged in their healthcare and, as a result, improved health status and reduced costs.

**References**

1. Skipper J. *Individuals’ Access to Their Own Health Information.* The Office of the National Coordinator for Health Information Technology; 2012.


A Semantic Parsing Method for Mapping Clinical Questions to Logical Forms

Kirk Roberts, PhD, Braja Gopal Patra, PhD
School of Biomedical Informatics,
The University of Texas Health Science Center at Houston, Houston, TX

Abstract

This paper presents a method for converting natural language questions about structured data in the electronic health record (EHR) into logical forms. The logical forms can then subsequently be converted to EHR-dependent structured queries. The natural language processing task, known as semantic parsing, has the potential to convert questions to logical forms with extremely high precision, resulting in a system that is usable and trusted by clinicians for real-time use in clinical settings. We propose a hybrid semantic parsing method, combining rule-based methods with a machine learning-based classifier. The overall semantic parsing precision on a set of 212 questions is 95.6%. The parser's rules furthermore allow it to “know what it does not know”, enabling the system to indicate when unknown terms prevent it from understanding the question’s full logical structure. When combined with a module for converting a logical form into an EHR-dependent query, this high-precision approach allows for a question answering system to provide a user with a single, verifiably correct answer.

Introduction

The wealth of information available in the electronic health record (EHR) of a patient can be both a blessing and a curse. While having the most information possible for a patient stored in the EHR is important for the many uses of EHR data (record keeping, billing, legal, research, etc.), the large amount of information can create difficulties for clinicians when they need to quickly locate specific information. For this reason, much focus has been placed on presenting clinicians with the most important information, either visually1 or in text2, 3. This is no replacement, however, for mechanisms to efficiently access any piece of information. Largely this is accomplished by (often cumbersome) graphical interfaces for structured information and textual search for unstructured notes. Textual search offers notable advantages: it is often quick to enter a query and far less training is required than vendor- and provider-specific graphical interfaces. While some effort has gone into a more semantic search to find structured data elements4, these methods still largely rely on a keyword-based approach that cannot deeply grasp the user’s information need in a manner that presents a single, verifiably correct response. Instead, a search engine-style listing of ranked results is presented, burdening the user with determining the relevance of each result in turn.

Instead of keyword queries, which are often ambiguous in regards to their precise information need, natural language questions provide a clear and intuitive form to query EHRs. For this reason, observational studies of clinician needs often use natural language questions as their representation of choice5, 6. Due to the utility of expressing information needs as questions, numerous automatic question answering (QA) methods have been proposed for medical data7, combining natural language processing (NLP) and information retrieval (IR) methods. However, comparatively little attention has focused on QA for patient-specific data in the EHR. Furthermore, the QA methods that have been proposed for EHRs are largely analogous to the keyword-based semantic search methods for general-purpose IR queries. This limits their abilities to search over structured data and furthermore limits their ability to provide a single, verifiably correct answer. For example, given the question “What are her three most recent glucose measurements?” a clinician would ideally want a single result (e.g., “145, 139, 156”) that can be verified (e.g., by providing proof the question was properly understood and links to the lab results in the EHR interface). Further, if a clinician were to ask a question without an answer (e.g., the patient only had two glucose measurements), or if the system did not fully understand the question (either through human error, e.g. a misspelling, or lack of system functionality, e.g. inability to understand “three most recent”) then appropriate responses can be shown instead of providing a list of completely incorrect search results.

To provide such an answer, the natural language question must be converted to a complete, semantically-grounded logical form using a pre-defined set of logical operations. While semantic grounding task such as concept normalization8 and semantic relation tasks such as semantic role labeling9 provide partial information toward a complete semantic grounding, they are insufficient in capturing the full meaning of a question. The NLP task for transforming natural
language into a logical form based on pre-defined logical operations is known as semantic parsing. While semantic parsing is a well-studied task in NLP, almost no work exists on semantic parsing for a task like EHR QA.

This paper presents a semantic parsing method for understanding EHR questions. As one might imagine, converting any form of natural language to a completely structured form is a difficult, NLP-complete task. To avoid the difficult task of enumerating the lexicon of logical operations, some semantic parsing methods utilize ungrounded logical forms\(^{10,11}\), where the words in the question themselves become the logical functions. However, this is generally only used for searching unstructured data, or assumes words are easily mappable to the database structure. This can further be problematic when mathematical operations (e.g., $\min$, $\text{less than}$) or knowledge-based operations (is positive, is significant) are mixed in with logical operations that correspond to fields in the database. As such, for tasks such as this, pre-defined grounded logical forms are preferable, as they are easier to convert to structured queries. Manually enumerating a pre-defined lexicon for natural language is largely intractable, so most grounded semantic parsing techniques thus focus on short texts (such as questions) and small domains, such as U.S. geography\(^ {12}\) (e.g., What states border Texas?). Even under closed domains, oftentimes thousands of manually annotated $<\text{question, logical form}>$ pairs are necessary to train a semantic parser to achieve satisfactory results. Our method, therefore, differs from many existing machine learning (ML) based semantic parsers due to the complexity of the domain and the difficulty of obtaining so many annotations. This is done by relying on existing clinical NLP resources and incorporating several rule-based elements to improve its customizability. Instead of performing semantically-embedded syntactic parsing as done by many semantic parsers (e.g., using a combinatorial categorical grammar (CCG) on the raw question input), the approach proposed here first leverages existing clinical NLP methods to simplify the question. Second, the simplified question is converted to a dependency-based tree structure. Third, a lexicon is used to convert lexico-syntactic sub-structures into logical operations, resulting in an initial set of logical trees. Fourth, a set of rules transform and then filter these trees. Fifth, a ML-based classifier chooses the best logical form tree for the question. Sixth, a separate classifier identifies the relevant temporal intervals for medical concepts in the question. Finally, the best tree is turned into a flat logical form and items withheld during the simplification step are substituted back in. We use a logical form designed to concisely represent the question while also being easily convertible to EHR query standards such as Fast Healthcare Interoperability Resources (FHIR). Examples of such logical forms are:

<table>
<thead>
<tr>
<th>Logical Form</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\delta(\lambda x. \text{has_treatment}(x, C0852255, \text{status}))$</td>
<td>How low has her blood pressure been?</td>
</tr>
<tr>
<td>$\min(\lambda x. \text{has_test}(x, C0005823, \text{visit}))$</td>
<td>What is the trend in hemoglobin?</td>
</tr>
<tr>
<td>$\text{positive}(\lambda x. \text{has_test}(x, C022885, \text{visit}))$</td>
<td></td>
</tr>
<tr>
<td>$\text{has}(\lambda x. \text{visit}(x, C0005823))$</td>
<td></td>
</tr>
<tr>
<td>$\text{min}(\lambda x. \text{has_test}(x, C0005823, \text{visit}))$</td>
<td></td>
</tr>
<tr>
<td>$\text{trend}(\lambda x. \text{has_test}(x, C0005823, \text{visit}))$</td>
<td></td>
</tr>
</tbody>
</table>

This paper addresses the critical second through fifth step described above—the actual semantic parsing steps. The first step can leverage any number of existing clinical NLP systems (e.g., MetaMap\(^ {13}\), cTAKES\(^ {14}\)); the sixth step is an important but tangential task to the semantic parsing; and the final step is a deterministic process re-incorporating the information extracted during the first step. Evaluating the semantic parsing steps in isolation enables identifying key issues in converting a natural language into a logical structure without the difficulties involved in error propagation from upstream components. Optimizing the semantic parser in the presence of errors in the first step and classifying the appropriate time intervals in the sixth step are both left to future work.

Background

As previously stated, numerous methods have been proposed for medical QA\(^ {7}\). Again, the bulk of these methods are not focused on patient-specific EHR QA, but rather on general medical knowledge or the scientific literature. Some works focus on questions asked by clinicians, while others focus on consumer questions (see Roberts et al.\(^ {15}\) for a brief discussion of these systems). The key reasons for the separation include the linguistic differences between medical professionals and consumers\(^ {16}\) as well as the appropriate information source (e.g., consumers are less likely to likely to understand scientific literature articles). Further, while QA for EHRs is less studied, IR for EHRs has received more attention\(^ {6,17-20}\).

QA methods for EHR data vary along a spectrum from keyword-based IR approaches to deep semantic approaches, though the only work involving a deep semantic approach that we are aware of falls under the project discussed in this paper. The IBM Watson system\(^ {21}\) uses a passage scoring method. While this is similar to IR approaches, QA-specific
The input question (Q) is first split into multiple questions via the Syntax Decomposition module, then each of the $n$ output questions are separately run through the remaining pipeline. After Concept Normalization, the question is simplified (SQ) by replacing concepts with consistent placeholders. A dependency parse is then generated from SQ and run through the semantic parsing components.

The dependency tree is run through the Lexicon, which produces all $N$ possible Lexicon Match Trees. Each of these is run through both generation rules (increasing the number of trees) and filtering rules (pruning trees), resulting in $M$ Logical Trees. A Tree Classifier then selects the best Logical Tree.

After the semantic parser, the concept CUIs are re-substituted back into the logical form, the TIME attribute is classified, and finally a query module uses the logical form as a template for interacting with the EHR (e.g., through FHIR queries).

Our current project aims to overcome the limitations of existing EHR QA methods to enable a true natural language interface for structured data in the EHR. We have previously described the structure of the logical form as well as an annotated corpus of $<$question, logical form$>$ pairs. The corpus furthermore contains gold-annotated concepts and other manual annotations in a layer-wise fashion to enable training the semantic parser on correctly simplified questions. There are three annotated layers: (1) a syntactic simplification step (i.e., question decomposition), (2) a concept recognition and normalization step, and (3) a logical form step (i.e., semantic parsing). We have further proposed an automatic method for syntactic simplification as well as an automatic method for distinguishing patient-specific questions (i.e., answerable with the EHR) and other types of questions (i.e., answerable with many of the existing QA approaches) so that multiple QA systems can be integrated into a single interface. In this work, we focus on the most critical NLP component of the overall project: the semantic parser that converts natural language into a logical form. Future work will cover the remaining tasks, such as mapping logical forms to FHIR queries for direct integration of a QA system into EHRs.

Materials and Methods

The overall QA framework for this project is shown in Figure 1. Most of the following subsections except the semantic parsing subsection describe elements of this framework at a high level, motivating certain choices made in semantic parsing and providing references to already-published descriptions.

1. Data

The dataset consists of 212 manually-annotated questions originally collected from actual clinicians by Patrick & Li. The dataset contains gold-annotated concepts and other manual annotations in a layer-wise fashion to enable training the semantic parser on correctly simplified questions. There are three annotated layers: (1) a syntactic simplification step (i.e., question decomposition), (2) a concept recognition and normalization step, and (3) a logical form step (i.e., semantic parsing). We have further proposed an automatic method for syntactic simplification as well as an automatic method for distinguishing patient-specific questions (i.e., answerable with the EHR) and other types of questions (i.e., answerable with many of the existing QA approaches) so that multiple QA systems can be integrated into a single interface. In this work, we focus on the most critical NLP component of the overall project: the semantic parser that converts natural language into a logical form. Future work will cover the remaining tasks, such as mapping logical forms to FHIR queries for direct integration of a QA system into EHRs.
function max. Most questions, roughly 97%, have only one λ-statement, while multi-λ questions can be particularly complicated for semantic parsers. For instance:

\[
\delta(\lambda x. \text{has treatment}(x, \text{C0087111}, \text{visit}) \land \delta(\lambda y. \text{is response}(x, y)))
\]

This logical form corresponds to the question “Did she have a reaction to the treatment?”. The second λ-statement is nested inside the first, essentially testing every treatment event \(x\) (\text{C0087111} is the high-level code for treatments) to see if any other event (denoted \(y\)) is a response to event \(x\). There is insufficient space here to discuss the reasoning behind every decision for the logical form annotations. We thus refer the interested reader to our previous work describing these decisions\textsuperscript{15,23}.

2. Initial Question Processing

The important insight behind the layer-wise approach is that while the concepts are present in the final logical form, they do not affect the structure of the logical form. As a result, the semantic parser need not understand the differences between “diabetes” and “hypertension”, only that “Does he have PROBLEM?” corresponds to the logical form \(\delta(\lambda x. \text{has problem}(x, \text{CUI}, \text{TIME}))\), where the UMLS code for the problem is substituted into the place of CUI. For this reason, the first two layers are referred to as simplification steps, as they reduce both the lexical and syntactic complexity of the question. Further, these steps correspond to clinical NLP tasks for which there is existing data. As a result, far less training data for the semantic parser is required. The layer-wise annotation on this dataset furthermore enables experimentation using gold standard simplification, isolating just the semantic parsing.

While the automatic methods behind these steps are beyond the scope of this paper, as previously stated, existing methods can largely be leveraged to perform this processing. A syntactic decomposition method, for instance, was proposed by Roberts\textsuperscript{24}. Concept recognition and normalization has been extensively studied, with both ML-based methods based on the SemEval task focused on disorders\textsuperscript{8} as well as rule-based methods such as MetaMap\textsuperscript{13}.

The final step prior to the semantic parser, which is performed automatically in this work, is the dependency parse. The Stanford dependency parser\textsuperscript{26} operates on the simplified questions. To ensure a proper dependency parse, the concepts in the simplified question are substituted with a part-of-speech appropriate replacement word. For example: a noun concept is replaced by the word “concept”; a past-tense verb concept is replaced by the word “ate”; and the reference to the patient is replaced by the pronoun “he” or “his” (if possessive). This replacement ensures a proper dependency parse on the simplified question, since replacing a phrase with a word from a different part-of-speech would likely result in an erroneous dependency parse. The original placeholders are then substituted back into the output dependency tree (e.g., “ate” is changed back to “concept” and “he” is changed back to “patient”).

To demonstrate the full extent of processing prior to the semantic parser, consider the following two questions:

**Lexical:** When was the patient most recently dialyzed? What was her lowest and highest blood sugar level?

**Syntax:** When was the patient most recently dialyzed? What was her lowest blood sugar level?

**Concept:** when was patient most recently vbd:procedure what was pos:patient lowest nn:procedure

**Dependency:**

![Dependency Diagram]

Note that the right-hand question is split into two questions, with the results of only one being shown here.

3. Semantic Parsing

The semantic parsing components are shown inside the box in Figure 1. The first component is a dependency tree-based lexicon. Note that most publicly available semantic parsers\textsuperscript{27,28} make use of a similar lexicon: while a lexicon can be learned directly from \(<\text{question, logical form}>\) pairs, such systems generally prefer the rely on lexicons for increased precision. In our case, since a high precision semantic parser is crucial, we thus follow suit.
The tree-based lexicon maps nodes and edges in the dependency tree to logical operations. The simplest case are individual nodes (i.e., single words/concepts):

<table>
<thead>
<tr>
<th>Node</th>
<th>Logical Operation</th>
</tr>
</thead>
<tbody>
<tr>
<td>concept</td>
<td>( \lambda x. \text{has}_{\text{concept}} )</td>
</tr>
<tr>
<td>lowest</td>
<td>( \text{min} )</td>
</tr>
<tr>
<td>temporal_ref</td>
<td>( \text{time}_{\text{within}} )</td>
</tr>
<tr>
<td>concept(der:the)</td>
<td>( \lambda x. \text{has}_{\text{concept}} )</td>
</tr>
<tr>
<td>done(aux:have, auxpass:been)</td>
<td>( \delta )</td>
</tr>
<tr>
<td>much(advmod:how)</td>
<td>( \text{sum} )</td>
</tr>
<tr>
<td>received(dobj:concept)</td>
<td>( \lambda x. \text{has}_{\text{concept}} )</td>
</tr>
</tbody>
</table>

Where null matches indicate words not useful for semantic parsing, but possibly useful for downstream tasks (e.g., “yet” is useful for time classification). Multi-word matches are also common. They are specified with the head node and specific edge types to dependent nodes:

<table>
<thead>
<tr>
<th>Multi-word Match</th>
<th>Logical Operation</th>
</tr>
</thead>
<tbody>
<tr>
<td>concept(der:the) ( \rightarrow ) ( \lambda x. \text{has}_{\text{concept}} )</td>
<td></td>
</tr>
<tr>
<td>done(aux:have, auxpass:been) ( \rightarrow ) ( \delta )</td>
<td></td>
</tr>
<tr>
<td>much(advmod:how) ( \rightarrow ) ( \text{sum} )</td>
<td></td>
</tr>
</tbody>
</table>

Note that while the overall focus of the system is on precision, lexicon building invariably must focus on recall. For this reason, multiple lexicon entries are often necessary for the same word (e.g., [was ⇒ null], [was ⇒ \( \delta \)]). The result is that most dependency trees have at least one overlapping lexicon match. When this happens, all possible combinations of lexicon matches are generated such that every node in the dependency tree is covered by exactly one lexicon match. A new tree is then created with nodes corresponding to the dependency relations (nodes and edges in multi-word matches are collapsed into a single node). Leaf nodes with null values are also discarded. These trees are referred to as lexicon match trees: the nodes contain logical operations, but the edges between those operations do not necessarily correspond to relations in the final logical form. The trees for the above examples include:

**Lexicon Match Tree:**

Note that while the two initial questions are quite different on the surface, by now the structures are quite similar.

At this point in the processing, questions without any lexicon match trees are considered unanswerable, as this means some part of the question has an unknown logical function. This is the source of the system’s high precision design: we would prefer the system returns answers to fewer questions than risk returning incorrect answers. Incorrect answers would result in diminished trust in the system by clinicians. When a question is unanswerable, the user can be presented with the unknown word(s) and be asked to rephrase the question.

The next component converts each lexicon match tree into zero or more logical trees. A set of rules using a grow-and-prune strategy is used. First, two “grow” rules generate new trees by manipulating related nodes in the initial lexicon match tree:

1. **FlipRule**: if a parent has one child, create a new tree with parent and child flipped (e.g., the tree \([A \rightarrow B \rightarrow C \rightarrow D]\) would also result in \([A \rightarrow C \rightarrow B \rightarrow D]\)).
2. **PromoteRule**: if a parent has more than one child, promote each child in a new tree (e.g., \([A \rightarrow B, C]\) would also result in \([A \rightarrow B \rightarrow C]\) and \([A \rightarrow C \rightarrow B]\)).

Trees generated by these rules are also run through the same rules, resulting in a final set of trees not only resembling the original lexicon match tree (which frequently contains nodes with multiple children), but also every possible unary tree (which are more typical of the logical forms in the dataset). Next, the “prune” rules remove invalid trees:

1. **NullRule**: null leaves are removed, and trees with null non-leaf nodes are filtered.
2. **TypeRule**: Every logical function has pre-defined input and output types. For instance, \( \lambda x. \text{has}_{\text{concept}} \) has no input and returns an \text{EVENTSET}; \text{latest} and \text{max} take in an \text{EVENTSET} and return an \text{EVENT}; \delta takes in an \text{EVENTSET} and returns \text{TRUE}/\text{FALSE}; and \text{time} takes in an \text{EVENT} and returns a \text{TIME}. Using these types, any tree that has an incompatible parent-child relationship (parent’s input must equal child’s output) is filtered.

These rules drastically reduce the number of trees, as the grow rules produce many invalid trees. As the trees are now compatible with logical forms, they are referred to as logical trees. Logical trees for the previous examples include:
Logical Tree:

```
  time
    latest
      λx.hua_concept
  min
    latest
      λx.hua_concept
```

This tree structure is useful for downstream processing, however, so they are not yet “flattened” into logical forms.

The final step in semantic parsing is to choose between the remaining logical trees. There are several statistical factors that dictate the true logical forms. The two most important are that some lexicon matches are more likely than others (e.g., “was” is more likely to be null than δ), and some arrangements of the logical form are more likely (corresponding to parent-child relations in the logical tree). For this reason, a ML-based approach using a support vector machine (SVM) is used. The two statistical factors described are the first two features considered by the SVM, along with a third feature emphasizing the importance of the top-level logical operation. The three features are:

1. **LexiconMatch**: IDs of the lexicon matches that were used to generate the lexicon match tree that, in turn, was used to generate the logical tree.
2. **ParentChild**: All parent-child pairs in the logical tree, including identifying the root and leaf nodes.
3. **Stem+Root**: The question stem (e.g., what, how) combined with the logical operation of the logical tree’s root node. For example, what-latest, how-many-count, when-time.

The SVM classifies all remaining logical trees for a question. The tree with the highest positive confidence is then selected as the output of the semantic parsing phase.

**4. Subsequent Question Processing**

After semantic parsing, several more steps are necessary. First, the TIME element needs to be classified, which can be done using traditional ML-based multi-class classification. Second, the concepts simplified after concept normalization need to have their CUIs inserted into their proper place (easily done with multiple concepts by tracking which logical tree node corresponds to which concept). This will produce the final logical form, which can be represented as a flat string.

To query an EHR, a query module needs to convert the elements in the logical form to either EHR-specific queries or functions within the module. For instance, given the logical form `time(latest(λx.has_test(x, C0392201, visit)))`, the query module might convert `λx.has_test(x, C0392201, visit)` into a FHIR query (likely multiple queries) to find all the blood glucose measurements occurring during the patient’s visit. Each of these results would be represented as an event, and the full set of events would be passed to the latest function to find the most recent matching event. Finally, that event would be given to the time function to return just that event’s timestamp.

Using this approach, a QA system can provide a single answer (resulting from the structured query and logical functions) that is verifiably correct in that a visualization of the logical form (for instance, see Figure 2) can be provided to the user to verify the system properly understood the question. If the question understanding was accurate, the answer should therefore be correct (at least from the NLP perspective, errors in data entry and EHR querying notwithstanding).

**Results**

We evaluate the semantic parsing accuracy on the questions with a leave-one-out validation to maximize data use (the small number of questions and features means this is still quite fast, around 15 seconds). Table 1 shows feature experiments to demonstrate the importance of each of the three features, along with several baselines. All methods reported in Table 1 use the same lexicon and logical tree rules, only the tree classifier is altered. The Stem+Root feature, which simply looks at the question stem and the root logical operation, has the highest performance. This is likely due to the type-based filtering rules: if the top-level logical operation is correct, this limits the potential children. Additionally, so many logical forms start either with latest or δ, after which the amount of ambiguity decreases, and the Stem+Root feature can identify high-likelihood root operations (e.g., is-delta, what-latest) and eliminate unlikely operations (e.g., what-delta, did-latest).
<table>
<thead>
<tr>
<th>System</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random Baseline</td>
<td>39.1%</td>
</tr>
<tr>
<td>Largest Tree Baseline</td>
<td>42.5%</td>
</tr>
<tr>
<td>Smallest Tree Baseline</td>
<td>48.9%</td>
</tr>
<tr>
<td>LexiconMatch</td>
<td>74.8%</td>
</tr>
<tr>
<td>ParentChild</td>
<td>73.1%</td>
</tr>
<tr>
<td>Stem+Root</td>
<td>78.2%</td>
</tr>
<tr>
<td>LexiconMatch, ParentChild</td>
<td>86.7%</td>
</tr>
<tr>
<td>LexiconMatch, Stem+Root</td>
<td>88.8%</td>
</tr>
<tr>
<td>ParentChild, Stem+Root</td>
<td>88.1%</td>
</tr>
<tr>
<td>LexiconMatch, ParentChild, Stem+Root</td>
<td>95.6%</td>
</tr>
</tbody>
</table>

**Table 1:** Experiments for tree classifier using baseline methods and feature combinations.

<table>
<thead>
<tr>
<th>System</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>+ All Rules</td>
<td>95.6%</td>
</tr>
<tr>
<td>- FlipRule</td>
<td>56.8%</td>
</tr>
<tr>
<td>- PromoteRule</td>
<td>72.1%</td>
</tr>
<tr>
<td>- FlipRule, PromoteRule</td>
<td>50.7%</td>
</tr>
<tr>
<td>- NullRule</td>
<td>95.6%</td>
</tr>
<tr>
<td>- TypeRule</td>
<td>89.8%</td>
</tr>
<tr>
<td>- NullRule, TypeRule</td>
<td>89.9%</td>
</tr>
<tr>
<td>- All Rules</td>
<td>49.7%</td>
</tr>
</tbody>
</table>

**Table 2:** Experiments removing various logical tree rules using best tree classifier.

Table 2 shows experiments to assess the impact of logical tree rules on the final result. Performance is hurt dramatically without the generation rules, achieving 56.8% and 72.1% without the FlipRule and PromoteRule, respectively. Removing both rules results in a performance of 50.7%. The filtering rules have far less of an impact on the final result. This makes sense as removing generation rules fundamentally limits recall. The filtering rules, however, are designed to help with precision, but the ultimate decision rests with the tree classifier. The NullRule, which removes logical trees with null nodes, removes trees that are easily detected by the SVM classifier, and so there is no effect on the final score. The TypeRule has some effect on the system, as without it the accuracy drops to 89.8%. While the classifier should ultimately learn inaccurate parent-child relationships, they can be quite sparse for some relations. So having a hard rule that forces agreement is useful. It also greatly speeds up the system: without the type rule, there is a mean of 206 and a median of 13 logical tree candidates per question. With the TypeRule, that drops to a mean of 11 and a median of 2 candidates. This further has the positive effect of reducing data imbalance, which likely has a positive impact on training the SVM.

To demonstrate the effect of data size, Figure 3 shows the leave-one-out accuracy (using all three features) with increasing percentages of the total data. With small amounts of data (<20%), there are random fluctuations in performance, as would be expected, but afterward there is a steady climb in performance. It appears likely that adding more data would continue to benefit the semantic parsing performance.

**Discussion**

The high performance of the semantic parser (>95%) on the annotated dataset demonstrates the potential for a high-precision question understanding system, converting natural language to a fully-structured logical form. However, the use of the lexicon means that the results reported above are high estimates of the performance of the semantic parser on unseen questions, since almost all questions had the necessary lexicon entries to generate their proper logical form. For this reason, the 95.6% accuracy is less of a measure of system accuracy than it is a measure of system precision: some unseen questions will contain unseen words, which have no lexicon match and therefore will return no answer at all (i.e., a false negative). Precision does not measure false negatives, however, so the results above are a decent estimate for precision. Recall, however, while not a focus of our method, would still be a concern if it were so poor that hardly any questions would be answered. The primary limitation of recall in our system is the lexicon: if it fails to find an entry for a term in the question, or the entries it does find fail to produce a valid logical tree, then no answer will be provided. Figures 4 and 5 provide a sense of the distribution of the entries in the lexicon. Figure 4 shows that the lexicon has a very long tail: around 20 lexicon entries are highly used when generating gold logical trees, while
dozens of entries are only associated with a single gold logical tree. Figure 5 shows that 72 questions rely on at least one lexicon entry that is not used for any other question. It is difficult to estimate recall exactly from these figures, but it would not be surprising if recall on unseen questions was in the 60-80% range, far below the high precision of the semantic parser. However, this assumes that users ask questions without knowledge of the system: over time a user would likely come to understand how to phrase questions based on the parser’s vocabulary.

Even on the annotated data for which we have a complete lexicon, however, the semantic parser is not completely perfect, so some discussion of the types of errors it makes is worthwhile. The most common type of error revolves around rare logical operations. For instance, the question “Is his Computed Tomography positive?” has the gold logical form \( \text{is\_positive(latest(\lambda x.\text{has\_concept}))} \). The predicted logical form, however, was \( \delta(\lambda x.\text{has\_concept} \land \text{is\_positive}(x)) \). The predicted form more closely resembles other uses of \( \text{is\_positive} \) in the data, but unfortunately in this case would actually answer the question “Does the patient have a positive Computed Tomography”, i.e., it will return true if any prior test is positive, whereas the original question is only interested in whether the most recent is positive. In order to maximize precision further, these rare predicates either need to be removed or generalized to a common concept.

Other, more difficult errors, involve some degree of ambiguity or implied answer type. While the intentional decision was made during annotation to take the questions as literally as possible, it was not possible to remove all real-world knowledge of the true intention of the question. For instance, the question “How high is his Troponin?” has the gold logical form \( \text{latest(\lambda x.\text{has\_concept})} \), which essentially is just asking for the result of the latest Troponin test. The phrase “How high” is roughly analogous to “What is the value of” in the question. While that interpretation can be added to the lexicon, the semantic parser still focuses on a strict interpretation of the phrase “How high” and predicts the logical form \( \text{max(\lambda x.\text{has\_concept})} \), i.e., “What is the highest Troponin level?”. Similarly, “How much sputum does he have?” is interpreted as the summation of all sputum tests (\( \text{sum(\lambda x.\text{has\_concept})} \)) and not simply the result of the most recent test (\( \text{latest(\lambda x.\text{has\_concept})} \)).

Many of these errors can be tolerated if the answer is placed in the proper visual context. For instance, the incorrect interpretation of the question about sputum level should result in a graphically produced answer that indicates there was a misunderstanding. In this case, not only showing a graphical query representation, but also having the actual answer that shows not just the sum, but all of the sputum tests below the sum, would be a good visual clue to the clinician that the question was misinterpreted. This is yet another case of how placing a noisy NLP system within a proper visual context...
interface can prevent NLP errors from propagating to clinical decision-making, and thus retain the trust of clinicians. Unlike most NLP systems, however, the logical forms created by the semantic parser provide an incredibly useful structure to visually represent information. NLP systems that rely on bag-of-words text categorization, for instance, are difficult to visually interpret beyond highlighting noteworthy words. Instead, the logical form, when displayed graphically as in Figure 2, allows the user to not only know an error occurred almost immediately, but also gives a strong clue of how to re-phrase the question in a less ambiguous manner. Such a trust-building combination of NLP and interface design can lead to a greater use of NLP to directly benefit clinicians in a real-time situation. Currently, however, most clinical applications that involve NLP focus on its offline use for research and administrative purposes.30 For the most part, this is largely due to the fact that NLP systems are not sufficiently precise to risk introducing errors into clinical decision-making. It is our hope, however, that methods like the one proposed here focus sufficiently on precision, while still offering substantial utility, that their adoption by clinicians will eventually take hold.

Conclusion

We have presented a semantic parsing method to convert clinical questions to a structured logical form. The method combines rule-based techniques (a lexicon and generation/pruning rules) to create tree-based structures that are then classified by a SVM classifier to select the tree corresponding to the best logical form. This semantic parsing method achieves an accuracy of 95.6% on a manually annotated set of questions, though as discussed this is probably better interpreted as the system’s precision. The method is intentionally high-precision, sacrificing recall in order to preserve user trust. Not only is the logical form approach an effective way to retrieve a single, verifiably correct answer, it also provides a useful opportunity to visualize the NLP prediction in a way that can further increase user trust. Future work on the semantic parsing includes expanding the number of questions, including questions with additional complexity. For the project as a whole, future work will focus on integrating the semantic parsing component into an end-to-end QA system, from natural language question to FHIR query.

References


Is the Application of SNOMED CT Concept Model sufficiently Quality Assured?

JM Rodrigues MD PhD1,2, S Schulz MD3, B Mizen MSc1, A Rector MD PhD4, S Serir MSc1

1INSERM LIMICS UPMC UP 13 Paris, France; 2University of Saint Etienne, CHU, Department of Public Health and Medical Informatics, Saint Etienne, France; 3Institute for Medical Informatics, Statistics and Documentation, Medical University of Graz, Austria; 4University of Manchester, UK

Abstract

The terminological content of SNOMED CT, the world’s largest clinical terminology is linked to description logics expressions, which give support to consider SNOMED CT a formal ontology. The Terminology Quality Assurance (TQA) of such a terminology resource is hampered by errors in modeling, which act as a barrier for the successful use of electronic health records to ensure semantic interoperability. One application case is the new version of ICD, now in its pre-final form, the content of which is based on a subset of SNOMED CT. The ongoing alignment exercise has highlighted significant modeling issues in more than one third of cases that contrasted SNOMED CT concept model instances with the intuitive meaning given by their Fully Specified Names or synonyms lexically mapped to ICD-11 class names. We recommend prioritizing SNOMED CT TQA on the subset of the core SNOMED CT content to constitute the always true common ontology between SNOMED CT and ICD-11.

Introduction

SNOMED CT1 is the world’s largest clinical terminology with about 300,000 representational units, called concepts. They are linked to terms in several languages, with English providing the highest coverage. All lexical entities in SNOMED CT are named Terms. They include Fully Specified Names, Preferred Terms, and Synonyms. SNOMED CT concept are described and defined by expressions following a formalism called Compositional Grammar (CG)2, which can be interpreted according to description logics (DLs) which allows SNOMED CT to be considered a formal ontology. SNOMED CT is introducing itself as “a terminological resource which consists of codes representing meanings expressed as terms, with interrelationships between the codes to provide enhanced representation of the meanings.”

Nearly half of SNOMED CT’s concepts constitute the large Procedure and Clinical Finding hierarchies. They are not covered by any Terminology Quality Assurance (TQA) mechanisms apart from an abstraction network derivation methodology which focuses on internal consistency only3,4,5.

We here propose to assess the quality of SNOMED CT concept model by assessing its ability to support a formalization based not only on the closed world of SNOMED or ICD but on the open world assumption6 by taking the external example of SNOMED CT ICD-11 semantic alignment work.

As explained in7 this semantic alignment needs several steps: step 1 is a lexical map between the ICD-11 class name with textual definition and SNOMED CT Fully specified name or synonyms: step 2 is a tentative match between ICD-11 textual definition and SNOMED CT Compositional Grammar (CG) expressions2. Further steps as queries on pre-coordinated or post-coordinated SNOMED CT concepts to take care of ICD-11 inclusions and exclusions rules are out of the scope of this paper.

The purpose is to assess the quality of the inferred CG expressions candidates to the common ICD-11 SNOMED CT ontology by measuring the need of modification of CG expressions in order to represent ICD 11 class definitions if and only if there is a possible lexical map between ICD 11 class and pre-coordinated or post- coordinated SNOMED CT concepts. This is justified by the seemingly high rate of the same strings of characters in both systems: it allows to avoid the more complex issues when the same meaning is expressed by different strings of characters.
We do not consider the current pre-final version of ICD-11 as a gold standard for healthcare terminological resources and the total or partial omission of a SNOMED CT concept that seems necessary to ICD 11 external authority. We neither assess the clinical consistency of ICD 11 textual definitions. We assess only the ability of the existing CG expression(s) to represent the ICD-11 textual definition when lexically mapped to SNOMED CT concept description terms or if necessary to modify this concept model instance according to assumptions, rules, and standards that are specific to the SNOMED CT concept model.

**Material**

The ICD-11 Mortality Morbidity and Standard (ICD 11-MMS) is extracted from the WHO Browser (date: 31 Jan 2017). The semantic alignment principles explained above are performed with SNOMED CT content as displayed by the IHTSDO browser, corresponding to the SNOMED CT release of the same date, which includes:

1. Number of direct parents
2. Number of role groups
3. Number of children
4. Whether the concept is primitive or fully defined
5. Depth (distance from root)
6. Compositional grammar (CG) expression

The SNOMED CT specific compositional grammar (CG) is a set of rules that govern the way in which SNOMED CT expressions are represented as a plain text string. It is a syntax that supports a wide range of clinical meanings to be captured in a clinical record, without requiring the terminology to include separate concept for each case. Clinical expressions using SNOMED CT concepts can be of two types: pre-coordinated expressions, which use a single SNOMED CT concept identifier and post-coordinated expressions, which contain more than one SNOMED CT identifier.

The IHTSDO browser provides the representation in plain text and in a diagram. For example, the SNOMED CT concept 31978002 *Fracture of tibia*, a fully defined pre-coordinated concept, is defined using a CG expression as being equivalent to *Injury of tibia* and *Fracture of lower leg*, with *Associated morphology Fracture* and *Finding site Bone structure of tibia*. This definition is shown below in Table 1 using the SNOMED CT CG formalism, together with the transcription into DL Manchester syntax and in Figure 1 for the CG diagram displayed by the browser.

To enhance readability, concepts are noted in *Italics* and relations are noted in **Bold**.

**Table 1.** Fracture of Tibia in Compositional Grammar (CG) and Manchester syntax

<table>
<thead>
<tr>
<th>Concept ID</th>
<th>CG Expression</th>
<th>Manchester Syntax</th>
</tr>
</thead>
<tbody>
<tr>
<td>31978002</td>
<td><em>Fracture of tibia</em></td>
<td><code>equiv Fracture of tibia fracture of lower leg injury of tibia bone structure of tibia morphology fracture attribute site bone structure of tibia morphology attribute site some</code></td>
</tr>
</tbody>
</table>

**Figure 1.** The SNOMED CT concept Fracture of tibia (ID = 31978002) as a CG diagram

The purpose of this study is to assess the quality of the SNOMED CT CG inferred expressions of concepts candidates to constitute the common ontology between ICD-11 and SNOMED CT. We measure the rate of issues in...
modeling when the SNOMED CT concept model instance is not aligned with the Fully Specified Name of the SNOMED CT concept.

The study is limited to 428 classes from ICD-11-MMS, covering the circulatory system, and 522 ones covering the digestive system. We exclude ICD-11 codes ending with a "Y" or "Z" code (other specified or unspecified term at the subchapter level): 206 of such residual codes are in the circulatory chapter and 250 in the digestive chapter (see Table 3). Residuals assure that in ICD hierarchies all sub-hierarchies are fully exhaustive, although they have no clear meaning outside the context of the ICD hierarchy.

For this study, we focused on SNOMED CT concepts that can be lexically mapped to ICD-11 MMS classes either by one pre-coordinated SNOMED CT concept or by more than one post-coordinated SNOMED CT concept as stated in the introduction and explained in the following subsection. We analyze only the CG expressions of these candidate SNOMED CT concepts lexically mapped to ICD-11 MMS classes.

The work was done by two different knowledge engineering master students one in charge of the circulatory chapter and one in charge of the digestive chapter. Both chapters were supervised by the same senior ICD-11 and SNOMED CT expert.

Methods

Semantic interoperability is a crucial point for using SNOMED CT and ICD in parallel, e.g. to address clinical documentation, epidemiology, research and decision support use cases.

The first step is to perform a lexical map and in a second step a match of meaning between ICD-11-MMS classes, supported by their subclass hierarchies and text definitions and SNOMED CT concepts. Both terminologies are accessed by the web browsers made available by the respective maintenance organizations (cf. Figure 2 and Table 2).

We introduce the following symbols for the mapping types: M (refined by M1 and M2), A (refined by A1 and A2), P and Z. We consider the mapping of a SNOMED CT Concept $SC_i$, described by terms $ST_{i1\ldots n}$ to an ICD class $IC_i$, described by a name $IT_i$.

Lexical map

The following rules apply for lexical map

- If there is a full lexical map between the ICD-11 class name $IT_i$ and one SNOMED CT description $ST_{i1\ldots n}$, considered as pre-coordinated in SNOMED CT it is classified as M (for Lexical map) type.
- If there is no lexical map between any $IT_i$ and $ST_{ik}$, but if mapping can be achieved to the post-coordination of two or more descriptions $ST_{i1\ldots n}$ of $SC_k$, it is classified as A (for Addition map) type.
- If only a part of $IT_i$ of $IC_i$ can be lexically mapped to any $ST_{ik}$ it is classified as P (for Partial) type.
- Finally, if not even a partial lexical mapping between any $IT_{i\alpha}$ of $IC_i$ and $ST_{ik}$ is possible, it is classified as Z (for Zero) type.

Match of meaning

Subsequently, the CG inferred expressions $^2$ of $SC_i$ are analyzed to check whether they correspond to the totality of the textual definition and hierarchy inheritance of $IC_i$ and of defining and constraining axioms of one or more than one $SC_i$.

The following cases are distinguished:

- M (lexical match) type:
  1. This expression fully represents the meaning of $IC_i$, a complete match meaning is assumed: the classification is refined to M1.
  2. This expression does not fully represent the meaning of $IC_i$, a new expression is produced according to CG: the classification is refined to M2.

- A (addition map) type:
  1. These expressions fully represent the meaning of $IC_i$, a complete match meaning is assumed: the classification is refined to A1.
2. These expressions do not fully represent the meaning of $IC_i$, a new expression is produced according to CG: the classification is refined to $A_2$.

- **P type:**
  For $IC_i$ it is then necessary to create a logical representation based on one existing CG expression plus an extended de novo CG expression.

- **Z type:**
  For this $IC_i$ it is necessary to create a logical expression in accordance with SNOMED CT CG.

In the following, only M and A types will be analyzed.

---

**Figure 2.** ICD-11 SNOMED CT semantic alignment principles

---

**Table 2.** The lexical maps types and meaning matches between the ICD-11 MMS classes and SNOMED CT compositional grammar.

<table>
<thead>
<tr>
<th>Lexical map and meaning match</th>
<th>Action</th>
<th>Compositional grammar</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lexical map and full meaning match (M1).</td>
<td>Take the representation expression of the SNOMED CT concept</td>
<td>The existing pre-coordinated inferred expression of</td>
</tr>
</tbody>
</table>
Lexical map and no full meaning match (M 2)  Take the representation expression of the SNOMED CT concept  Modify the existing pre-coordinated inferred expression of SNOMED CT concept

Post-coordinated lexical map possible and full meaning match (A 1).  Take the representation of two or more pre-coordinated existing representations of SNOMED CT concepts  Post-coordination of two or more pre-coordinated existing inferred expression of SNOMED CT concepts

Post-coordinated lexical map possible but no full meaning match (A 2).  Take the representation of two or more pre-coordinated existing representations of SNOMED CT concepts  Post-coordination and modification of two or more pre-coordinated existing inferred expression of SNOMED CT concepts

Partial lexical map (P)  Take the representation of one pre-coordinated existing representation of SNOMED CT concept  One pre-coordinated existing inferred expression of a SNOMED CT concept plus an extended de novo CG expression

No lexical map (Z).  Create a logical CG expression  A new logical CG expression

Results
Table 3 provides an overview of the results.

Table 3. Numbers of codes in the circulatory chapter and Digestive chapter, from ICD 11 MMS 2017 to SNOMED CT 31 January 2017 release by map and meaning match types

<table>
<thead>
<tr>
<th>Map and meaning match types</th>
<th>Circulatory system count</th>
<th>Circulatory system rate (%)</th>
<th>Digestive system count</th>
<th>Digestive system rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1</td>
<td>243</td>
<td>56.77</td>
<td>275</td>
<td>52.68</td>
</tr>
<tr>
<td>M2</td>
<td>94</td>
<td>21.96</td>
<td>85</td>
<td>16.28</td>
</tr>
<tr>
<td>A1</td>
<td>51</td>
<td>11.92</td>
<td>29</td>
<td>5.56</td>
</tr>
<tr>
<td>A2</td>
<td>13</td>
<td>3.04</td>
<td>104</td>
<td>19.92</td>
</tr>
<tr>
<td>P</td>
<td>27</td>
<td>6.31</td>
<td>10</td>
<td>1.92</td>
</tr>
<tr>
<td>Z</td>
<td>0</td>
<td>0.00</td>
<td>19</td>
<td>3.64</td>
</tr>
<tr>
<td>Total (M+A+P+Z)</td>
<td>428</td>
<td>100</td>
<td>522</td>
<td>100</td>
</tr>
<tr>
<td>Total (M+A+P+Z)</td>
<td>428</td>
<td>67.51</td>
<td>522</td>
<td>67.62</td>
</tr>
<tr>
<td>&quot;Other&quot; and &quot;Unspecified&quot;</td>
<td>206</td>
<td>32.49</td>
<td>250</td>
<td>32.38</td>
</tr>
<tr>
<td>Total</td>
<td>634</td>
<td>100</td>
<td>772</td>
<td>100</td>
</tr>
</tbody>
</table>

The two most frequent lexical map types are M (M1 plus M2) for full lexical map with a pre-coordinated SNOMED CT concept and A (A1 plus A2) full lexical map with a post-coordinated SNOMED CT concept: 93.69 % for the circulatory chapter and 94.44% for the digestive chapter.

The most frequent type is M1 for for both
The less frequent types are Z for no possible lexical map for the circulatory chapter (0%) and P for partial lexical map for the digestive chapter (1.92%).

The other main differences between the two chapters are the higher rate for the match types A2 for the digestive chapter (19.92% against 3.04%): These differences can be explained by interrater differences (the work was done by two different knowledge engineering master students supervised by the same senior terminology expert) or quality differences between these two chapters either in WHO ICD 11 or in SNOMED CT or in both.

We found for these two chapters nearly the same percentage found on a previous work on the circulatory chapter with browsers based on WHO ICD-11 and SNOMED CT one year before the current study (31 January 31 2016).

The goal of our study is to measure the quality of SNOMED CT concept model for the SNOMED CT concepts candidates for a lexical map with ICD 11 class names and definitions i.e lexical map types M and A which correspond to more than 90% of ICD-11 classes for the circulatory and the digestive chapters.

As an example for M1 type ICD-11 DA 40.4 Perforation of esophagus is defined by: “perforation of esophagus is a penetration or hole of the wall of the esophagus, resulting in luminal contents in esophagus flowing into the mediastinum and/or thoracic cavity”.

The full lexical map is with SNOMED CT 23387001 | Perforation of esophagus (disorder) | which is a fully defined concept with the following pre-coordinated SNOMED CT inferred expression in Figure 4 and the diagram in Fig. 5.

Figure 4. Pre-coordinated CG expression for Perforation of Esophagus

```plaintext
37657006 | Disorder of esophagus (disorder) | + 51875005 | Gastrointestinal perforation (disorder) | + 300286002 | Lesion of esophagus (finding) |
| { 363698007 | Finding site (attribute) | = 32849002 | Esophageal structure (body structure),
  116676008 | Associated morphology (attribute) | = 36191001 | Perforation (morphologic abnormality) } |
```

Figure 5. Pre-coordinated CG diagram for Perforation of Esophagus

As an example for M2 type ICD-11 BA00 Essential hypertension is defined as “Essential (primary) hypertension, accounting for 95% of all cases of hypertension, is defined by ICD as high blood pressure for which a secondary cause cannot be found”.

The full lexical map is with SNOMED CT 59621000 | Essential hypertension (disorder) | which is not a fully defined concept but a primitive meaning that the SNOMED CT concept FSN has no full representation with the following pre-coordinated SNOMED CT inferred expression in Figure 6 and the diagram in Figure 7.

Figure 6. Pre-coordinated CG expression for Essential hypertension

As an example for M2 type ICD-11 BA00 Essential hypertension is defined as “Essential (primary) hypertension, accounting for 95% of all cases of hypertension, is defined by ICD as high blood pressure for which a secondary cause cannot be found”.

The full lexical map is with SNOMED CT 59621000 | Essential hypertension (disorder) | which is not a fully defined concept but a primitive meaning that the SNOMED CT concept FSN has no full representation with the following pre-coordinated SNOMED CT inferred expression in Figure 6 and the diagram in Figure 7.

Figure 6. Pre-coordinated CG expression for Essential hypertension
This is only the representation of hypertensive disorder, systemic arterial (disorder) which has the same representation and there is no representation of the ICD-11 class definition statement: a secondary cause cannot be found. We will present later how CG can represent this statement.

Table 3 shows how SNOMED CT concept model represents logically ICD-11 classes

### Table 3. Quality of SNOMED CT concept model to represent logically ICD-11 classes

<table>
<thead>
<tr>
<th>Map and meaning match types</th>
<th>Circulatory system count</th>
<th>Circulatory system rate (%)</th>
<th>Digestive system count</th>
<th>Digestive system rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>M 1</td>
<td>243</td>
<td>56.77</td>
<td>275</td>
<td>52.68</td>
</tr>
<tr>
<td>M 2</td>
<td>94</td>
<td>21.96</td>
<td>85</td>
<td>16.28</td>
</tr>
<tr>
<td>Rate M2/(M1+M2)</td>
<td>94/337</td>
<td>27.89</td>
<td>85/360</td>
<td>23.61</td>
</tr>
<tr>
<td>A 1</td>
<td>51</td>
<td>11.92</td>
<td>29</td>
<td>5.56</td>
</tr>
<tr>
<td>A 2</td>
<td>13</td>
<td>3.04</td>
<td>104</td>
<td>19.92</td>
</tr>
<tr>
<td>Rate A2/(A1+A2)</td>
<td>13/64</td>
<td>18.75</td>
<td>104/133</td>
<td>78.19</td>
</tr>
<tr>
<td>Poor quality Rate= (M2+A2)/(M1+M2+A1+A2)</td>
<td>107/401</td>
<td>26.68</td>
<td>199/493</td>
<td>40.36</td>
</tr>
</tbody>
</table>

To summarize the quality issue there is a rate of poor quality of 306/894 = 34.22 %

It is interesting to compare the poor quality rate to represent logically ICD-11 classes with the poor quality rate to represent SNOMED concepts which can be the rate of primitive SNOMED CT concept out of the candidates SNOMED CT concept to represent ICD 11 classes as shown in Table 4.

### Table 4. Primitive SNOMED CT concepts by map and meaning match types

```sql
<< 38341003 | Hypertensive disorder, systemic arterial (disorder);
  36370008 | Has definitional manifestation (attribute) =
  24184005 | Finding of increased blood pressure (finding);
  363698007 | Finding site (attribute) = 51840005 | Systemic circulatory system structure (body structure) |
```
The types with full map and meaning match (M1 and A1) have a low rate of SNOMED CT primitive concepts (from 0% to 33%) and the types with no full match (M2 and A2) have a high rate of SNOMED CT primitive concepts (from 81% to 100%). It can be considered that the quality issue with SNOMED CT concept model concerns mainly the rate of primitives with few exceptions.

Discussion
The semantic alignment between ICD-11-MMS classes and SNOMED CT concepts provides a good opportunity to trace the quality issues with the SNOMED CT concept model.

According to our study, 107 out of 401 SNOMED CT concepts (26.68%) in the circulatory chapter and 199 out of 493 SNOMED CT concepts (40.36%) in the digestive chapter from the Clinical finding hierarchy that lexically mapped to ICD-11 classes show modeling issues resulting in misalignments not only between ICD-11 MMS class meaning and SNOMED CT concept model instances, but as well between SNOMED CT Fully Specified Names and concept model instances. We here list some typical examples: As shown in Table 4 in most of the cases this is related to the high number of primitives, i.e. not fully defined SNOMED CT concepts for more than 80% of SNOMED CT concept candidates to lexically map with ICD-11 class names.

Misalignment between SNOMED CT concept FSN and primitive representation
In the example of Essential hypertension (ICD-11 class BA 00), which is the main category of Arterial hypertension which is one of the most frequent diseases in the world, Figure 7 shows that the existing SNOMED CT expression does not represent the lack of secondary cause, which is the meaning of “essential” or “idiopathic”. The CG provides the possibility to represent the lack of secondary cause by adding the following attribute and domain value: 370135005 | Pathological process (attribute) | 54690008 | Unknown (origin) (qualifier value) | and the diagram of Figure 8.

Figure 8. Actual SNOMED CT CG candidate representation of ICD 11 class BA 00 and proposed fully defined representation.
There are several other cases with the wording “of unknown etiology” or, e.g. as 85598007 | Constrictive pericarditis (disorder) | with no representation of “constrictive”, 373945007 | Pericardial effusion (disorder) | with no representation of “effusion”, 706882009 | Hypertensive crisis (disorder) | with no representation of “crisis”.

**Misalignment between SNOMED CT concept FSN and fully defined representation**

ICD-11 DA52.51 Allergic gastritis due to IgE-mediated hypersensitivity can be fully represented by the SNOMED CT concepts 1824008 | Allergic gastritis (disorder) | and 422076005 | Immunoglobulin E-mediated allergic disorder (disorder) |, both of which are fully defined. The role of Immunoglobuline E is not represented in the present version of the SNOMED CT concept model.

**Conclusion**

To answer the question of the title, *viz.* whether the application of the SNOMED CT concept model is sufficiently quality assured, we can state the following points as a route to an answer.

Semantic interoperability between a clinical terminology as SNOMED CT and a statistically aggregated classification as ICD-11 will be more and more necessary in the future.

The SNOMED CT concept model is a good opportunity to provide a logical basis to automate the coordination.

Problems in matching SNOMED and ICD are a way of highlighting quality issues when applying the SNOMED CT concept model.

We recommend prioritizing SNOMED CT TQA on the subset of the core SNOMED CT content by using ICD11 definition coming from WHO closed world.
This can contribute to develop a common SNOMED CT and ICD-11 ontology always true in an open world assumption6.

References

Abstract

Automated literature analysis could significantly speed up understanding of the role of the placenta and the impact of its development and functions on the health of the mother and the child. To facilitate automatic extraction of information about placenta-mediated disorders from the literature, we manually annotated genes and proteins, the associated diseases, and the functions and processes involved in the development and function of placenta in a collection of PubMed/MEDLINE abstracts. We developed three baseline approaches to finding sentences containing this information: one based on supervised machine learning (ML) and two based on distant supervision: 1) using automated detection of named entities and 2) using MeSH. We compare the performance of several well-known supervised ML algorithms and identify two approaches, Support Vector Machines (SVM) and Generalized Linear Models (GLM), which yield up to 98% recall precision and F1 score. We demonstrate that distant supervision approaches could be used at the expense of missing up to 15% of relevant documents.

Introduction

The placenta is the most important organ in human pregnancy. It plays the role of lungs and kidneys for the developing fetus, supplies substrates for its development, and regulates complex immune functions to allow the cohabitation of two different organisms - the mother and the fetus - during the pregnancy. Defects in the placentation process are known to be associated with a wide range of pregnancy related complications such as preeclampsia, uterine growth restriction, and premature rupture of membranes, fetal growth retardation, placenta abruption, spontaneous abortion, and fetal death. The association between defects in the placenta, the placentation process and most of the above-mentioned diseases was established in the 1950’s. More recently additional associations between the functionality of the placenta and other maternal/fetal diseases have become evident. For example, in gestational diabetes, adipokins, leptin and TNFα produced in placenta are implicated in gestational insulin resistance and possibly in insulin resistance in the adult life of the fetus through a process known as “fetal programming”. The functionality of the placenta influences maternal and fetal health and development and impacts future health for both mother and baby, not only during pregnancy but for the lifetime of both (1) (2) (3) (4). The study of the organ in real time using invasive procedures to sample the placenta and the placenta insertion site can be dangerous for the mother and the fetus, even with the use of guided imaging technology. Safer study approaches during pregnancy have only become available recently with the advancement of genetic and molecular sciences. The full potential of now-available biomedical research tools has not yet been applied to the study of the placenta. Recognizing the importance of multidisciplinary collaboration for the study of the placenta, the National Institutes of Health (NIH) through the National Institute of Child Health and Human Development (NICHD) established the Human Placenta Project (HPP) “to understand the role of the placenta in health and disease and to develop new tools to learn how it develops and functions throughout the pregnancy” (2).

Discovery of gene pathways and biochemical mechanisms that help explain disease causality can be facilitated by automated extraction of relevant information from the biomedical literature (5). One essential step in understanding complex pathways has been the compilation of disease-specific gene candidates extracted from the literature. Data analysis then makes it possible to create candidate gene assays and translate the results to the wet bench for biochemical research (6).

Our long-term goal is to extract genes, gene pathways, biomarkers, and related events from the human placenta literature to create a specialized human placenta gene repository and to identify pathways that can uncover target genes and gene therapies for pregnancy-related diseases. In this study, we present a machine learning approach to extract and identify sentences describing gene-disease relationships and gene and protein activity from a collection of studies in the placenta literature. We also make the collection of PubMed® (7) abstracts annotated for this study publicly available.
available. In addition, we evaluate if the manually annotated relations among genes, proteins and diseases could have been captured by the rules based on MeSH® indexing – human annotation of MEDLINE® abstracts by NLM indexers.

**Background**

Perch and Altman point out that all important biomedical knowledge is described in the published research literature, but to operationalize this knowledge we need computational algorithms that can efficiently extract, aggregate, annotate, and store information from the raw text (8).

To develop approaches for the automated extraction of genes and gene pathways studied in the human placenta, it is necessary to identify specific genes, their mechanisms of action, and related diseases in the human placenta literature. Having information on already-identified cause/effect relationships between genes and diseases should expedite the extraction process. Extraction of this information from the literature will allow the discovery of novel relationships as new genes are identified in the human placenta. Some hereditary aspects of such associations have already been explored in the clinical setting, where women who have developed pregnancy-induced hypertension or gestational diabetes appear to have a higher incidence and family history of chronic hypertension and Type II diabetes, respectively (1). Similar approaches to automated extraction of information from the literature have been useful in Alzheimer disease research: literature-based gene enrichment and prioritization tools were used for the discovery of novel genes related to Alzheimer’s disease (AD) (9) and protein-protein interaction networks constructed using data from literature sources, among others, were advantageous in providing objective prioritization of disease-gene candidate criteria (10).

The understanding of biological systems requires not only the ability to extract entities such as cells, proteins, genes, diseases, etc., but also to establish associations between different identified entities and cause/effect relationships (11). It is necessary to understand the role of each entity in a given relationship to establish a cause/effect association. Reports in the literature and shared natural language processing tasks have focused on DNA methylation (12), protein modifications (13), and gene expression mentions in anatomical locations (14), among others. Automatic sentence classification has been used to support evidence based medicine (15), for automatic extraction of clinically useful sentences in clinical evidence resources (16), and for automatic retrieval of abstracts on randomized controlled trials (17).

The goal of the current study is to develop approaches to facilitate automatic text identification of gene associations, their mechanisms of action, and their effects on diseases by employing automated extraction of entities, their roles, and the relationships reported in the human placenta literature. In this work, we present the annotated collection of documents needed for supervised machine learning approaches for extraction of the above information and a comparison of supervised and distantly supervised approaches for extraction of contexts that contain this information.

**Methods**

In our previous short communication (18), we described the dataset obtained for the study. In this section, we summarize information about the data and focus on the information extraction methods.

**Data**

The corpus was downloaded from PubMed on December 8, 2015, searching for: placenta AND (human OR woman OR women) AND (gene OR genes OR biomarker* OR polymorphism* OR enzyme*) AND (gestational diabetes OR hypertension OR preeclampsia OR pre-eclampsia OR eclampsia OR SGA OR growth restriction OR preterm OR HELLP OR acute fatty liver OR DVT OR anemia OR placenta abruption OR placenta previa OR stillbirth OR miscarriage) in the articles that have abstracts. The terms preeclampsia, hypertension, and other pregnancy and fetus-related diseases were included in the search because they are the most common causes of maternal-fetal morbidity in pregnancy. The search retrieved 428 MEDLINE citations. The search was specifically aimed at retrieving human placenta-related studies; however, the manual review revealed that only 300 documents referred to human placenta, while the remaining 128 were human placenta gene studies in animal models. The 128 documents with mentions of gene studies in animal placentas were excluded from the study.

**Manual Annotation**

We set out to annotate gene and protein activity events and all biomarkers, genes, and disease entities involved in the events, along with the roles the genes play in the events. We based our annotation schema on those developed for event annotation in the GENIA corpus (19) and for gene-drug relationships extraction (20). Table 1 presents the
entities and events annotated in our corpus, along with their counts. Events were marked as negated or speculation when clear negation or hedging was stated in the text. To differentiate between genes and proteins with the same name, we annotated entity mentions involved in increased or decreased levels or amount of events as protein. We annotated mentions involved in increased or decreased expression or activity of events as gene. We used BRAT online annotation tool (21) installed in a secured server and configured with the annotation schema designed for this study.

The four authors all with biomedical informatics experience and training--two physicians (DF and LR), a biologist (SM), and a cellular and molecular scientist (KG)--annotated the first 20 documents individually, reconciled the differences together, and finalized the guidelines for annotation. The remaining documents were annotated individually by two annotators each, and the differences were reconciled in pairs. We computed inter-annotator agreement using F1-score, as proposed by Hripcsak and Rothschild (22). The inter-annotator agreement between the pairs of annotators was fair. The agreement between the clinician and the geneticists was higher (58.5% on average), than between the geneticists whose F1-score was 40.3%. The better agreement between the clinician and the geneticists could be explained by learning, as for annotating the second half of the documents we paired the geneticists with the clinician, so the agreement between the geneticists was measured earlier in the process. The final modest scores reflect the difficulty of the task. The gold standard should not be affected by the modest initial agreement, as all differences were discussed and carefully reconciled in group meetings.

The resulting collection of annotated documents was then used for machine learning experiments to extract and identify sentences containing information about biomarkers and the activity of genes and proteins in the human placenta that are associated with diseases in different stages of pregnancy. Figure 1 shows an example of a fully annotated sentence.

Figure 1: Annotation of protein activity mention associated with disease in human placenta study. In this example, correlates is an event of type is-associated, and angionic regulators and oxidative stress biomarkers are entities of the type protein_enzyme and gene_family respectively. These entities have the roles of objects of the correlates event.

We explored two distant supervision approaches on the same collection as an alternative to manual annotation. We explored using the NLM PubTator web-based tool (23) to automatically annotate entity types: Disease, Species, Mutation, and Gene. We then compared the use of these automatically derived annotations as features in machine learning to manual annotation-based features.

We also explored exploiting MeSH indexing. Jimeno Yepes et al. demonstrated the use of MeSH headings and subheadings in extraction of gene-disease relationships (24). In the MeSH indexing process, NLM indexers assign headings (terms that describe the topic of the document) and subheadings (qualifiers) from the MeSH controlled vocabulary to MEDLINE citations. The subheading creates coordination among the headings, e.g., in the document PMID: 25305692 the indexing terms Pre-Eclampsia/genetics* and 14-3-3 Proteins/genetics* potentially indicate involvement of 14-3-3 Proteins in Pre-Eclampsia (disease); the asterisk indicates that both headings are the major topics in the article. For each one of the 300 documents we: 1) extracted indexing terms with the same subheadings; 2) derived indexing-based rules that establish associations among entities involved in the events presented in Table 1 and roughly correspond to these events; and 3) evaluated the relations against our gold standard as described below. Note that this approach is complementary to the initial PubMed retrieval strategy, which is recall-oriented and designed to find as many potentially relevant articles as possible. The MeSH indexing rules are geared towards finding specific relations in the set of the retrieved articles.
The following rules presented in the form subheading (Heading type, Heading type) capture associations corresponding to the events. For example, the rule genetics (disease, chemical) corresponds to association and finer grained events involving genes and proteins. In document PMID 11766889 this rule captures genetics (Pre-Eclampsia, HLA Antigens), which approximately captures the decreased_level_of event manually annotated in “A deficit in levels of the HLA-G3 transcript was observed in mild pre-eclampsia compared to normal placentas.”

- enzymology(disease, anatomy)
- enzymology(anatomy, cellular structure)
- biosynthesis(chemical, chemical)
- genetics(disease, chemical)
- genetics(disease, cellular structure)
- genetics(Physiological Phenomena, disease)
- immunology(anatomy, disease)
- immunology (Physiological Phenomena, disease)
- metabolism(disease, chemical)
- metabolism(anatomy, disease)
- metabolism(Physiological Phenomena, disease, chemical)
- pathology | physiopathology (disease, anatomy)
- pathology | physiopathology (anatomy, chemical)
- physiology (anatomy, chemical)
- physiology (Physiological Phenomena, chemical)
- physiology (Physiological Phenomena, anatomy)

We evaluated the correspondence between the associations captured by the above rules and the manually annotated gold standard (GS) as follows: 1) the MeSH-based association corresponds to at least one event in the GS (True Positives), 2) the events in the GS are not captured in MeSH indexing (False Negatives), 3) No events related to a disease are annotated in GS and the rules produced no associations (True Negatives), 4) A MeSH-based association has no corresponding events in GS (False Positive).

**Machine Learning**

We used R statistical software core package for data preprocessing and machine learning experiments, and the packages RTextTools, E1071, and MxNet for the machine learning experiments (25) (26) (27).

In the machine learning (ML) experiments, we used sentences with entities and events (963), sentences with only entity mentions (246), and a random sample of 200 sentences with no entity mentions or events. We used WordToVec filter to process the sentences eliminating English stop words, numbers, punctuation, and extra white spaces.

For classification purposes we used supervised machine learning algorithms known to perform well on text with settings briefly described below. We used a stepwise gradual approach adding features to measure performance with each of the ML algorithms. The features used for the stepwise approach were: bag of words, manually annotated events, entities, roles and attributes (negation and speculation). To test the distant supervision approach based on automatic annotation of entities, we used entities identified using PubTator, PubTator types and PubTator codes.

We used the R implementation of Support Vector Machine (SVM) (28) based on LibSVM with default settings for the radial kernel.

*Glmnet:* an R package that fits a generalized linear model via penalized maximum likelihood (29). We used the default settings with alpha set to 0.2, which makes it closer to ridge penalty.

*LogitBoost:* is an algorithm in R adapted to add a convex optimization to Boosting (30) that uses decision stumps, one node decision trees as weak learners (31).

*Max Entropy:* R implementation of Low-memory Multinomial Logistic Regression with Support for Text Classification. The regularizers were turned off in our experiments.

*Neural Networks:* (NNet) we used an R implementation of the single layer neural network with the default settings (32).
Evaluation

We evaluated Recall, Precision and F1 scores for extraction of sentences containing entities and events using 10-fold cross validation on the set of 963+246 + 200 sentences labeled respectively as containing events, entities only, and not relevant. We evaluated Sensitivity and Specificity for gene-protein/disease relations in MeSH indexing.

Results

A summary of the manual annotation of the entities and events in 300 documents is shown in Table 1.

The most commonly mentioned genes were LEP, PHLDA2, FLT1, PGF, STOX1, and VEGF. The LEP gene, which provides instructions for making a protein called leptin, is involved in the regulation of body weight, and in our collection it was associated with pre-eclampsia (33) (34). The PHLDA2 gene, generally associated with tumor growth suppression and placenta growth, was associated with placenta insufficiency and fetal growth retardation in our collection (35) (36). The FLT1, VEGF, and PGF (placenta growth factor) genes, known to be associated with development of embryonic vasculature, the regulation of angiogenesis, cell survival, and cell migration, were associated with pre-eclampsia and intrauterine growth restriction in our collection (37) (38) (39). The most frequent events were associations between gene mentions and the diseases preeclampsia and intrauterine growth retardation followed by HELLP syndrome and preterm delivery.

Extraction of sentences containing entities and events:

The results of classifying sentences as containing gene-placental mediated events, entities only, and not relevant are shown in Table 2. Adding manually annotated entities improved the results for all classifiers by at least 60%. For the automatically extracted entities, the boost was much smaller, and the difference in the improvements provided by the automated and manual entity annotation was preserved even when other manual features were added to PubTator annotations. The second visible improvement in the results is associated with adding the event annotations as features, whereas the roles and the modality features either do not contribute anything or slightly worsen the results.

Table 1. Event Types and Entities in manual annotation

<table>
<thead>
<tr>
<th>Event Types</th>
<th>#</th>
<th>Entities</th>
<th>#</th>
</tr>
</thead>
<tbody>
<tr>
<td>is_associated</td>
<td>260</td>
<td>Gene</td>
<td>1324</td>
</tr>
<tr>
<td>increases_activity</td>
<td>214</td>
<td>Disease</td>
<td>1189</td>
</tr>
<tr>
<td>decrease_activity</td>
<td>151</td>
<td>Protein_Enzyme</td>
<td>633</td>
</tr>
<tr>
<td>increases_levels_of</td>
<td>142</td>
<td>Organ</td>
<td>374</td>
</tr>
<tr>
<td>is_different</td>
<td>106</td>
<td>Tissue</td>
<td>184</td>
</tr>
<tr>
<td>decrease_levels_of</td>
<td>93</td>
<td>PhysiologicProcess</td>
<td>139</td>
</tr>
<tr>
<td>is_expressed</td>
<td>72</td>
<td>Cell</td>
<td>110</td>
</tr>
<tr>
<td>affects_expression</td>
<td>49</td>
<td>GeneFamily</td>
<td>104</td>
</tr>
<tr>
<td>affects_modifies</td>
<td>33</td>
<td>Process</td>
<td>74</td>
</tr>
<tr>
<td>regulates</td>
<td>22</td>
<td>PhysiologicState</td>
<td>29</td>
</tr>
<tr>
<td>cause</td>
<td>19</td>
<td>Cell_Component</td>
<td>26</td>
</tr>
<tr>
<td>inhibits</td>
<td>16</td>
<td>Substance</td>
<td>10</td>
</tr>
<tr>
<td>is_active</td>
<td>10</td>
<td>Chromosome</td>
<td>4</td>
</tr>
<tr>
<td>activates</td>
<td>7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>stimulates</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>synthesizes</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>suppresses</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>induces_expression</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>interacts</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total instances</td>
<td>1206</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 2. Contribution to sentence-level classification results of progressively adding entities, events, roles and attributes features to the bag of words. $R$ – Recall, $P$ – Precision. PubTator entities+ includes manually annotated event, roles and attributes.

<table>
<thead>
<tr>
<th>Bag of words</th>
<th>+ entities</th>
<th>+events</th>
<th>+roles</th>
<th>+attributes</th>
<th>PubTator entities</th>
<th>PubTator entities+</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$P$</td>
<td>$R$</td>
<td>$F_1$</td>
<td>$P$</td>
<td>$R$</td>
<td>$F_1$</td>
</tr>
<tr>
<td>SVM</td>
<td>0.70</td>
<td>0.55</td>
<td>0.67</td>
<td>0.83</td>
<td>0.85</td>
<td>0.98</td>
</tr>
<tr>
<td>LogitBoost</td>
<td>0.53</td>
<td>0.53</td>
<td>0.87</td>
<td>0.82</td>
<td>0.84</td>
<td>0.94</td>
</tr>
<tr>
<td>GLMNet</td>
<td>0.60</td>
<td>0.51</td>
<td>0.83</td>
<td>0.75</td>
<td>0.78</td>
<td>0.97</td>
</tr>
<tr>
<td>NNetWork</td>
<td>0.47</td>
<td>0.47</td>
<td>0.69</td>
<td>0.70</td>
<td>0.70</td>
<td>0.93</td>
</tr>
<tr>
<td>MaxEnt</td>
<td>0.59</td>
<td>0.56</td>
<td>0.67</td>
<td>0.73</td>
<td>0.76</td>
<td>0.92</td>
</tr>
</tbody>
</table>

Evaluation of MeSH Indexing-based event extraction:

Not surprisingly, MeSH headings contained the following diseases included in our search: *preeclampsia*, *HELLP syndrome*, *intrauterine growth retardation*, *gestational diabetes* and *premature rupture of membranes*. The chemicals included: *Nitric Oxide Synthase*, *Plasminogen Activators*, *Nucleic Acids*, *Nucleotides*, and *Nucleosides* among others. The most common subheadings were: *genetics*, *metabolism*, *physiology*, *pathology*, and *enzymology*. Three documents did not have MeSH indexing. One was published prior to the date indexing started for the publishing journal; the two others are in journals not indexed for MEDLINE. The non-indexed documents were discarded from the calculations. The sensitivity for the MeSH based association rules was 0.87 and specificity was 0.97. Table 3 presents the evaluation results.

Table 3. Comparison of manual annotations (GS) and associations based on MeSH indexing rules for document-level classification

<table>
<thead>
<tr>
<th>MeSH-GS Category</th>
<th># documents</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rule-based associations capture at least one event annotated in GS (True Positive)</td>
<td>256</td>
</tr>
<tr>
<td>GS event is not captured by any rule (False Negative)</td>
<td>36</td>
</tr>
<tr>
<td>No events related to a disease are annotated in GS and the rules produced no associations (True Negative)</td>
<td>4</td>
</tr>
<tr>
<td>A rule-based association has no corresponding events in GS (False Positive)</td>
<td>1</td>
</tr>
<tr>
<td>Documents with no MeSH indexing</td>
<td>3</td>
</tr>
</tbody>
</table>

Discussion

Our results demonstrate that an SVM or a GLM model can be used to reliably identify sentences containing information about genes associated with placenta-mediated disorders. Our collection provides a means to further extract several types of events, such as *increased/decreased activity* and *increased/decreased levels of*. Additionally, our exploration of MeSH indexing shows that MeSH headings used in conjunction with subheadings can be used to identify documents that establish gene or protein activity in certain diseases. MeSH controlled vocabulary does not include specific gene or protein terms, but it can be used in conjunction with entity annotation tools such as PubTator.

The field of genetic studies in the human placenta in real time has emerged in recent years due to availability of tissue testing on smaller sample sizes, which decreases the need for invasive intrauterine procedures that can be deleterious to the fetus and the mother. The Human Genome Project has made it possible to make important advancements in the study of genetic related diseases; knowledge grows every day on gene activity and physiology (40). With an estimated 19,000-20,000 human protein-coding genes, the task of finding genes associated with disease in the human placenta would be daunting without automated support. We have demonstrated that it is possible to use machine learning algorithms to extract knowledge from the literature in this field. As the literature grows and reports increase in size and volume, it is necessary to make use of automatic tools to guide the wet-bench research. Likewise, automatic extraction of genes, gene disease relationships, and activity allows mapping to annotated genetic databases for knowledge extraction and further guidance.
We show that even sophisticated literature searches alone are not sufficient to extract only human placenta studies: at least 30% of the studies retrieved by our search were researching human tissues in animal models. Our results also show that although distant supervision using the state-of-the-art tools such as PubTator for named entity recognition is possible, about 15% of the documents with gene mentions will be missed using this approach. Similarly, relying on MeSH indexing alone will also result in missing 13% of the relevant documents. However, MeSH indexing is an alternative source of high quality information that we plan to combine with the machine learning approaches in the future.

Limitations and Future Work

The largest limitation of the study is the size of the corpus and the paucity of some events, such as activates, stimulates, and suppresses. The collection, however, reflects the state of this area of research: a search on PubMed in January, 2017 retrieved 488 documents; only 60 more than the study corpus. The genetic human placenta literature is growing at a relatively slow pace, but we anticipate that as the knowledge increases, the rate will also increase. The specific events extracted from the identified sentences can be used to inform future research projects investigating gene pathways related to the identified gene-disease associations and potential precision medicine approaches for the mitigation of placenta-mediated disorders.

Conclusion

Our study demonstrates the possibility of automatic extraction of sentences containing information about genes associated with placenta-mediated disorders in the human placenta literature. We compare the performance of several well-known supervised ML algorithms and identify two approaches, SVM and GLM, which yield up to 98% recall precision and F1 score. We demonstrate that distant supervision approaches based on MeSH indexing and automatic extraction of entities could be used at the expense of missing up to 15% of relevant documents. Our annotation schema, guidelines, and the annotated documents will be available at https://bionlp.nlm.nih.gov/PlacentaCollection.

References


Electronic Surveillance For Catheter-Associated Urinary Tract Infection Using Natural Language Processing

Patrick C Sanger, MD PhD1,2, Marion Granich, RN2, Robin Olsen-Scribner, RN MPH CIC2, Rupali Jain, PharmD2, William B Lober, MD2, Ann Stapleton, MD2, Paul S. Pottinger, MD2
1UCSF, San Francisco, CA; 2University of Washington, Seattle, WA

Abstract
Catheter-associated urinary tract infection (CAUTI) is a common and costly healthcare-associated infection, yet measuring it accurately is challenging and resource-intensive. Electronic surveillance promises to make this task more objective and efficient in an era of new financial and regulatory imperatives, but previous surveillance approaches have used a simplified version of the definition. We applied a complete definition, including subjective elements identified through natural language processing of clinical notes. Through examination of documentation practices, we defined a set of rules that identified positively and negatively asserted symptoms of CAUTI. Our algorithm was developed on a training set of 1421 catheterized patients and prospectively validated on 1567 catheterized patients. Compared to gold standard chart review, our tool had a sensitivity of 97.1%, specificity of 94.5%, PPV of 66.7% and NPV of 99.6% for identifying CAUTI. We discuss sources of error and suggestions for more computable future definitions.

Introduction
Catheter-associated urinary tract infection (CAUTI) is a common healthcare-associated infection (HAI) in the US, contributing to the deaths of up to 13,000 patients and costing at least $400 million annually.1,2 Especially in light of new reporting requirements and payment reforms penalizing hospitals for development of certain HAIs, CAUTI identification and prevention have become top priorities for hospitals.3-5

Despite the economic and medical importance of CAUTI, standardized measurement of CAUTI incidence is challenging for two reasons. First, the standard Centers for Disease Control and Prevention (CDC) National Healthcare Safety Network (NHSN) CAUTI definition is complex and subjective.6-8 Second, CAUTI surveillance has historically been done through manual chart review, which is subject to significant inter-observer variability and is resource-intensive at a time when infection control departments have fewer resources for surveillance.9-11 As of 2008, fewer than half of hospitals had an established, house-wide surveillance system for monitoring CAUTI.12

Given the need for cheaper, more reliable surveillance in the electronic health record (EHR) age, there has been a trend toward development of automated HAI surveillance tools.5,13,14 However, previous attempts to automate CAUTI surveillance have modified the NHSN definition to exclude subjective symptoms because they are not typically stored as coded EHR data.15-20 Some recent approaches have used natural language processing (NLP) to identify the presence of a urinary catheter and/or fever.21,22 However, to date no published approach has implemented the NHSN definition fully, including all subjective symptoms (e.g., urgency, dysuria).

We have developed an automated surveillance tool capable of applying the NHSN definition as written, using NLP to capture subjective symptoms documented in free-text clinical notes. Herein we describe our approach, characterize its performance compared to manual chart review, and present opportunities for improvement, especially in terms of a more computable and reliable definition of CAUTI.

Methods
We developed a CAUTI surveillance tool through an iterative process that involved a multidisciplinary team of infectious disease specialists, an infection preventionist (IP), quality improvement (QI) staff, a statistician, and a programmer. The study was conducted at the University of Washington Medical Center, a regional tertiary academic...
medical center serving five states. This effort was reviewed by the University of Washington IRB, who exempted this work from further IRB involvement due to its objective of improving healthcare quality.

Data sources and study population

Patient data from our EHR (a Cerner implementation known as “ORCA”) were deposited daily into a Microsoft Amalga UIS database, which our CAUTI surveillance tool queried as needed (Figure 1). Billing and University HealthSystem Consortium (UHC) data were obtained from a hospital administrative database. The study population consisted of all inpatients ≥18 years of age with length of stay ≥ 24 hours with a urinary catheter in place within 48 hours prior to a urine culture, from February 2010 through July 2011 (training set) and August 2011 through August 2012 (validation set). Within this population, no patients were excluded.

Figure 1. Data sources and study population. All data is discretely encoded except for items with a single asterisk (free text) or double asterisk (mix of discrete and free text). UCx: urine culture. ASB: asymptomatic bacteriuria

CDC CAUTI definition implementation

We sought to implement the NHSN definition⁶ as faithfully as possible using an iterative development process. Figure 2 graphically represents our resulting algorithm. We began by identifying all urine cultures during the training timeframe. We then examined documentation of urine output on indwelling urinary catheters (charted discretely by RNs in our EMR) surrounding each urine culture. Patients were included if a catheter was present within 48 hours prior to the urine culture (UCx). Other kinds of catheters (e.g., suprapubic, in and out, condom) and nephrostomy tubes were not included.

We then used urine culture and urinalysis (UA) data per Figure 2. We defined a positive UA as meeting at least one of the following: positive or trace leukocyte esterase, positive nitrite, pyuria, or bacteria visualized on unspun urine. Pyuria was defined as >5 white blood cells per high power field (based on our lab’s cutoff point). If the UCx identified 3 or more distinct organisms, or described “mixed flora”, the specimen was excluded for contamination. Fever was defined as temperature >38°C; if a fever was present, we looked at blood culture (BCx) +/- 72 hours from UCx.

If the BCx was positive with a non-matching (by genus) organism, the fever was considered to be caused by the organism growing in blood and did not count towards CAUTI. If the blood culture was negative or caused by a
matching (by genus) organism, it was considered to represent urosepsis and the fever counted towards CAUTI. If the blood culture was negative or caused by a matching (by genus) organism, it was considered to represent urosepsis and the fever counted towards CAUTI.

Urinary symptoms other than fever were extracted from clinical notes using natural language processing (NLP; described further below) or discovered from a discreetly encoded “genitourinary symptoms” (GU) field in our EMR. We looked for symptoms within +/- 48 hours of the UCx based on the logic of the NHSN Transfer Rule, assuming that a CAUTI takes at least 44 hours to develop from inoculation. The presence of fever or any allowable symptom (i.e. some symptoms like dysuria do not count while the urinary catheter is present) in either a clinical note or the discrete GU field defined a candidate CAUTI.

A candidate CAUTI was then excluded if it met any of the following criteria: if a matching genus was identified from any previous UCx within 14 days (then the CAUTI was considered incomplete treatment); if the reference UCx occurred up to 24h after admission (then the CAUTI was considered present on admission); if the reference UCx occurred up to 24h after urinary catheter insertion (then the CAUTI was considered present on insertion); if a CAUTI had previously been identified within the preceding 14 days (then the CAUTI was considered in treatment).

Natural Language Processing

In the final algorithm, all clinical notes associated with catheterized patients within +/- 2 days of the reference UCx were searched for positively asserted CDC symptoms (dysuria, urgency, etc.). We developed our NLP algorithm as follows: a random sample of 500 notes from the training period containing each symptom were returned and then cropped to the 100 characters before and after the symptom (resulting in a 200 character “substring”). On this training set of substrings, by manually examining the context around the symptom, “modifiers” were defined that qualify the symptom as present, absent, or

Figure 2. CDC algorithm (numbers in green circles represent number of urine cultures leaving each node); ASB: asymptomatic bacteriuria.
other. For example: complains of dysuria (present), denies dysuria (absent), or history of dysuria (other). Synonyms and variations of symptoms were also defined, e.g. “suprapubic pain”, “tender over suprapubic”. Intermediate terms were also defined that would be allowable between the modifier and the symptom, e.g. “[modifier] left-sided [symptom]”. By running this rule-based algorithm in a sequential manner (Figure 3), that is, first looking for neutral, then negative, then positive qualifiers, we were able to reduce false positives by removing negative and neutral uses from further consideration. We also assessed confidence based on proximity of the modifier to the symptom (closer is higher confidence); we only used high confidence results. See Table 1 for examples of modifiers defined.

Our EMR also has a discreetly encoded genitourinary symptoms field. Preliminary analysis showed poor agreement between this field and clinical notes/NLP, so both approaches were combined (union) in the algorithm for maximum sensitivity (see Table 2, bottom).

Table 1. Types of modifiers defined with examples.

<table>
<thead>
<tr>
<th>Modifier category</th>
<th>N defined</th>
<th>Universal</th>
<th>Sx specific</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neutral</td>
<td>74</td>
<td>25</td>
<td>49</td>
</tr>
<tr>
<td>Absent</td>
<td>20</td>
<td>19</td>
<td>1</td>
</tr>
<tr>
<td>Present</td>
<td>91</td>
<td>87</td>
<td>4</td>
</tr>
</tbody>
</table>

Sub-types

- admission (“admitted with…”), history (“chronic…”), wrong context (“Fecal urgency”), other cause (“menses” near hematuria)
- direct (“denied…”), indirect (“no c/o of…”)
- direct (“complains of…”), indirect (“persistent…”), problem list (“#…”)

Validation and Analysis

We verified all data elements gathered by the tool against their corresponding entries in the primary sources (EMR, microbiology lab data) to ensure reliability. During the year-long prospective validation period, the tool was used in parallel to the IP’s existing workflow. This existing workflow was semi-automated in that the IP received a monthly report of patients with urinary catheters with positive culture results, and then conducted manual chart review on these patients. We compared the tool’s performance (sensitivity, specificity, PPV, NPV) to this “gold standard” of manual chart review. For the purposes of these analyses, we included only cultures ≥10^3 because any culture <10^3 would be excluded by the NHSN definition.

Results

The training set included 2460 cultures from 1421 patients; the validation set included 2547 cultures from 1567 patients. Table 2 provides patient demographics and hospital stay information as well as incidence of signs and symptoms of CAUTI, for all patients and for patients meeting various criteria (no growth, asymptomatic bacteriuria/ASB, CAUTI).
Table 2. Patient characteristics of the training population.

<table>
<thead>
<tr>
<th>Per patient stay</th>
<th>All</th>
<th>No growth</th>
<th>ASB</th>
<th>CAUTI</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Demographics</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>1519</td>
<td>904</td>
<td>173</td>
<td>95</td>
</tr>
<tr>
<td>Age, mean</td>
<td>56.1</td>
<td>55.2</td>
<td>57.5</td>
<td>56.9</td>
</tr>
<tr>
<td>Female sex</td>
<td>52.9%</td>
<td>45.8%</td>
<td>57.2%</td>
<td>76.8%</td>
</tr>
<tr>
<td><strong>Hospital stay</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cost, mean</td>
<td>$45,531</td>
<td>$46,394</td>
<td>$70,886</td>
<td>$49,279</td>
</tr>
<tr>
<td>Expected LOS (days)</td>
<td>12.0</td>
<td>12.0</td>
<td>15.3</td>
<td>12.3</td>
</tr>
<tr>
<td>LOS, mean (days)</td>
<td>15.3</td>
<td>13.7</td>
<td>28.0</td>
<td>20.5</td>
</tr>
<tr>
<td>LOS excess (days)</td>
<td>3.3</td>
<td>1.6</td>
<td>12.7</td>
<td>8.1</td>
</tr>
<tr>
<td>Case-mix index, mean</td>
<td>4.3</td>
<td>4.6</td>
<td>4.8</td>
<td>4.6</td>
</tr>
<tr>
<td>Number discharge diagnoses, mean</td>
<td>18</td>
<td>17</td>
<td>23</td>
<td>19</td>
</tr>
<tr>
<td><strong>Urinary catheter</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No of catheters/stay, mean</td>
<td>1.2</td>
<td>1.1</td>
<td>1.4</td>
<td>1.5</td>
</tr>
<tr>
<td>UCx/stay, mean</td>
<td>1.62</td>
<td>1.40</td>
<td>2.55</td>
<td>2.79</td>
</tr>
<tr>
<td>+UCx /stay, mean</td>
<td>0.46</td>
<td>0.00</td>
<td>1.60</td>
<td>1.66</td>
</tr>
<tr>
<td>Any NHSN symptom during stay</td>
<td>10.5%</td>
<td>5.6%</td>
<td>6.9%</td>
<td>45.3%</td>
</tr>
<tr>
<td><strong>Per urine culture</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of UCx</td>
<td>2460</td>
<td>1620</td>
<td>182</td>
<td>97</td>
</tr>
<tr>
<td>urinary catheter</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dwell time before UCx, mean (days)</td>
<td>6.1</td>
<td>5.4</td>
<td>9.0</td>
<td>6.8</td>
</tr>
<tr>
<td>Positive urinalysis</td>
<td>59.2%</td>
<td>51.0%</td>
<td>88.5%</td>
<td>83.5%</td>
</tr>
<tr>
<td><strong>Signs (+/- 48h of UCx)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fever unexplained by BCx</td>
<td>33.3%</td>
<td>39.3%</td>
<td>0.0%</td>
<td>63.9%</td>
</tr>
<tr>
<td>+BCx</td>
<td>10.7%</td>
<td>9.9%</td>
<td>12.6%</td>
<td>8.2%</td>
</tr>
<tr>
<td>Hypothermia</td>
<td>36.8%</td>
<td>35.1%</td>
<td>48.9%</td>
<td>27.8%</td>
</tr>
<tr>
<td>WBC, mean (x1000)</td>
<td>12.5</td>
<td>12.7</td>
<td>13.0</td>
<td>12.2</td>
</tr>
<tr>
<td>Low WBC</td>
<td>10.0%</td>
<td>10.5%</td>
<td>7.1%</td>
<td>7.2%</td>
</tr>
<tr>
<td>High WBC</td>
<td>51.7%</td>
<td>53.6%</td>
<td>48.9%</td>
<td>46.4%</td>
</tr>
<tr>
<td><strong>Symptoms (+/- 48h of UCx)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>From notes (NLP)</td>
<td>4.1%</td>
<td>2.0%</td>
<td>0.0%</td>
<td>28.9%</td>
</tr>
<tr>
<td>From discrete GU sx field</td>
<td>4.2%</td>
<td>2.7%</td>
<td>0.0%</td>
<td>26.8%</td>
</tr>
<tr>
<td>From either notes or GU field</td>
<td>6.8%</td>
<td>4.1%</td>
<td>0.0%</td>
<td>40.2%</td>
</tr>
</tbody>
</table>

ASB: asymptomatic bacteriuria; LOS: length of stay; UCx: urine culture; BCx: blood culture; WBC: white blood cell; NLP: natural language processing; GU: genitourinary

Based on comparison to gold standard of semi-automated manual review in a series of 346 catheterized patients with urine cultures $\geq 10^3$, our tool had a sensitivity of 97.1%, specificity of 94.5%, PPV of 66.7% and NPV of 99.6% (see Table 3).
Table 3. Validation (only including cultures \(\geq 10^3\)).

<table>
<thead>
<tr>
<th>Electronic surveillance tool results</th>
<th>CAUTI</th>
<th>Not CAUTI</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAUTI</td>
<td>34</td>
<td>17</td>
</tr>
<tr>
<td>Not CAUTI</td>
<td>1</td>
<td>294</td>
</tr>
</tbody>
</table>

After the tool was adopted for routine use (at which time the existing parallel workflow ceased), the IP kept a log of all CAUTI cases identified by the tool which were subsequently rejected based on chart review. Out of 100 tool-identified CAUTIs, 62 were confirmed, resulting in a PPV of 62%. See Table 4 for top reasons for rejection of tool-identified CAUTIs.

Table 4. Top reasons for rejection of tool-identified CAUTIs by IP.

<table>
<thead>
<tr>
<th>Top reasons for rejection of first 100 tool-identified CAUTIs</th>
<th>N (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Other explanation for fever/symptoms</td>
<td>12 (29%)</td>
</tr>
<tr>
<td>Pre-existing urinary tract infection (present on admission)</td>
<td>5 (12%)</td>
</tr>
<tr>
<td>Non-catheter source of microorganism (e.g. peritonitis)</td>
<td>4 (10%)</td>
</tr>
<tr>
<td>Kidney or bladder manipulation</td>
<td>4 (10%)</td>
</tr>
</tbody>
</table>

We compared the sensitivity and PPV of the tool using fever alone vs all signs/symptoms (see Table 5). Using fever alone, the tool identified 64% of all cases, compared to 97% when subjective symptoms (from NLP and GU field) were incorporated. PPV increased slightly from 60% to 62% when subjective symptoms were included. We found that subjective symptoms tended to have lower sensitivity but higher PPV for CAUTI than fever (Table 6).

Table 5. Comparison of performance characteristics of surveillance methods.

<table>
<thead>
<tr>
<th>Surveillance methods</th>
<th>Performance characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manual chart review</td>
<td>Sensitivity</td>
</tr>
<tr>
<td>Automated w/ objective (fever) only</td>
<td>59%(^{10})</td>
</tr>
<tr>
<td>Automated w/ objective + subjective sx</td>
<td>64%</td>
</tr>
<tr>
<td>PPV: positive predictive value</td>
<td>97%</td>
</tr>
</tbody>
</table>

1495
Table 6. Sensitivity and PPV of subjective symptoms among catheterized patients.

<table>
<thead>
<tr>
<th>Subjective symptoms</th>
<th>All patients</th>
<th>Patients with CAUTI</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Incidences</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Burning</td>
<td>0.5%</td>
<td>6.2%</td>
</tr>
<tr>
<td>Frequency</td>
<td>3.4%</td>
<td>21%</td>
</tr>
<tr>
<td>Dysuria</td>
<td>1.8%</td>
<td>16%</td>
</tr>
<tr>
<td>Urgency</td>
<td>2.6%</td>
<td>13%</td>
</tr>
<tr>
<td>Costovertebral angle tenderness</td>
<td>0.4%</td>
<td>2.1%</td>
</tr>
<tr>
<td>Suprapubic pain</td>
<td>0.7%</td>
<td>1.0%</td>
</tr>
<tr>
<td><strong>Objective signs</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fever &gt;38°C</td>
<td>33%</td>
<td>64%</td>
</tr>
</tbody>
</table>

Discussion

We report a novel use of natural language processing to mine unstructured full text in clinical notes to find subjective symptoms indicative of CAUTI. With this method, we were able to apply the NHSN definition as written, and analyze how it applied to our population. We found our tool to have excellent sensitivity, specificity and NPV, and moderate PPV (67%), suggesting that, given its high sensitivity, it would serve as an ideal screening tool, eliminating the vast majority of potential cases while leaving a highly enriched set of potential cases (2/3 true CAUTI, 1/3 false positives) for manual confirmation. Our sensitivity (97%) was much greater than manual review which was found in a multicenter study to average just 59%. In addition, we estimate significant time savings of >100 hours per year based on elimination of review of 300 cases per year with time savings of 20 minutes per case; estimated cost savings are up to $10,000 yearly. Additional time savings would likely result from bringing all pertinent data (e.g., cultures, clinical notes) into one display for those cases which do undergo manual review.

In our population, we found objective data alone to lack sensitivity—36% of CAUTI cases had subjective symptoms without documented fever. Others have found as little as 0-10% decreased sensitivity when using objective findings alone. It is unclear whether different documentation practices, patient populations, or study methods led to these differences, however it is likely that removing subjective symptoms from the definition may have a disproportionate effect in different settings.

In comparing two sources of subjective symptoms in our EMR (clinical notes and a discrete GU field), we found they had little overlap (see Table 2, bottom). About 4% of patients had associated symptoms according to each method, but 6.8% had symptoms when both were combined. This finding should serve as a caution to hospitals that even if their EMR has a discrete data field representing GU symptoms, symptoms often may only be charted in free text notes.

Related work

Much progress has been made in developing automated surveillance tools, however previous work has been limited to use of diagnostic codes or objective clinical data to identify CAUTI. Due to poor coding consistency, ICD codes were found to be insensitive. Increased sensitivity has been obtained with the addition of objective clinical data such as presence of fever, culture and urinalysis results.

More recent approaches have used NLP to determine the presence of a urinary catheter, finding moderate sensitivity (65%) and PPV (54%) but high specificity and NPV (>99%). Branch-Elliman et al. also used NLP in this work to identify documentation of fever, however they do not present any data to show that discretely documented body temperature is unreliable. They state in online supplemental material that capturing symptoms using NLP was not
feasible due to variation in documentation practices. In their manuscript, they state that none of the 7 cases that were missed by their algorithm were attributable to subjective symptoms, and conclude that it remains unclear whether including subjective symptoms would improve their algorithm’s operating characteristics.

Gundlapalli et al.25 show the feasibility of using NLP to detect the presence of a urinary catheter as well as a full complement of urinary symptoms, however they do not appear to have incorporated their work into a functioning surveillance tool. Their results are encouraging, showing sensitivity of 50-63% and PPV of 96-97% in determining presence/absence of urinary catheter; and sensitivity of 100% and PPV of 97% in determining presence of urinary symptoms. They conclude that it is possible to reliably extract symptoms from clinical notes and recommend further work towards standardizing and structuring documentation to improve extraction of information necessary for surveillance.

In contrast to prior work, we used NLP only for capturing subjective symptoms; we did not attempt to extract language related to whether a catheter was present or whether a fever was documented because both of these data elements are already documented discretely and reliably in our EMR. It is unclear whether a mention of a “fever” without an accompanying body temperature reading should count as a qualifying sign. Additionally, discrete documentation of the presence of an indwelling urinary catheter is increasingly common in major EMRs to enable QI initiatives to track dwell times and prompt early removal of unnecessary catheters.26 Therefore, we suggest that future work would best be directed toward extracting urinary symptoms which are likely to remain, at least partially, in free text clinical notes. As mentioned previously, our EMR includes a discrete field for urinary symptoms but it was only 67% sensitive compared to an approach that also included NLP.

More computable definitions

Definitions for HAIs originated in the pre-EHR era in which computability was not a concern. For our work, we found the NHSN CAUTI definition to be difficult to apply algorithmically for two main reasons. First, it contains several subjective elements; most notably that symptoms should not be related to “another recognized cause”. We implemented this in such a way that if a febrile patient has a positive blood culture from a different genus than present in urine culture, we discounted the fever, but it was less clear how to handle, e.g., a contemporaneous positive sputum or wound culture. It was also challenging to define rules about what would disqualify subjective symptoms such as urgency, e.g., if the patient has a known history of an enlarged prostate. The second reason for difficulty in algorithmic implementation was a lack of specific detail around timeframes. For example, how near in time must a symptom be to a positive urine culture? How soon can a patient develop a second CAUTI? What if a patient has had a previously positive non-catheterized UCx with a matching organism—how much time must pass before the patient can be said to develop a CAUTI with the same genus?

Due to these limitations, different implementations and interpretations of the definition will lead to undesired variability in reported HAI rates; at the same time these rates have taken on increased significance as markers of care quality and factors impacting payment. We suggest that HAI definitions be reconsidered for digital implementation, especially if those definitions will be used for comparison between hospitals. Computable definitions should be both meaningful (capturing true rates of infection) and able to be applied uniformly across hospitals.

We recognize that an approach like ours is not feasible on a large scale across many institutions due to variability in documentation practices and technical resources. We would therefore recommend a simpler approach that maximizes the use of discretely charted data. To be most reliable and uniform, we suggest changes to documentation practices and to the CAUTI definition itself. Specifically, hospitals should ensure that urinary catheters and urinary symptoms are documented discretely (in addition to culture results, body temperature, etc), including ongoing training and QI activities to ensure this remains reliable. Including urinary symptoms (not just fever) is important because in our population they increased sensitivity considerably. The NHSN definition itself should be modified to minimize subjectivity by identifying specific allowable “other recognized causes”, especially for fever, given that it is so nonspecific (39% of our patient population with no growth on UCx had a fever). Because fever is so non-specific, we do not suggest removing “other recognized causes” from the definition entirely as this would trade off too much specificity for increased objectivity. Finally, the NHSN definition should spell out all relevant timeframes to enable an algorithmic approach to surveillance.
CAUTI surveillance will likely never be fully automated due to its inescapable subjectivity; tools like ours should strive to prescreen and rule out as many true negatives as possible to allow IPs to efficiently review the remaining cases. One reasonable step would be for the NHSN to establish criteria for definite (not requiring further review), probable (requiring further review), and non-CAUTI (not requiring further review) to standardize the types of cases that must be reviewed.

Limitations

Our study was limited to a single center and had a relatively small validation sample. Our NLP engine should be validated in other settings with different documentation patterns and patient populations. We used our best judgement about applying the “no other recognized cause” language and timing parameters that were not specifically articulated in the NHSN definition; more detail around these elements of the definition from the NHSN would improve objectivity and reproducibility of tools like ours. However, we found that even an experienced IP struggled to determine whether patients met criteria, suggesting that completely accurate discrimination of complex real-world patients will likely remain beyond the scope of this, or any, tool’s logic.

Future work

Future work should address implementation of similar tools across a variety of settings to identify sources of variability contributing to artificial inter-institutional differences in HAI rates. Such work could inform modifications to HAI definitions to make them more meaningful and uniform when applied electronically. Additionally, real-time prospective use of such tools should be explored to enable infection control staff to interact with clinicians while diagnosis and treatment is still ongoing to facilitate continuous quality improvement.

Conclusion

CAUTI is a common and costly healthcare-associated infection yet it is challenging to measure accurately due to the complexity and subjectivity of the NHSN definition. We demonstrate a novel use of NLP to fully implement the definition, including subjective symptoms found in clinical notes. Compared to manual chart review, we show excellent sensitivity and specificity in large part due to the inclusion of symptoms. Our tool has been adopted clinically, screening out the vast majority of potential CAUTIs and saving at least a hundred hours of chart review per year. Reflecting on our experience, we suggest a generalizable approach that maximizes the use of discretely charted data—requiring changes to hospital documentation practices and ultimately to the NHSN definition itself.

References


The Problems of Realism-Based Ontology Design: a Case Study in Creating Definitions for an Application Ontology for Diabetes Camps

James C. Schuler, BSc¹, PhD student and Werner M. Ceusters, MD¹
¹Department of Biomedical Informatics, University at Buffalo, Buffalo, NY, USA

Abstract

A requirement of realism-based ontology design is that classes denote exclusively entities that exist objectively in reality and that their definitions adhere to strict criteria to ensure that the classes are re-usable in other ontologies while preserving their ontological commitment. Building realism-based ontologies is therefore quite challenging and time-consuming, demanding considerable training. Although the top-level in the form of the Basic Formal Ontology (BFO) is worked out very well, and also the upper levels of certain domains, there is still a disconnect with the bottom-up or middle-out approach which is typical, and more practical, for application ontologies. Using the development of an application ontology for diabetes management in diabetes camps as an example, we present an overview of problems trainees in realism-based ontology design can be confronted with and offer some guidelines on how to deal with them in case no ideal solution is available.

Introduction

Realism-based ontology design¹ is often criticized as being too difficult². That it is difficult indeed is witnessed by the many ontologies that attempt adherence to the principles of Ontological Realism³ but then fail at one or other stage and do not become accepted as an Open Biomedical Ontologies (OBO) Foundry ontology⁴. There are at least two reasons for this difficulty. One is that ontological commitment in realism-based ontologies is exclusively towards what has (or once had) objective ‘mind-independent’ existence³. This includes entities and their types (the latter called ‘universals’), for example organisms and cells, and certain combinations thereof (called ‘configurations’; entities, universals and configurations all being ‘portions of reality’) such as all living and deceased diabetes patients. The second one is that definitions for representational units in realism-based ontologies must adhere to very strict criteria: each one must be defined by means of an Aristotelean definition which states all necessary conditions that are also jointly sufficient and satisfiable for some entity to be an instance of the class denoted by the representational unit¹. This is extremely important for reference ontologies, i.e. ontologies that are intended to be comprehensive for a given domain and that consist of representational units that can be re-used in other ontologies while preserving their ontological commitment to the original entities in reality. Representational units in the form of terms – in contrast to, for example icons or numerical identifiers – should also have general face validity independent of context⁵.

Concept-based ontologies such as SNOMED CT⁶, in contrast, require mere adherence to formal logical criteria and an ontological commitment not to reality, but to a ‘universe of discourse’. The advantage of concept-based ontology design is that the notion of ‘concepts as shared meanings’ is perceived as more intuitive and closer to the language used in the domains that are modeled than the quite ‘abstract’ notions of ‘universals’ and ‘portions of reality’ as employed by realists. While, for example, there is no place for terms such as ‘prevented abortion’ and ‘absent nipple’ in realism-based ontologies, they can be allowed in concept-based ontologies as, from such a perspective, it does not involve rocket-science to grasp their meaning. Therefore, metaphorically, conceptualists can ‘see’ both absent nipples and persons with absent nipples, while realists can in that context only see persons with absent nipples. While the challenge for conceptualists is not to misclassify prevented abortions and absent nipples as special kinds of abortions, resp. nipples, the challenge for realists is to find ways for formally describing in what way persons with absent nipples are different from persons with prevented abortions without resorting in these descriptions to absent nipples and prevented abortions⁷. Such considerations are very hard for trainees without a solid education in philosophy.

Whereas strict adherence to the realism-based principles is an absolute requirement for reference ontologies³, one might wonder whether that must also be the case for application ontologies, i.e. ontologies which are designed to accomplish some specified local task or application within a specific context. Application ontology development usually follows a bottom-up or middle-out approach: fundamental terms in the application domain are addressed first before moving up to more generic and down to more specific terms. One could argue that for such ontologies it is sufficient to introduce terms that have face validity only within the interface terminology⁸ of the intended application. One might even be tempted to simplify definitions to what is relevant for the logical semantics of the underlying
application as advocated by concept-based ontologies. With less stringent criteria advanced ontological and philosophical analyses can be avoided, thus leading to quicker development times. The drawback is, however, that such ontologies are more prone to contain errors\textsuperscript{9, 10}, and that formally underspecified terms with local face validity hamper their re-usability in other semantic applications so that data collected through them cannot automatically be merged and integrated with data coming from other applications, even if they cover the same domain. Therefore, it is still worthwhile to define the mid-level application domain representational units of application ontologies in terms of upper-level units from realism-based ontologies. Unfortunately, here lies an additional level of complexity for trainees in realism-based biomedical ontology design: whereas the top-level in the form of the Basic Formal Ontology (BFO) is worked out very well\textsuperscript{1}, there is a shortage of adequate mid-level representational resources that follow realism-based principles strictly, perhaps with the exception of the Ontology for General Medical Science (OGMS)\textsuperscript{11}.

In this paper, we demonstrate the sort of problems this introduces for trainees on the basis of a use case: an application ontology for glycemic control of attendees of diabetes camps. We offer some guidelines on how to deal with them in case no ideal solution is available. The goal is to identify and prioritize issues that needs to be addressed by experts to make realism-based ontology development more accessible.

**Background**

The American Diabetes Association (ADA) holds many residential camps (‘diabetes camps’) for children age 7 to 18 with diabetes throughout the United States. These are typical summer camps in the sense the campers do all the typical camp activities over the course of a week, including activities such as arts and crafts, singing songs, building campfires, performing skits, rock climbing, archery, athletics, eating meals in a dining hall, and participating in an evening program revolving around a theme for the week. The atypical aspect of these camps is that all of the campers, and much of the staff and counselors, have diabetes. These camps have been held for over 90 years\textsuperscript{12}.

Managing diabetes is a challenging task for many people with diabetes in a normal day-to-day scenario\textsuperscript{13}. Adding the spectrum of high activity and a different routine than normal at camp only increases the difficulty of diabetes management. To keep campers safe, the management of the campers’ diabetes is achieved by means of teams composed out of medical professionals (e.g., physicians, nurses, and pharmacists, and medical professionals in training such as medical, nursing, pharmacy, and physician assistant students), the counseling staff, and most importantly, the campers themselves. Presently at camp, medical professionals working with the campers record key data on paper forms which describe the camper’s glycemic control and management (Figure 1).

![Figure 1: The current paper-based medical records used at Camp Aspire](image)

The tracking of this data allows the medical professionals to perform analysis and make adjustments to the various aspects of care which determine glycemic control\textsuperscript{12}. Moreover, the tracking on records is important for communication between staff members, and engages staff to provide continuous safe and high-quality diabetes management.
The current state of medical record keeping and tracking of vital healthcare data at diabetes camp is archaic and thus ripe for innovation. The data tracked on paper medical records include, but are not limited to: time of observation, blood glucose level, carbohydrates consumed in grams, doses of insulin, basal insulin delivery rate, and urine/blood ketone level. The paper records are cheap, easily modifiable, and relatively secure. However, there are many downsides to their use. For example, paper records only modestly facilitate tracking and analysis of data and are ill-suited to detect patterns indicating potentially dangerous developments. Furthermore, the records can often be unreadable, and unavailable in a reasonable time frame in an emergent situation. Finally, there is no standard of use of paper records, or fields present in those records throughout the camps run by the ADA. Indeed, no two camps use the same forms for tracking of data (Figure 2). While each camp has its own unique forms, there is a standard within a camp, there is no standard between camps, as, for example, certain entities are referred to by different terms and possible ‘allowed’ values as instances. This lack of standard hinders research about diabetes management at all camps as data are not easily collated and analyzed what hampers the advancement in elucidation and understanding of the factors relevant to diabetes management in a camp setting and potentially places campers at risk for severe emergency events related to their diabetes. It is unknown to what extent an electronic medical record system implemented at a diabetes camp would impact the ease of recording and tracking data, elucidation of patterns in the data, and thus would increase the safety, care, and efficiency and efficacy of diabetes management at camp. With an ontology-based mobile app for diabetes care in such camps, this question might be answerable.

Methods

One option for the final assessment of students in the 500-level 3-credits course ‘Introduction to Biomedical Ontologies’ co-organized by the Departments of Biomedical Informatics and Philosophy of the University at Buffalo is the development of an ontology. The first author of this paper, a student in this course, being a regular collaborator at diabetes camps, envisioned the benefits of a small custom-made ontology-based electronic medical record app for glycemic control in such camps. It should be able (1) to support collecting, recording, presenting and managing data about the campers’ so that (2) the medical professionals would be able to analyze and interpret the data collected more easily. To keep it manageable, it would (3) allow collection of the data fields pertinent to the diabetes management of adolescents in a camp setting as, thus covering what is currently recorded on paper. This custom, micro electronic medical record would have less functionality than typical electronic medical records used in traditional inpatient or outpatient settings, but was hypothesized to have high value for use at diabetes camps. For reasons of reusability and comparability of data and integration with other relevant technologies, another design criterion envisioned would be (4) to keep the ontology – henceforth called the ‘Diabetes Camp Ontology’ (DCO) – underlying the system compatible
with the growing number of realism-based ontologies\textsuperscript{1}. The application would, for instance, be able to deal with the following scenario from camp which forms the core of the relationship between a camper and the medical staff, and the majority of what is currently recorded in the paper records: ‘\textit{Rob Smith checks his blood sugar when he wakes up on Monday. His sugar is 208. Rob’s medical staff for the week, Jim, records his blood sugar and based on this blood sugar, calculates Rob needs to take 2.8 units of insulin, and Rob takes his insulin. Immediately afterward, Rob Smith goes to the dining hall to eat breakfast. At breakfast he eats 65 grams of carbohydrates. Based on the number of carbohydrates Rob eats, Jim calculates Rob should deliver 4.3 units of insulin, so Rob does.’’ Such a scenario would be recorded in the paper records as displayed in Figure 3, a part of the form displayed in Figure 1.

In a first step, the student collected from the available paper record templates of various diabetes camps all the data elements used for the campers’ follow ups. In a second step, he inspected existing ontologies for the presence of representational units denoting types of entities in terms of which the entities represented by the terms that were collected could be described following the provisions of the Relation Ontology\textsuperscript{15}. Whenever in a feeder ontology an representational unit was found that denotes an entity also referenced by one of the collected data elements, that representational unit became part of the DCO, otherwise, a new representational unit was defined.

The resulting ontology and the documentation thereof – a short paper and presentation were part of the assignment – was then inspected by the instructor for adherence to realism-based quality principles\textsuperscript{1-3} and the total effort graded. At that stage, it became clear that failure to adhere was not only because of gaps in understanding on the side of the student, but also because of gaps in the available resources. The definitions and terms of the application ontology were therefore afterwards in a short turn-around modified and improved in discussion with the instructor to the extent possible and then analyzed together with the student’s arguments for his decisions with the goal (1) to find out what problems this student – and perhaps, to be determined through future studies, students in general – encountered with the application of realist principles, (2) to determine the root causes for these problems, and (3) to propose ‘easy fixes’, if any at all, that can be applied to definitions so that the corresponding classes, despite coming from an application ontology, can be more reliably, though not necessarily automatically, compared with relevant ones in other application ontologies within the same or a closely related domain.

Results

Building this application ontology turned out to be a difficult task and several issues were encountered. The biggest problem was not only to find adequate ontologies, but also adequate terms within them. The Basic Formal Ontology (BFO)\textsuperscript{4} as a top-level ontology, and the Information Artifact Ontology (IAO)\textsuperscript{16}, the Ontology for General Medical Science (OGMS)\textsuperscript{11}, the Ontology for Biomedical Investigations (OBI)\textsuperscript{17}, and the Ontology of Medically Related Social Entities (OMRSE)\textsuperscript{18} as high-level domain ontologies turned out to be good choices despite the wide range of incompatible options for existing representational units (RU) to choose from. Whereas the ‘middle-out’ development approach, i.e. starting with terms that need to be part of the interface terminology of the application in line with the envisioned scenarios, provided a good basis, way more terms than originally anticipated needed to be included as well and this for purposes of (1) crafting adequate definitions – a challenge in its own right, (2) providing additional functionality to the app as compared to what is possible with paper forms and (3) the ones required to drive the logic of this app. A major challenge was to identify the exact ontological types of the entities that figured in the scenarios we had in mind, a problem which not only seems to be a problem for students and trainees in realism-based ontology development, but also for the experts. For the purposes of this paper, detailed analyses are limited to representational units and definitions related to camps and the roles played by participants (Table 1) and to blood glucose levels and insulin administration (Table 2). For each RU, a textual definition precedes a semi-formal definition. The latter uses the formatting principles of the Relation Ontology\textsuperscript{15}: (1) each representational unit starts with the acronym of the ontology in which it is included, (2) SMALL CAPS is used for types, (3) \textit{italic bold font} for formal relationships. This formatting is, where important for correct understanding what we are referring to, also used in the discussion section.

Discussion

Finding relevant domain ontologies. Several projects report to have attempted the creation of an ontology for diabetes, its treatment, its complications or its diagnosis often with the eventual end goal to create clinical decision support systems to aid in numerous diabetes related issues for both patients with diabetes and healthcare providers\textsuperscript{19-25}.

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{image.png}
\caption{Part of a diabetes camp form}
\end{figure}
Table 1. Terms and definitions related to camps and roles played by participant

<table>
<thead>
<tr>
<th>Class and definitions</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>DCO: CAMP</strong></td>
</tr>
<tr>
<td>D1a: an organization which provides people (often children) specialized facilities and activities, often related to a central theme or goal, and often providing overnight accommodations.</td>
</tr>
<tr>
<td>D1b: OBI:PLANNED PROCESS which (1) realizes the BFO: ROLE of some OBI: ORGANIZATION at some BFO: SITE which comprises overnight accommodations, and (2) has process parts (2a) in which instances of OBI: HOMO SAPIENS participate and (2b) some of which are of a sort as described in instances of IAO: OBJECTIVE SPECIFICATION describing some central theme or goal.</td>
</tr>
<tr>
<td><strong>DCO: MEDICALLY-FOCUSED CAMP</strong></td>
</tr>
<tr>
<td>D2a: a camp where an instance of some disease inheres in all instances of attendees, and whose mission often includes creating a sense of normalcy for the attendees and teaching self-management of the disease, while providing necessary and critical medical care.</td>
</tr>
<tr>
<td>D2b: DCO: CAMP which (1) has attendees of OBI: HOMO SAPIENS all of which are inherers of some instance of some OGMS: DISEASE $D$, (2) has process parts some of which are of a sort as described in instances of IAO: OBJECTIVE SPECIFICATION describing the objectives to create a sense of normalcy for HS and to teach self-management of $D$, and (3) has process parts some of which provide necessary medical care to HS.</td>
</tr>
<tr>
<td><strong>DCO: DIABETES CAMP</strong></td>
</tr>
<tr>
<td>D3a: medically-focused camp, where the specific which inheres in all instances of attendees is diabetes.</td>
</tr>
<tr>
<td>D3b: DCO: MEDICALLY-FOCUSED CAMP which has attendees of OBI: HOMO SAPIENS all of which are inherers of some instance of diabetes.</td>
</tr>
<tr>
<td><strong>DCO: CAMP ASPIRE</strong></td>
</tr>
<tr>
<td>D4b: DCO: DIABETES CAMP organized by the OBI: ORGANIZATION called ‘American Diabetes Association’.</td>
</tr>
<tr>
<td><strong>DCO: DIABETES CAMP CAMPER ROLE</strong></td>
</tr>
<tr>
<td>D5a: a patient role borne by a human being being realized by participating in the events held by some diabetes camp.</td>
</tr>
<tr>
<td>D5b: OMRSE: PATIENT ROLE which inheres in OBI: HOMO SAPIENS which are attendees of some DCO: DIABETES CAMP.</td>
</tr>
<tr>
<td><strong>DCO: DIABETES CAMP MEDICAL STAFF ROLE</strong></td>
</tr>
<tr>
<td>D6a: a health care provider role borne by some human being and realized by managing or directing of management the diabetes care of a diabetes camp camper.</td>
</tr>
<tr>
<td>D6b: OMRSE: HEALTH CARE PROVIDER ROLE which inheres in OBI: HOMO SAPIENS and which is realized by OBI: PLANNED PROCESS(es) of managing or directing the diabetes care of DCO: CAMPER(s).</td>
</tr>
<tr>
<td><strong>DCO: CAMPER</strong></td>
</tr>
<tr>
<td>D7a: human being which is bearer_of some diabetes camp camper role.</td>
</tr>
<tr>
<td>D7b: OBI: HOMO SAPIENS which is bearer_of DIABETES CAMP CAMPER ROLE.</td>
</tr>
</tbody>
</table>

The existence of them may raise the question whether yet another diabetes ontology is needed? Closer inspection reveals that only few have potentially pertinent terms and relations\textsuperscript{26, 27}. Unfortunately, rarely do these projects follow the realism-based guidelines and only one\textsuperscript{28} expresses interest in following guidelines to become a part of the OBO Foundry. While all are meaningful work within a concept-based approach and for their own specific uses, none encapsulate the reality of diabetes care and management in a camp setting. Therefore, for our purposes, a new realism-based ontology with high potential for integration with OBO Foundry compliant ontologies to address the specific needs of serving as a model of reality of caring for children with diabetes in a camp setting is indeed needed.

What terms to include? In line with the purposes the ontology has to serve, a first type of representational units to be included are those that represent entities to which the terms from the interface terminology of the application stand in an aboutness relation\textsuperscript{16}. This includes entities referenced by the terms on the paper forms used in diabetes camps, as well as the ones for which the application might offer some possible values. How clear cut this might seem, it isn’t in light of the specific ontological commitment of the BFO and the ontologies that intend to be compatible with it. An example is what is denoted by the abbreviation ‘BG’ on the form in Figure 1, i.e. blood glucose levels, and what is expected to be entered as values to denote a specific blood glucose level. One could allow just numbers such as ‘98’ and ‘289’, standing for ‘blood glucose level of 98 mg/dl (milligrams per deciliter)’ and ‘blood glucose level of 289 mg/dl’ resp. But does BFO commit to the existence of numbers with objective, ‘mind-independent’ existence? This is to say, if we write ‘98’, does this term denote some $x$ in reality? If so, would that $x$ be an instance of the type number?
Table 2. Terms and definitions related to blood glucose and insulin administration

<table>
<thead>
<tr>
<th>Class and definitions</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>DCO:BLOOD GLUCOSE LEVEL</strong></td>
<td></td>
</tr>
<tr>
<td>D8a: A physical quality which inheres in blood by virtue of the concentration of glucose in the serum.</td>
<td></td>
</tr>
<tr>
<td>D8b: BFO:QUALITY which <strong>inheres in</strong> a portion of blood by virtue of the concentration of glucose in the serum.</td>
<td></td>
</tr>
<tr>
<td><strong>DCO:BLOOD GLUCOSE READING</strong></td>
<td></td>
</tr>
<tr>
<td>D9a: a scalar measurement datum that is the specified output of measuring glucose concentration in serum, often displayed on a glucometer.</td>
<td></td>
</tr>
<tr>
<td>D9b: IAO:SCALAR MEASUREMENT DATUM that <strong>is the specified output</strong> of a OBI:MEASURING GLUCOSE CONCENTRATION IN BLOOD SERUM, often displayed on an OBI:GLUCOMETER, and <strong>which is about</strong> some DCO:BLOOD GLUCOSE LEVEL</td>
<td></td>
</tr>
<tr>
<td><strong>DCO:INSULIN PUMP</strong></td>
<td></td>
</tr>
<tr>
<td>D10a: a device which stores insulin and participates in a process of subcutaneous injection with or without human intervention, according to some pump setting datum.</td>
<td></td>
</tr>
<tr>
<td>D10b: OBI:DEVICE in which <strong>inheres</strong> the BFO:FUNCTION(s) (1) to store insulin and (2) to participate in instances of OBI:SUBCUTANEOUS INJECTION, according to some DCO:PUMP SETTING DATUM</td>
<td></td>
</tr>
<tr>
<td><strong>DCO:INSULIN PEN</strong></td>
<td></td>
</tr>
<tr>
<td>D11a: a device which stores insulin and participates in a process subcutaneous injection with human intervention.</td>
<td></td>
</tr>
<tr>
<td>D11b: OBI:DEVICE in which <strong>inheres</strong> the BFO:FUNCTION(s) (1) to store insulin and (2) to participate in instances of OBI:SUBCUTANEOUS INJECTION carried out by an instance of OBI:HOMO SAPIENS.</td>
<td></td>
</tr>
<tr>
<td><strong>DCO:INSULIN ADMINISTRATION SITE</strong></td>
<td></td>
</tr>
<tr>
<td>D12a: a part of the host body and entrance point where insulin is administered</td>
<td></td>
</tr>
<tr>
<td>D12b: BFO:SITE which <strong>is part of</strong> a DCO:CAMPER through which insulin is administered</td>
<td></td>
</tr>
<tr>
<td><strong>DCO:INSULIN INJECTION SITE</strong></td>
<td></td>
</tr>
<tr>
<td>D13a: an insulin administration site that is used for the injection of insulin</td>
<td></td>
</tr>
<tr>
<td>D13b: DCO:INSULIN ADMINISTRATION SITE that is used for injection of insulin</td>
<td></td>
</tr>
</tbody>
</table>

Or would that x be a type itself so that there are entities which are instances of that x. Are then 98 apples, a heart rate of 98 beats per minute and a blood glucose level of 98 mg/dl all instances of x? Trainees might have these questions, yet find no answers in available documentation, or perhaps even more confusing, different positions which nevertheless all fall under a realist perspective. It is of course here that pragmatism comes into play by leaving out anything not directly relevant to the application while still staying under the realist perspective. So it is documented that BFO commits to entities on the side of the patient that are of types such as ‘blood glucose level of 98 mg/dl’ – they existed before anybody knew about glucose in blood, how to measure it, etc. – and all these entities are also of the type ‘blood glucose level’. Because blood glucose level is a quality, it allows convenient reasoning and accurate representations of reality at instance levels. For example, consider an instance of camper, Rob Smith. Rob Smith’s blood glucose level exists, and is an instance of blood glucose level for all the time during which Rob Smith exists. At some time, t₁ = 7:30 am, Rob Smith’s blood glucose level instantiates a blood glucose level of 208 mg/dl. At some later time, t₂ = 10:30 am, Rob Smith’s blood glucose level instantiates a blood glucose level of 115 mg/dl. Software can deal appropriately with these issues without the need to overload the ontology.

Should one include terms such as ‘hyperglycemia’ or ‘hypoglycemia’? These are medical terms describing blood glucose which often have strict definitions and cut-off values in certain settings and situations which often dictate action or treatment protocols. For example, in the emergency department of a hospital if a patient’s blood glucose is described as ‘hypoglycemia’ (physiologically low), it might prompt a provider to obtain some sugary food or drink for the patient to raise the blood glucose level back to ‘euglycemia’ (physiologically normal). Kids attending diabetes camp may have different blood glucose levels which they describe as ‘high’, ‘normal’, or ‘low’, a statement often referring to a blood glucose level in some range of blood glucose levels which is decided by the child’s physician. Furthermore, ‘High glycemia’ and ‘hyperglycemia’ do not always coincide, and neither ‘normal glycemia/euglycemia’, and ‘low glycemia/hypoglycemia’. One might be tempted to include terms such as ‘within target range blood glucose’, ‘above target range blood glucose’, and ‘below target range blood glucose’. But then, targets might be different for each patient. Thus in two distinct kids might inhere blood glucose levels which are of exactly the same determinate type, yet one could be within target, while the other one wouldn’t be. The issue here is that such terms when used on the form of some patient are not ‘just’ about the blood glucose level of that individual patient, but also
about the population of human beings or about the disease course of that patient as defined in the OGMS\textsuperscript{11}. When it is decided to use these terms, it is mandatory that the ontology contains a second type of representational units, i.e. those which are required to provide adequate ontological definitions\textsuperscript{29}. But there is not always the need for these terms to become also part of the interface terminology of the application.

The third type of representational units to be included are those required for the app to add additional functionality as compared to what is possible with paper forms. Thus it is critical for the functioning at diabetes camp and the representation thereof in the ontology for the app that certain roles be defined and used in precise manner. For example, when reviewing the records of diabetes management at camp, it is important to know who ordered what dose of which medication, as certain roles confer certain powers to individual persons to make certain decisions regarding the diabetes management of campers. Consider the following: a medical director of a diabetes camp may order all injections of insulin be cut by 25\% for a certain day because there will be a large amount of activity in the afternoon. This action by a medical director will likely result in the campers’ blood glucose level remaining in a safe range, as activity and insulin both act to decrease blood glucose level. A specific general medical staff working with a specific camper may decide that camper should cut the dose of an injection by 25\%, but cannot order such a change on a camp-wide scale. In table 1, D5 and D6 are examples of roles. Important subtypes of DCO:DIABETES CAMP MEDICAL STAFF ROLE include: diabetes camp medical director role, diabetes camp medical coordinator role, diabetes camp lead medical staff role, and diabetes camp general medical staff role. This then leads to the introduction of a fourth type of representational unit: the ones required to drive the logic of the application. Again, care must be taken, to the extent possible, that these units denote entities a realism-based ontology can commit to. In this case, it can be achieved by reference to entities of the type BFO:RELATIONAL QUALITY, such as the relational quality of authority between a person in whom an instance of diabetes camp medical director role inheres and other human beings in whom instances of other diabetes camp medical staff roles inheres. Analogously, there is a relational quality of authority between a person in whom an instance of diabetes camp medical coordinator role inheres and other human beings in which instances of other diabetes camp medical staff roles inheres, except for the person in whom an instance of diabetes camp medical role inheres. This allows to create an authority hierarchy among the diabetes camp medical staff.

Finding relevant mid-level ontologies. A major problem trainees are confronted with when attempting to define mid-level classes of their application ontology in terms of existing classes in other ontologies is the wide range of options, unfortunately mostly incompatible ones, which are suggested when searching through ontology repositories such as the BioPortal (http://bioportal.bioontology.org/) and Ontobee (http://www.ontobee.org/). Even more possibilities become available through literature searches.

Clearly diabetes camps in particular and camps in general are organized. Does that make such camp an ‘organization’? Over 30 existing ontologies use that term for a variety of distinct entities, the most frequent ones being: (1) the process of organizing something, e.g. organizing AMIA 2017, (2) the result of such process, e.g. the conference AMIA 2017 itself, and (3) some collection of people (and for some ontologies also other entities such as buildings, statutory documents, …), such as (3a) those involved in organizing in the first sense, e.g. the AMIA 2017 Organization Committee and (3b) everything involved once AMIA 2017 is ongoing such as its attendees, hosts, people involved in preparing the halls and rooms, etc. The student, in our use case, selected option (3a) which is the one at first sight suggested by the OBI, but with which the instructor, as we will explain further, disagreed (Table 1, D1a and D1b).

A similar situation occurs for ‘blood glucose level’ where the term is used to denote (1) an act of measuring glucose levels, (2) the value obtained through such measurement or (3) a quality of the blood independent of whether it is measured or not. In this case, the student correctly included all three types of entities in his analysis and definitions (Table 2, D9). Although several of the terms presented in the Clinical Measurement Ontology\textsuperscript{26} may be of use to this work, including ‘blood glucose level’ itself, the OBI describes better the relations between the various entities. For example, the original OBI publication describes in great detail the process of collecting a sample of blood and analyzing the glucose content in the sample\textsuperscript{17}. Thus, this ontology contains a class for OBI:MEASURING GLUCOSE CONCENTRATION IN SERUM so that the resulting output DCO:BLOOD GLUCOSE READING could be defined as a IAO:SCALAR MEASUREMENT DATUM which is linked to the former through the relationship has\textunderscore specified\textunderscore output.

Decisions on which ontologies to build further on should not only be taken on the basis of suitability of individual representational units for one’s purpose, but on the quality of the ontology as a whole. Acceptance of the ontology in the OBO Foundry is often used as a quality argument but it needs to be taken with a pinch of salt. So was at some point the Vaccine Ontology\textsuperscript{30} (VO) considered useful for the DCO. However, the VO considers the USA and Canada subtypes of ‘country’ rather than instances, which is too egregious a mistake for the quality label to be taken serious.
For the use case described, several ontologies with realism-based foundations are used. OMRSE is an ontology ‘cover[ing] the domain of social entities that are related to health care, such as demographic information and the roles of various individuals and organizations’\textsuperscript{18}. Notably, subtypes of the OMRSE:HEALTH CARE ROLE include OMRSE: PATIENT ROLE and OMRSE:HEALTH CARE PROVIDER ROLE as used in D5 and D6.

The OBI\textsuperscript{17} turned out to be very useful for many definitions, for example to address ways in administering insulin. Historically, insulin is introduced into the body via a subcutaneous route. Most campers use insulin pumps, and a minority use insulin pens or syringes. Insulin pens are made of cartridges of insulin and a screw top upon which one can screw a needle to inject the insulin. Insulin pumps are small (pager-size and shape), battery-operated entities which can push insulin held in a syringe through a tube. The tube is inserted into the hypodermis of a person with a needle. A similarity between the two methods of insulin injection is that insulin pumps and insulin pens all have as part an OBI:SYRINGE. A key difference is that every instance of insulin pump has part some instance of OBI: COMPUTER. Since there are several sorts of insulin pumps, it is important for the diabetes camp medical staff to be aware of the manufacturer of a camper’s insulin pump, and what model it is, such that the medical staff will then be able to navigate the appropriate menu options in the envisioned application. For example, Rob Smith’s Medtronic Paradigm 751 would be an instance of DCO:MEDTRONIC INSULIN PUMP MODEL PARADIGM 751, which is a subtype of DCO:MEDTRONIC INSULIN PUMP, which is a type of DCO: INSULIN PUMP (Table 2, D10) which OBI:is manufactured by value Medtronic Minimed which itself is an instance of OBI: ORGANIZATION. Whether a DCO: CAMPER receives insulin via a DCO: INSULIN PUMP or DCO: INSULIN PEN, he or she is participating in the process OBI: SUBCUTANEOUS INJECTION.

The IAO makes it possible to keep entities on the side of the patient and caregivers, e.g. blood glucose levels, strictly separate from entities to describe such entities, e.g. blood glucose readings on the screen of a glucometer. So there is IAO: SETTING DATUM defined as ‘a datum that denotes some configuration of an instrument’. Insulin pumps are devices with programmable settings. Examples of settings which are programmed by a user include: insulin to carbohydrate ratio, correction factor (sensitivity), basal rates, and blood glucose level targets. The DCO includes these terms referring to the analogous setting datum as subtypes of a new term DCO: PUMP SETTING DATUM (Table 2, D10), which itself is a subtype of an IAO: SETTING DATUM. These setting datum provide the information to determine how much insulin should be given in a given situation, for example based on current blood glucose level of a camper and carbohydrates eaten by a camper. Other representational units based on IAO include DCO: BLOOD GLUCOSE READING (Table 2, D9) as a subtype of IAO: SCALAR MEASUREMENT DATUM.

Deciding what types of entities selected terms denote. This aspect of realism-based ontology development is undoubtedly the hardest nut for trainees to crack as it does not allow objects to be ‘anything perceivable or conceivable’ as advocated in concept-based approaches\textsuperscript{31}. So it would be tempting to create subtypes of OBI: SUBCUTANEOUS INJECTION for what terms such as ‘food bolus injection’ and ‘correction bolus injection’ would denote. A food bolus injection would be administered when a camper eats carbohydrates, and a correction bolus injection when a camper’s blood glucose is high. However, it is easy to conceive two processes of bolus injection which are of exactly the same type (same sort of needle, same concentration, same type of injection site, …) whereby only the intention of the clinician would be different, but such intention is not part of the process of injecting. Or consider a carbohydrate count as subtype of IAO: SCALAR MEASUREMENT DATUM on the argument that such count is the result of a measurement of carbohydrates in food eaten. Carbohydrates are counted based on standard guidelines from the FDA and the nutrition labels on food. A standard may be: a medium size apple has 15 grams carbohydrates, and 1 slice of bread has 15 grams of carbohydrates. The question the trainee has to answer here is whether this method is more and estimate than a measurement, and whether OBI accepts estimates as kind of measurements.

As a last example, an important consideration when giving an injection of insulin is the location of the injection on the body. This is important to know, as different spots on the human body absorb insulin to varying degrees. For example, insulin injected into the abdomen is absorbed faster into the bloodstream than insulin injected into the buttocks, and a camper’s blood glucose level may be influenced by choosing to perform injections into certain locations. One could follow the example of the Vaccine Ontology\textsuperscript{30} (VO), a domain ontology which represents the entities in reality as pertaining to vaccines and vaccinations. Since vaccines can be given in various locations in the human body and via various methodologies, for example, an injection of a vaccine into the deltoid muscle of the arm, or an inhalation of a vaccine through the nose, the VO defines a VO: VACCINATION SITE, and a subtype VO: VACCINE INJECTION SITE. Thus in a similar way DCO: INSULIN INJECTION SITE is defined (Table 2, D12, D13). The question to be answered here is whether such terms denote truly extra entities or whether such terms are mere ‘ways of talking’. After all: there is no injection site in BFO sense unless the injection has been given and one can therefore not use it to describe where an injection is to be given; the ‘site used for injection’ is simply a body part. It is a bit different for inhalation, because in that case, the body part (nostril, mouth, …) is indeed a site in BFO sense. But would that site
behave suddenly an instance of ‘insulin inhalation site’ after inhalation of insulin? What did change after the inhalation that warrants this? It is acceptable, for pragmatic reasons, to introduce what are called ‘defined classes’ for these purposes, but the resulting ontologies are then not ‘realism-based’ in the strict sense as defined by some authors.

Closely related to this is determining whether entities are types or instances. Instances, also called ‘particulars’ or ‘individuals’, are entities that carry identity and do not come in various sorts. Canada and the USA are instances: there are no two different kinds of Canada or different kinds of USA. Is ‘Camp Aspire’, acronym for ‘Always Sharing Priceless, Inspirational, Rewarding Experiences’ as organized by the ADA a type or an instance (Table 1, D4)? Interestingly, that there is another ‘kind of’ ‘Camp Aspire’, acronym for ‘Autism Social Skills Program for Interaction and Relationship Education’ organized by the Autism Society of Central Illinois is not an argument for typehood but a simple case of homonymy. But that the ADA ‘Camp Aspire’ is organized annually, in distinct locations and at some locations at different times are all arguments for typehood, each specific camp, at a specific site and during a specific time period being an instance of that type. And of course, both the ADA ‘Camp Aspire’ and the autism ‘Camp Aspire’, which also is organized multiple times, are instances of DCO:CAMP. This raises the question whether ‘camp’ should indeed be defined in DCO, or whether it should not move up to a higher level in the hierarchy of domains.

What if the experts don’t know? To what ontology a representational unit ideally belongs is not a clear cut matter. It is proposed that the distinction between occurrents (entities that evolve, such as processes) and continuants (entities that don’t have temporal parts), and, amongst the latter, between independent continuants (e.g. objects) and dependent continuants (e.g. qualities, roles, functions, …) should be a guide. The Gene Ontology has been used as an example of how that would work. But what if even the experts disagree – or can’t decide – about what the highest subsumer of some type might be. It was tempting to define DCO:CAMP (Table 1, D1) as a type of OBI:ORGANIZATION until the fine print was read: ‘The definition summarizes long email discussions on the OBI developer, roles, biomaterial and denrie branches. It leaves open if an organization is a material entity or a dependent continuant, as no consensus was reached on that. The current placement as material is therefore temporary, in order to move forward with development.’ (http://www.ontobee.org/ontology/OBI?iri= http://purl.obolibrary.org/obo/OBI_0000245). Also using OBI:HOMO_SAPIENS to denote human beings is not unproblematic as experts seem to disagree about whether this OBO-term just lacks face validity by using the name of a specific species (an individual) for a type or truly represents a species. The difference is important as there are no different kinds of homo sapiens (yet?).

Conclusion

Building realism-based ontologies is a daunting task. Although the first version of BFO was created fifteen years ago, the current version consists of a mere 35 classes with a maximum hierarchy depth of 5. This is not only because BFO as top-level reference ontology is by design restricted to representational units which are relevant to all domains and thus is expected to be rather small, but also because of the extremely detailed philosophical analyses that were required to bring it into its present shape. Although not the same level of philosophical enquiry is needed for application ontologies, it cannot be avoided. The number of realism-based ontologies is growing enormously, but so also the number of inconsistent uses, even incompatibilities. Although ‘Building Ontologies with Basic Formal Ontology’ is a useful resource, it does not contain all the answers to practical questions trainees are confronted with.

References


31. Shortliffe EH, Cimino JJ. Biomedical informatics : computer applications in health care and biomedicine2014.


Reconciliation of multiple guidelines for decision support: a case study on the multidisciplinary management of breast cancer within the DESIREE project

Brigitte Sérousso, MD, PhD1,2,3, Gilles Guézennec, MSc1, Jean-Baptiste Lamy, PharmD, PhD4, Naiara Muro, MSc1,4,5, Nekane Larburu, PhD4,5, Booma Devi Sekar, PhD6, Coralie Prebet7, Jacques Bouaud, PhD1,8

1 Sorbonne Universités, UPMC Univ Paris 06, INSERM, Université Paris 13, Sorbonne Paris Cité, UMR_S 1142, LIMICS, Paris, France
2 AP-HP, Hôpital Tenon, Département de Santé Publique, Paris, France
3 APREC, Paris, France
4 eHealth and Biomedical Applications, Vicomtech-IK4, Donostia-San Sebastian, Spain
5 Biodonostia, Donostia-San Sebastian, Spain
6 Computer Science Research Institute, Ulster University, Newtownabbey, United Kingdom
7 AP-HP, Hôpital Tenon, Service d’Oncologie Médicale, Paris, France
8 AP-HP, DRCI, Paris, France

Abstract

Breast cancer is the most common cancer among women. DESIREE is a European project which aims at developing web-based services for the management of primary breast cancer by multidisciplinary breast units (BUs). We describe the guideline-based decision support system (GL-DSS) of the project. Various breast cancer clinical practice guidelines (CPGs) have been selected to be concurrently applied to provide state-of-the-art patient-specific recommendations. The aim is to reconcile CPG recommendations with the objective of complementarity to enlarge the number of clinical situations covered by the GL-DSS. Input and output data exchange with the GL-DSS is performed using FHIR. We used a knowledge model of the domain as an ontology on which relies the reasoning process performed by rules that encode the selected CPGs. Semantic web tools were used, notably the Euler/EYE inference engine, to implement the GL-DSS. “Rainbow boxes” are a synthetic tabular display used to visualize the inferred recommendations.

Introduction

Breast cancer is the most common cancer among women in France with about 50,000 new cases per year. After having doubled between 1985 and 2005, the incidence rate of breast cancer seems now to be stabilized. However, if it is decreasing for women aged 50-79 that benefit from the nationally organized screening, it has increased of more than 60% for the women aged 30-39 and 40-49\(^1\). With 12,000 deaths per year (figures consolidated in 2012), the mortality rate of breast cancer in France is decreasing. Breast cancer is one of the cancers with the best survival rate at five and 10 years (87%, resp. 76%). This may be explained by the evolution of therapeutics and the development of endocrine and targeted agents as well as to the reduction of menopausal hormone therapy prescription. However, breast cancer comprises a complex and heterogeneous group of diseases at the clinical, morphological, and molecular levels. For some of the subtypes, especially the triple-negative and HER2+ breast cancers, margins for improvement are both possible and necessary\(^2\).

Clinical practice guidelines (CPGs) are elaborated to provide best evidence-based recommendations for the management of patients with specific conditions. Studies showed that implementing oncology CPGs does improve clinical outcomes in both overall and recurrence-free survivals\(^3\). This is especially true in the case of breast cancer\(^4\). However, despite the publication of CPGs and the provision of state-of-the-art recommendations, cancer management remains subject to variable practices, and poor levels of compliance with oncology CPGs are observed\(^5,6\). For instance, Wöckel et al. reported a 51.9% guideline adherence rate for the complete treatment received by primary breast cancer patients\(^6\).

In the last decade, organizational measures such as multidisciplinary tumor boards (MTBs) have been introduced to promote quality in care delivery to cancer patients\(^7\). The aim is to gather the various cancer specialists (surgeons, medical oncologists, radiologists, pathologists, radiotherapists, etc.) to promote the collective discussion of cancer patient

*The DESIREE project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 690238.
clinical cases and the decision of the best care plan for the patients. MTBs are expected to improve CPG implementation and to help capture cases for clinical trials. If studies on cancer care generally associate multidisciplinary tumor boards with improvements of guideline compliance rates\textsuperscript{8–10}, daily practice in MTBs is hampered by the complexity of the disease and the vast amount of patient and disease data available. Implementing CPGs is difficult since a high case load is actually discussed in MTBs and individual cases usually only receive a very limited amount of time for review. Thus, the impact of multidisciplinary tumor boards has been questioned\textsuperscript{11}, suggesting that decision support tools could improve MTB efficiency.

Indeed, CPGs are usually developed by health professional societies and national health agencies as textual documents in a narrative format. The sole dissemination of narrative guidelines poorly helps CPG implementation and has almost no impact on physician decisions\textsuperscript{12}. On the contrary, embedding CPGs in clinical decision support systems (CDSSs) has shown some benefit on the improvement of physician decision compliance with CPGs\textsuperscript{13,14}. Different prototypes of CDSSs supporting the management of breast cancer patients have been developed. However to the authors’ knowledge, only few systems have been routinely used to support decisions made for actual breast cancer patients in real life breast MTBs or breast units (BUs), and only some of them evidenced they were improving compliance of BU decisions with CPGs. Among them, we can mention MATE\textsuperscript{15} which uses the PROforma language, OncoCure\textsuperscript{16} based on Asbru-encoded protocols of pharmacological therapies for breast cancer, and OncoDoc\textsuperscript{17,18} that offers to navigate through a knowledge base structured as a decision tree to obtain patient-specific recommendations. Different formalisms have been proposed to translate the narrative recommendations into computer-interpretable guidelines\textsuperscript{19}. More recently, research works using web semantic approaches have been conducted, e.g. Abidi \textit{et al.} proposed to use ontologies in a rule-based reasoning process to manage breast cancer patients\textsuperscript{20}.

However, all previous propositions only rely on the modeling and implementation of a single CPG applied to a unique pathology, in this case breast cancer. Other research works have developed solutions to handle the concurrent application of different CPGs in order to manage patients with comorbidities that need different treatment regimens. Wilk \textit{et al.}\textsuperscript{21} have proposed a framework based on first order logic to represent CPGs and to mitigate possible adverse interactions (drug-drug or drug-disease) between the recommendations provided by the different CPGs. Galopin \textit{et al.}\textsuperscript{22} have implemented an ontological reasoning process to allow for the flexibility necessary to deal with patients suffering from both hypertension and type 2 diabetes. In these cases, CPGs are reconcile on the basis of competition.

DESIREE is a European-funded project\textsuperscript{†} which aims at developing a web-based software ecosystem for the personalized, collaborative, and multidisciplinary management of primary breast cancer by BUs. DESIREE would offer guideline-based, case-based, and experience-based\textsuperscript{23} decision support. The system is expected to be used by the clinical partners of the consortium on actual breast cancer patients and in real life BUs. In this perspective, we have translated the different contemporary breast cancer CPGs that clinical partners are implementing in their BUs in a computer-interpretable format. The aim is then to reconcile breast cancer CPGs on the basis of complementarity and to concurrently apply all of them. The idea is to avoid the CDSS silence observed when no recommendation is issued for a clinical case not covered by single CPGs, and to extend the clinical coverage of the CDSS by considering additional breast cancer CPGs.

In this paper, we present the guideline-based decision support system (GL-DSS) of DESIREE and describe how we used a domain ontology as the conceptual and terminological structure on which relies the reasoning process performed from the rules that model the selected breast cancer CPGs. We also propose an original visualization of the output of the concurrent application of CPGs as “rainbow boxes”\textsuperscript{24}.

**Material and methods**

**General design of the GL-DSS and basic workflow**

The aim of the GL-DSS of DESIREE is to provide state-of-the-art guideline-based recommendations structured as action plans issued from multiple breast cancer CPGs for any clinical case discussed by BUs. Informed BU participants may then choose to comply of not with one of the options provided. Since patient clinical cases discussed during BUs may be more or less complete, and because clinical data and reasoning process may be described at various levels

\textsuperscript{†}http://www.desiree-project.eu

1511
of abstraction, we chose to use semantic web tools allowing for subsumption. A data model and a knowledge model
dedicated to breast cancer management were consistently developed. The data model relies on a standard entity-
relationship model. The Breast Cancer Knowledge Model (BCKM) is a specification of all concepts that have been
identified for the management of breast cancer patients, structured as an OWL ontology. It enables the multi-level
description of patient cases and recommended action plans. The BCKM is a central static resource shared by all
data-oriented modules and components of DESIREE, so that they can be semantically interoperable.

To elaborate the knowledge base of the GL-DSS, the first step of the method was to select the different breast cancer
CPGs to be used. Once validated by the clinical partners of the DESIREE consortium, each of the considered CPGs
has been first structured as a set of human-readable decision rules, then encoded into a formal model-driven rule
language (NRL). Each formalized CPG was then processed in order to check its consistency regarding the BCKM and
to generate an internal computer-interpretable representation of the rules in the N3\textsuperscript{25} notation\textsuperscript{1}.

In the DESIREE project, a patient data repository, part of the DESIREE Information Management System (DESIMS),
constitutes a local breast-cancer-specific Electronic Medical Record (EMR) to store all patient cases according to the
common data model. In order to be interoperable beyond the DESIREE system, we decided to align both the knowl-
edge and the data models with standard reference terminologies to address semantic interoperability, and we chose the
FHIR messaging standard for communicating with external components to warranty syntactic interoperability\textsuperscript{26}.

Finally, we expected the GL-DSS to be used in real life BUs for actual patients. Thus, we needed to demonstrate
the good performance of the system and chose to use a fast and powerful rule based inference engine (Euler/EYE\textsuperscript{27}).
The GL-DSS module is implemented as a servlet providing web services through the use of a REST API. Figure 1
illuminates the different internal components of the GL-DSS, the external resources, and data flows. At run time, the
GL-DSS queries a FHIR server, linked to DESIMS, for patient data. Then, it converts the received bundle of FHIR
resources into an internal N3 representation. On the basis of N3 rules and of the N3 version of the BCKM, the
inference engine produces new facts, including the set of recommended action plans, which are returned to the user.

Clinical practice guideline collection and structuration

Various CPGs have been published for the management of breast cancer both at the international and national levels.
Our goal was to identify up-to-date CPGs that would be implemented within the GL-DSS module of DESIREE. The
selection criteria were: primary breast cancer; no restriction on the management step (including diagnosis, therapy,
and follow-up); published in English, or in the languages of the project’s clinical partners (Spanish and French);
most recent version when several versions existed. We conducted a search through different on-line resources: (i)
biomedical literature, using PubMed, (ii) international and national web repositories of CPGs including governmental
health care agencies (e.g. NGC, G-I-N, NCI...),\textsuperscript{5} (iii) websites of professional societies and associations, health care
institutions (e.g. NCCN, ESMO, ASCO, SEOM...), and (iv) custom local CPGs used as references by clinical partners’
BUs.

Eight CPGs were finally validated by the clinical partners of DESIREE and chosen for implementation within the
project: two NCCN guidelines Version 2.2016 “Breast Cancer” and “Genetic/Familial High-Risk Assessment: Breast
and Ovarian” [English]; the ESMO guidelines published in Annals of Oncology published in 2015 [English]; the
SEOM clinical guidelines for early-stage breast cancer published in 2015 [English]; three CPGs from the Paris public
hospitals (AP-HP), guidelines for “breast cancer management” (2016), for “fertility preservation” (2016), and for the
“management of cancer and pregnancy” (2014) [French]; the Onkologikoa’s breast cancer guidelines 2016 [Spanish].

Each CPGs has been structured by a medical oncologist specialized in the management of breast cancer as human-
readable IF-THEN decision rules. The rules built have been double-checked by a public health physician. Then, they

\textsuperscript{1}N3 is a compact and human readable syntax for RDF serialization but it has several features that go beyond a serialization for RDF models,
such as support for RDF-based rules. It currently has the status of a W3C submission whose specifications dated back to 2011 can be found at
http://www.w3.org/TeamSubmission/2011/SUBM-n3-20110328/

\textsuperscript{5}NGC = The National Guideline Clearinghouse (http://www.guideline.gov), G-I-N = Guideline International Network (http://www.g-i-n.net),
European Society for Medical Oncology (http://www.esmo.org), ASCO = American Society of Clinical Oncology (http://www.asco.org), SEOM =
Spanish Society of Medical Oncology (http://www.seom.org).
have been normalized to comply with the domain concepts defined by the breast cancer knowledge model.

*The Breast Cancer Knowledge Model*

The purpose of the knowledge model is to provide a common reference and terminology for all modules and components with the guaranty that clear, unique, and accredited concepts and definitions are being used. All the relevant concepts used for a patient case description and for the reasoning process are stored in a single place as an ontology in the OWL format. The choice to define a specific terminology focusing on the breast cancer domain was mainly made in order to be able to easily add new concepts. Nevertheless, the model remains tightly coupled with some existing established authoritative terminologies such as the NCI thesaurus, LOINC, and SNOMED CT, to further warranty semantic interoperability with potential hospital information systems (HISs) or EMRs. Alignment is maintained through links whenever it is possible by means of systematic annotations (see figure 2). All the relevant clinical procedures, possible examinations, clinical findings, and observations are present in the model.

The ontology also accounts for the breast cancer data model which includes the main concepts such as Patient, Side, and Lesion which are the main entities characterized by series of parameters or attributes, and allows for the description of a breast cancer patient case. Consequently, from the decision support point of view, the BCKM ontology is divided into two main hierarchies. The first one contains all the concepts specific to the DESIREE environment, namely the concepts used to define the entity parameters, and the second one contains all reference concepts defined in the breast cancer domain. The reference concepts are mainly used as potential values for the former DESIREE concepts.
Decision rules formats

We followed the approach promoted by the HL7 CDS group on the notion of “knowledge artifact” to represent rule sets. Rules are first encoded in a formal language, human readable, and possibly editable by some trained clinicians. We used NRL which is a model-driven language allowing to write logical expressions matching model components and rules, independently of any implemented inference engine. In our approach, the BCKM ontology acts as the reference model; an XSD model is automatically derived from the OWL model to feed the NRL parser that validates rule specifications. NRL has already been used to manipulate clinical information models and is similar to the GELLO language. NDL rule sets are then transformed into an XML representation, depending on the model, but independent from the source format (NRL) and the target computer-interpretable format. This intermediate representation is close to the HL7 Clinical Decision Support Knowledge Artifact Specification and is used to be shared by other modules (e.g., the experience-based DSS). Another transformation yields an N3 representation of the NRL rules which is the target computer-interpretable format.

Data representations

Patient-related data is stored outside the GL-DSS, within the DESIMS component. Input and output data exchange with the GL-DSS is performed using FHIR. No specific FHIR profile was defined and only standard FHIR resources were used, mainly Patient, Observation, Condition, BodySite, and Specimen. In order to preserve the data model through the FHIR transport, the main assumption has been that the observation resource was used to convey parameter-value pairs of the model. DESIREE codes defined in the DESIREE namespace of the knowledge model are used to particularize these observations and make them describe the same concepts as those used by the GL-DSS. Once parsed and decoded from the FHIR messages, patient data are converted to N3 notation and the data model is instantiated as a set of triples. This N3 representation is matched with the N3 executable rules by the inference engine at run time.

Execution engine

To combine rule-based reasoning and ontological reasoning, we adopted Euler, more specifically the EYE implementation by De Roo et al. This provides the system with description logic (DL) reasoning facility along with some classical logic powerful features. Euler is notably using Notation3 (N3), it is directly interoperable with the knowledge model since it belongs to the same family of semantic web tools able to deal with graphs of triples. Moreover, it appeared to be among the fastest reasoners with a full OWL-DL.

In the GL-DSS module, the engine is fed with a set of N3 rules and a set of N3 facts produced from the patient data.
according to the knowledge model concepts and properties. It takes the full knowledge model as additional input and using a generic query it then saturates the knowledge base until no more new triple can be produced generating new inferred facts. The content of the triple store is then explored in order to build new recommendations that have been produced with the basic concepts and building primitives from the knowledge model.

**Visualization of recommendations**

In order to display the multiple recommendations that may be produced by the different CPGs, we adapted the rainbow boxes visualization technique designed by Lamy et al. for the comparison of drug properties. It performs overlapping set visualization, a visualization problem that considers elements and the sets made out of these elements. The objective of the visualization is to clearly display which elements a given set includes, and which sets a given element belongs to, to facilitate the discovery of new knowledge such as similarities between elements or sets.

Rainbow boxes represents the elements in columns and each set is displayed as a rectangular labeled box that covers the columns associated with the elements belonging to the set. The boxes are stacked vertically as pieces in a Tetris game, and ordered by size (largest ones at the bottom). “Holes” can occur in a box when the elements belonging to a given set are not in consecutive columns. However, columns are ordered using a heuristic algorithm that minimizes the number of holes.

The presentation of multiple recommendations can be seen as an overlapping set visualization problem where each recommendation is a set of therapeutic units. Thus, in this context, a set is a recommendation and an element is a therapeutic unit. Therapies are grouped in six categories (pre-operative chemotherapy, pre-operative endocrine therapy, surgery, post-operative chemotherapy, radiotherapy, post-operative endocrine therapy) displayed from left to right, and each of the four types of therapies (surgery, chemotherapy, radiotherapy, endocrine therapy) were distinguished by color hues. In addition, some recommendations have a temporal dimension, e.g. chemotherapy with Adriamycin-Cyclophosphamide followed by Docetaxel. Color saturation is used to emphasize the first step of the recommendation.

**A case study**

As an illustration of how the GL-DSS operates, we present a case study processed with two of the breast cancer management CPGs that were selected in the DESIREE project, namely the 2016 version of NCCN CPGs “Breast cancer” and the 2016 version of AP-HP CPGs “Breast cancer management”. Both guidelines have been formalized as human readable IF-THEN rules by an oncologist specialized in breast cancer.

We consider the case of a patient, aged 67, diagnosed with an invasive bifocal breast cancer of the lower outer quadrant of the left breast. The first lesion is 35 mm, the second 12 mm, and the distance between the two is 18 mm. There is no clinical axillary lymph node (cN0). The two lesions have the same pathological profile: Estrogen receptors (ER)=95%, Progesterone receptors (PR)=40%, HER2-, SBR 3, and KI67=25%. The patient has no contra-indication to chemotherapy. Patient data is transmitted using a FHIR message to the GL-DSS and translated into the N3 format. Some preliminary inference rules are triggered to enrich the patient profile and assess the postmenopausal status, the TNM staging as T3 N0 M0, the stage as IIA, and the positive hormonal receptor status (HR). Figure 3 displays an excerpt of NCCN and AP-HP rules triggered on this patient case. NCCN guidelines recommend two management strategies, either surgery (lumpectomy and sentinel node excision or mastectomy and sentinel node excision) or systemic therapy: endocrine therapy with aromatase inhibitors or chemotherapy with the proposition of two different protocols (AC followed by Paclitaxel and TC). In the same way, AP-HP guidelines propose both surgery and systemic therapy (see figure 4). However, both surgeries and systemic therapies are different: AP-HP CPGs recommend a surgery by mastectomy with axillary dissection, there is no recommendation of endocrine therapy, and if some of the chemotherapies are those recommended by NCCN CPGs, there are two additional protocols.

The rainbow boxes visualization of the case study is displayed in figure 5. The box label indicates the guideline(s) leading to the recommendation. When both CPGs produce the same recommendation, there are two labels, and thus the box is larger. Consequently, it is easy to determine at a glance the most consensual recommendations. Similarly, the use of color makes it easy to see that most recommendations involve chemotherapy. It is also noticeable that all these chemotherapies include Cyclophosphamide.
We have developed the GL-DSS module of the DESIREE project applied to the management of breast cancer patients.

**Discussion**

We have developed the GL-DSS module of the DESIREE project applied to the management of breast cancer patients. The aim was to process the reconciliation of contemporary CPGs developed on a given pathology on the basis of complementarity to extend the coverage of patient profiles for which the GL-DSS provides recommendations. Although BU clinicians would have the opportunity to tick the CPGs they want to select to feed the GL-DSS, they may chose to work with all of them to reduce the silence of the system or see when different CPGs are consistent.

The choice of developing the knowledge model as an ontology presents two major advantages on top of providing a unique vocabulary to all the DESIREE project components: on the one side, the model resolves multilingual issues through label language tags, which is strongly required in an international project and, on the second side, the inference engine benefits from the ontological reasoning capabilities of the ontological model in addition to those provided by the rule engine. For instance, the use of the subsumption relationship leads in some cases to decrease the number of rules that have to be written and permits to match parameter values expressed at various levels of abstraction. In the same way, the issued recommendations may have different levels of specialization depending on the granularity of the information contained in CPGs for a particular patient case.

Different approaches have been developed to represent and execute CPGs\(^\text{19}\). However, in the case of breast cancer management, the choice of IF-THEN rules is appropriate to represent CPG contents since patient clinical data are easily formalized and recommended actions are given as care plans described at a high level of abstraction (to be refined and implemented by care providers outside the BU. In the same way, the temporal nature of CPG recommendations is easily represented by the ordered sequence of the therapeutic steps that compose the recommended care plan. We
could have chosen the HL7 Clinical Decision Support Knowledge Artifact Specification to define Condition-Action rules but we considered the standard was not stable enough to be used in an international project stressed on the production of outputs and outcome measures. In addition, a end-user formal language would still have been required, but we considered the standard was not stable enough to be used in an international project stressed on the open world assumption and the degraded performances obtained on a real scale with these techniques, we preferred to use the NRL language which accepts any declared information model like our BCKM, instead of using existing dedicated formalisms (like GELLO or Arden Syntax). IF-THEN rules have been manually built which sets the issue of the scalability of the approach. If natural language processing (NLP) methods to semi-automatically extract IF and THEN parts of rules have been proposed, the result is not yet satisfactory and solutions for the development of structured CPG contents should rely on the initial authoring of CPGs in these formats.

As for the execution engine, using tools from the semantic web domain as a base should have led to some kind of solution based on SWRL rules associated to a classical OWL reasoning engine to produce inferences. However, because of the lack of expressiveness of the latter, the impossibility to deal with non monotonicity and negation within the open world assumption and the degraded performances obtained on a real scale with these techniques, we preferred to adopt an alternative solution with Euler/Eye which does not have these limitations. When running the GL-DSS, the different CPGs are operated, which may lead to intra and inter-CPG conflicting recommendations. Only intra-CPG conflicts are resolved in the way proposed by. The resolution of inter-CPG conflicts are left to BU participants.

Currently, only two CPGs have been implemented, and work is still in progress with the other CPGs. It’s only when all CPG rule bases would be developed that we will be able to assess the performance of the semantic reasoner Euler/EYE and validate the choices we made. In addition, further evaluation is needed to assess the use of rainbow boxes to visualize the set of concurrent recommendations provided by the different CPGs.

Figure 4: Recommendations inferred by NCCN and AP-HP guidelines on the patient clinical case.
Figure 5: Rainbow boxes visualizing the eight recommendations obtained for the case study. Therapies are presented in columns and there is one recommendation per rectangular box (not one per line). One box includes a hole in the Docetaxel column. Colors identify the four categories of therapies (green, yellow, red, blue), the temporal dimension (the vivid part of a box corresponds to the first step, while the dimmed part corresponds to the other steps if any), and the guidelines (black and white labels).

Acknowledgments

The authors would like to thank all the partners of the project, especially the clinicians from Onkologikoa, Ander Urrutioechea Ribate, Jon Bellos Linacisoro, and Gerardo Cajaraville Ordoñana, as well as the participating members of Bilbomatica, among which Amaia Ugarriza Pinedo, Ander Elzo Zarraga, and Roberto Álvarez Sánchez.

References


Harnessing Biomedical Natural Language Processing Tools to Identify Medicinal Plant Knowledge from Historical Texts

Vivekanand Sharma, PhD¹, Wayne Law, PhD²,³, Michael J. Balick, PhD², Indra Neil Sarkar, PhD, MLIS¹

¹Center for Biomedical Informatics, Brown University, Providence, RI
²New York Botanical Garden, Bronx, NY
³College of Arts and Sciences, Lynn University, Boca Raton, FL

Abstract

The growing amount of data describing historical medicinal uses of plants from digitization efforts provides the opportunity to develop systematic approaches for identifying potential plant-based therapies. However, the task of cataloguing plant use information from natural language text is a challenging task for ethnobotanists. To date, there have been only limited adoption of informatics approaches used for supporting the identification of ethnobotanical information associated with medicinal uses. This study explored the feasibility of using biomedical terminologies and natural language processing approaches for extracting relevant plant-associated therapeutic use information from historical biodiversity literature collection available from the Biodiversity Heritage Library. The results from this preliminary study suggest that there is potential utility of informatics methods to identify medicinal plant knowledge from digitized resources as well as highlight opportunities for improvement.

Introduction

Evidence from archaeological studies demonstrate the existence of knowledge regarding plant use even before writing had evolved.¹ Over centuries, such knowledge has been transferred both verbally as well as in the form of written text or scriptures. One of the oldest mentions of medicinal plant use dates back to 2600 BC written on clay tablets in cuneiform describing approximately 1000 plants.² Although the term “ethnobotany” was coined in 1895,³ long before that time, botanists and physicians were traveling to diverse destinations to collect and document plant use knowledge. An example of one of the early documentations of such information is *De Materia Medica*, written by Pedanius Dioscorides, a Greek physician, pharmacologist, and botanist, between 50 to 70 C.E.,⁴ which remained the core of western pharmacopoeia as late as 1865.⁵ His accounts of around 600 plants were based on travels to the Mediterranean. Since then, a rich repository of data has been compiled from accounts of botanical explorations across the globe aimed at collecting specimens and documenting their cultural use. Botanical explorations reached their peak during the 19th century and, by the latter part, drew interest of academicians who recognized the importance of systematic examinations of indigenous societies and their knowledge of plant use.⁶ Studies of the historical use from such texts have aimed to show correspondence between textual and present day folk traditions as well as guide discovery of potential medicines.⁷

Due to the impact of modernization, there is a loss of traditional knowledge transferred over generations. A major goal of ethnobotany is to document and investigate the knowledge about plant use by indigenous communities. Ample evidence exists that reflect the success of natural products research (aimed at drug discovery) directed by knowledge gained from ethnobotanical explorations.⁶ Considering the diminishing corpus of knowledge of traditional culture, it is imperative to document and preserve elements of indigenous knowledge.⁷ As a result of one such contemporary effort, the Palau and Pohnpei Primary Health Care Manuals were written based on information gathered from The Republic of Palau and Pohnpei State, Federated States of Micronesia, both island groups in Micronesia. Such documentation of traditional botanical knowledge is important both in terms of conservation of biodiversity and for supporting public health.⁸ Cataloguing plant species information and their respective use in treating ailments is an essential task that remains challenging. The information present in historical texts can complement contemporary ethnobotanical studies.⁹,¹⁰ However, historical texts need to be in computer-analyzable form to facilitate the identification of relevant knowledge. One such supportive initiative is the effort being led by the Biodiversity Heritage Library (BHL).¹¹ The BHL houses digitized versions of biodiversity literature covering 115,124 titles. The digitization of text describing medicine provide opportunities to carry out analyses over different periods of time.¹² Automated analyses and integration of data from such sources with contemporary biomedical data may lead to the generation of testable hypothesis as well as provide preliminary support for bioprospecting studies.
Extracting relevant information from digital archives can help mobilize information that would otherwise remain sequestered. Use of controlled vocabularies, such as the Medical Subject Headings (MeSH) in biomedical domain, has been shown to aid efficient retrieval of relevant information. In the biodiversity domain, approaches have focused on highlighting and leveraging the species-centric nature of how knowledge is sought in the discipline. Use of “taxonomically intelligent” approaches has been shown to facilitate retrieval and linking of organism-specific information across resources. There have been significant advances in the biomedical domain with the development of automated methods aimed at gaining insight and hypothesis generation from narrative data. In addition to tools and algorithms for mining the text, due consideration has been put into the development of ontologies and terminological resources. Natural Language Processing (NLP) systems have been developed that determine concept equivalence of terms or phrases from unstructured text by mapping to standardized terminologies. NLP systems have also been developed for mining entity-relation extraction (e.g., BioMedLEE and SemRep). The main approaches in biomedical entity recognition involve lexicon-based, rule-based, and machine learning based methods, with combinations of these approaches resulting in better performance. Potential applications of such methods in the biodiversity domain has been described by Thessen et al. Availability of annotated corpora is important for training NLP tools to derive key patterns that can be used subsequently. NLP tasks have benefited from the availability of annotated corpora in biomedical domain. However, the lack of similar resources, including annotated corpora, make retrieval of knowledge from biodiversity texts challenging.

This study aimed to test the feasibility of using biomedical terminologies and NLP tools to extract therapeutic information from contemporary and historical ethnobotany texts. The ability of the NLP tool MetaMap to extract and normalize mentions of medicinal use to Unified Medical Language System (UMLS) Metathesaurus was evaluated, focusing on plants identified from the ethnobotanical surveys of the Micronesian islands of Palau and Pohnpei. The results from this study reveal the challenges and opportunities regarding the applicability of informatics based methods for supporting large-scale analysis of historical texts, such as those that are included in the BHL. The experience gained from this preliminary study motivates the need for developing efficient strategies for reliable extraction of medicinal use information from historical texts.

Methods

The primary goal of this study was to examine the feasibility of applying biomedical text mining tools and methods to facilitate the extraction of plant-associated therapeutic indications from historical texts. A reference standard based on the manual annotation of Primary Health Care Manuals (PHCMs) provided the basis for optimizing the approach and further evaluation. Following evaluation of the approach, historical text pertaining to the plant list from PHCMs was analyzed for possible therapeutic uses. A general overview of the steps of the developed pipeline is depicted in Figure 1.

Annotation of Primary Health Care Manuals. The Primary Health Care Manuals (PHCMs) document indigenous knowledge related to plants commonly used for medicinal purposes for selected medical conditions on Palau and Pohnpei. These documents were a result of extensive ethnobotanical surveys carried out as a part of the program, Biodiversity and Human Health in Micronesia. The health conditions included in these manuals were selected based on frequent diagnoses by traditional healers and from the data collected during the surveys. These manuals serve as a reference for those interested in the use of plants for healing by cataloguing traditional medicinal knowledge attained first-hand from the field explorations. The Palau and Pohnpei PHCMs each include medicinal use information for more than 80 plant species. The botanical uses described covers use of individual plant species as well as recipes involving combinations of herbs. The scope of this study focused only on descriptions of medicinal use involving individual herbs.

A reference standard was developed that consisted of therapeutic annotations associated with individual plants mentioned in the PHCMs. An annotation guideline was used to direct the annotation process of PHCMs, which were organized into individual text files consisting of one plant per document. Each individual plant document included: (1) Local Name(s), (2) Scientific Name, (3) Family, (4) Description of Herb, (5) Traditional Uses, (6) Pharmacological Properties, and (7) Toxicology. The scope for the annotation process was limited to the sections Traditional Uses and Pharmacological Properties. The annotation “therapeutic indication” was used to include all single or multi-word expressions that related to the medicinal use of a given plant of interest. In addition to specific medical conditions (such as disease or symptoms) medically-relevant terms (e.g., “anti-bacterial”, “analgesic”, and “stypic”) were also included. If a given plant species was described as active against certain organisms (e.g.,
microbes or parasitic worms) those were also annotated. The annotation was performed using Stanford Manual Annotation Tool\textsuperscript{26} in batch mode, which resulted in a tagged document that was parsed using a program written in Julia.

![Figure 1: Study Overview.](image)

Figure 1: Study Overview. Text from PHCMs (Palau and Pohnpei) were manually annotated to create a reference standard against which MetaMap output was evaluated. A custom filter was designed by optimizing the combination of semantic types to achieve highest F-score. This custom filter was used to distill the MetaMap output for plant related texts identified from BHL to identify potential treatments that were manually examined.

Extraction and Optimization of Biomedical Concept Identification. The individual descriptions associated with plants from the PHCMs (Palau and Pohnpei) were processed using MetaMap, with plant associated mentions of UMLS concepts and their respective semantic types being extracted by a Julia program. The processed output was used to generate and optimize a semantic type filter that to identify UMLS concepts of interest. A default filter for concepts based on those that belonged to the UMLS Semantic Group ‘Disorders’ was used\textsuperscript{27}. The MetaMap output was compared to the output from manual annotation, and the correctly identified UMLS Concepts were grouped by semantic type. A cutoff was defined that included semantic types that accounted for at least 5% of correctly identified concepts entities. All possible combinations of the semantic types that were within the cutoff criteria were analyzed relative to the reference standard, calculating Precision, Recall, and F-score. The combination of semantic types with highest F-score was selected as custom filter. The optimized custom filter obtained from Palau dataset was used for evaluation of MetaMap output from Pohnpei dataset and vice versa. Finally, the semantic types from the custom filter obtained from both dataset was merged and all combinations were tested on the combined Palau and Pohnpei dataset. The concepts from filtered MetaMap output was compared to manually annotated PHCM reference standard. The best performing (according to F-score) semantic type combination was used as the final custom filter for further use.
Table 1: Optimum Combination of Semantic Types. After applying a cutoff
Identifying Therapeutic Uses of
(see methods), all combinations of remaining semantic types were evaluated
Plants from Historical Biodiversity
for identifying optimal filter. A comparison of default filter (semantic group:
Texts.
Biodiversity
Heritage
disorder) with custom filter identified from the optimization step is provided.
Library
(BHL)
data
were
Semantic Type Filter
Dataset
Precision
Recall F-score
downloaded on Jan 27, 2017 that
acab, antb, bact, dsyn, fngs, inpo,
Palau
43.52
67.12
52.80
identified digitized titles, as well as
lbtr, patf, sosy, and virs
Pohnpei
33.97
67.06
45.02
metadata about the volumes and
(derived from Palau dataset*)
Combined
37.41
66.76
47.95
scientific names identified using
antb,
bact,
dsyn,
fngs,
inpo,
moft,
Palau
41.44
56.61
47.85
the Global Names Architecture’s
Pohnpei
35.18
64.24
45.39
sosy, and virs
Global Names Recognition and
Combined
37.37
60.77
46.28
(derived from Pohnpei dataset)
Discovery28. These data were
Semantic Group: Disorder
Palau
17.92
58.31
27.41
loaded into a local MySQL
(default semantic type filter)
Pohnpei
18.34
62.35
28.31
database that was used to query for
Combined
18.68
61.06
28.61
titles and identifier barcodes
*Also the optimum combination from both the dataset combined
associated with plant species. Title
and barcode data were retrieved for the combined Palau and Pohnpei plant list from the PHCMs. The titles were
filtered to include only those that contained the string “%medicine%” within their subject keywords. Using the
barcodes identified for titles related to PHCM plant list, the OCR text files in XML format were processed and
paragraphs that had mention(s) of plant species names of interest were subjected to processing using MetaMap. The
MetaMap output was then filtered using the custom semantic type filter previously described. The final output from
the filtered MetaMap output was manually examined. The evaluation was performed by going through the identified
scope of text and examining whether a plant specific treatment indication extracted using MetaMap was ‘True’ or
‘False’.
Table 2: Treatments Identified from BHL. The list of plants from

PHCMs were used to located relevant text in BHL. The identified
Results
scope of text was processed and output was manually examined to
identify potential treatments. The respective BHL identifier(s) for
Manual Annotation of PHCMs. From the
related text(s) are also listed.
Palau PHCM, a total of 65 plant species
Plant
Indication
BHL identifier
were included in this study, with manual
Barringtonia
racemosa
skin,diseases
illustrateddicti00gouluoft
annotation identifying 295 treatment
diarrhea
illustrateddicti00gouluoft
indications. Similarly, from Pohnpei
Carica papaya
round,worms
Malaypoisonscha00Giml
PHCM, 80 plant species were identified
heart,disease
b21443038;
and the annotation of descriptions
illustrateddicti00goul;
resulted in identification of 427 treatment
illustrateddicti00gouluoft
indications. In total, there were 129
abscess
cu31924001136872;
unique plant species from both the
veterinarymateri00wins;
veterinarymateri01wins;
PHCMs. 
Semantic Type filter and MetaMap
evaluation. A total of 2798 distinct CUIs
grouped into 109 semantic types were
identified by MetaMap processing of
Palau dataset. The true indications
accounted for 202 distinct CUIs grouped
into 43 semantic types found within the
reference standard. Using the cutoff of
5% or greater true indications within
each semantic type, 25 semantic types
were selected from Palau dataset. From
Pohnpei dataset, 3427 distinct CUIs
grouped into 112 semantic types were
identified. The true indications accounted
for 232 distinct CUIs grouped into 44
semantic types. Applying the cutoff
resulted in selection of 22 semantic types
for further analysis. Optimization of

indigestion

Curcuma longa
Musa paradisiaca
Mangifera indica

dropsy
ulcers
epizootic,disease
foot,and,mouth,disease
catarrh
skin,diseases

Terminalia catappa

1523

diarrhea
catarrh

veterinarymateri02wins;
veterinarymateri04wins;
veterinarymateri05wins;
veterinarymateri06wins;
veterinarymaterie8wins
veterinarymateri00wins;
veterinarymateri01wins;
veterinarymateri02wins;
veterinarymateri04wins;
veterinarymateri05wins;
veterinarymateri06wins;
b21297034
b21297034
illustrateddicti00gouluoft
illustrateddicti00gouluoft
b21443038;
illustrateddicti00goul
b21443038;
illustrateddicti00goul
illustrateddicti00gouluoft
illustrateddicti00gouluoft


semantic type combinations on MetaMap processed Palau dataset identified a combination of ten semantic types: acab, antb, bact, dsyn, fngs, inpo, lbtr, patf, sosy, and virs. This combination accounted for a maximum F-score of 52.8 (Precision: 43.52 and Recall: 67.12). This custom semantic type filter when applied on MetaMap processed Pohnpei dataset resulted in an F-score of 45.39. The optimum custom filter obtained from MetaMap processed Pohnpei dataset consisted of eight semantic types: antb, bact, dsyn, fngs, inpo, moft, sosy, and virs. This combination was selected as it resulted in a maximum F-score of 45.39 (Precision: 35.18 and Recall: 64.24). When applied to the MetaMap processed Palau dataset this custom filter resulted in a F-score of 47.85 (Precision: 37.41 and Recall: 66.76). This custom filter was same as the one derived from semantic type optimization on Palau dataset. A comparative evaluation using a default filter (semantic types included within the semantic group ‘Disorder’) resulted in F-score of: 27.41 on Palau dataset; 28.31 on Pohnpei dataset; and 28.61 on combined dataset. A summary of Precision, Recall, and F-score obtained from optimum custom filters and default filter is presented in Table 1.

**Treatments Identified from BHL.**
For the 129 distinct plant species identified, 72 BHL texts were identified after filtering based on subject keywords followed by scope of text containing mention of plant names (217 text segments). Manual evaluation of MetaMap output with custom filter resulted in identification of 14 distinct treatment indications associated with six plants from 14 different BHL titles. The resulting indications are listed in Table 2 along with their respective BHL barcodes. The true indications identified from BHL accounted for 17.54% of total associations. An additional 24 distinct treatment indications associated with nine plants were identified that belonged to semantic types ‘phsu’ and ‘fndg’ (Table 3). From the associations identified with these two semantic types the true indications described in BHL texts accounted for 11.70%. Table 4 lists the titles, links to BHL texts and DOIs for the barcodes mentioned in Tables 2 and 3. Additional treatments were identified from BHL which add to the list compiled from ethnobotanical survey of Palau and Pohnpei (as described in PHCM). A comparative list of such treatments are presented in Table 5.

**Discussion**
Starting from generational transfer of knowledge through verbal means to ancient documentation by explorers, it is important to preserve historical knowledge for guiding future discoveries. The field of pharmacognosy has benefitted by historical leads from such sources\(^\text{[10]}\). Since the need for systematic investigation of indigenous plant

<table>
<thead>
<tr>
<th>Plant</th>
<th>Indication</th>
<th>STY</th>
<th>BHL Identifier</th>
</tr>
</thead>
<tbody>
<tr>
<td>Allium cepa</td>
<td>stimulant</td>
<td>phsu</td>
<td>b21443038</td>
</tr>
<tr>
<td></td>
<td>expectorant</td>
<td>phsu</td>
<td>b21443038</td>
</tr>
<tr>
<td></td>
<td>diuretic</td>
<td>phsu</td>
<td>b21443038</td>
</tr>
<tr>
<td>Areca catechu</td>
<td>astringent</td>
<td>phsu</td>
<td>b21443038; illustrateddicti00gouluof</td>
</tr>
<tr>
<td></td>
<td>demulcent</td>
<td>phsu</td>
<td>b21443038; illustrateddicti00gouluof</td>
</tr>
<tr>
<td></td>
<td>anthelmintic</td>
<td>phsu</td>
<td>b21443038; illustrateddicti00gouluof</td>
</tr>
<tr>
<td></td>
<td>vermifuge</td>
<td>phsu</td>
<td>illustrateddicti00gouluof</td>
</tr>
<tr>
<td>Barringtonia racemosa</td>
<td>diarrhea</td>
<td>fndg</td>
<td>illustrateddicti00gouluof</td>
</tr>
<tr>
<td>Carica papaya</td>
<td>abortifacient</td>
<td>phsu</td>
<td>Malaypoisonscha00Giml</td>
</tr>
<tr>
<td></td>
<td>vermifuge</td>
<td>phsu</td>
<td>Malaypoisonscha00Giml</td>
</tr>
<tr>
<td></td>
<td>freckles</td>
<td>fndg</td>
<td>Malaypoisonscha00Giml</td>
</tr>
<tr>
<td></td>
<td>emetic</td>
<td>phsu</td>
<td>b21443038</td>
</tr>
<tr>
<td></td>
<td>anthelmintic</td>
<td>phsu</td>
<td>illustrateddicti00gouluof</td>
</tr>
<tr>
<td>Curcuma longa</td>
<td>intermittent,fever</td>
<td>fndg</td>
<td>illustrateddicti00gouluof</td>
</tr>
<tr>
<td></td>
<td>stimulant</td>
<td>phsu</td>
<td>illustrateddicti00gouluof</td>
</tr>
<tr>
<td>Luffa cylindrica</td>
<td>diuretic</td>
<td>phsu</td>
<td>illustrateddicti00gouluof; illustrateddicti00gouluof</td>
</tr>
<tr>
<td></td>
<td>purgative</td>
<td>phsu</td>
<td>illustrateddicti00gouluof; illustrateddicti00gouluof</td>
</tr>
<tr>
<td></td>
<td>emetic</td>
<td>phsu</td>
<td>illustrateddicti00gouluof; illustrateddicti00gouluof</td>
</tr>
<tr>
<td>Mangifera indica</td>
<td>astringent</td>
<td>phsu</td>
<td>b21443038; illustrateddicti00gouluof</td>
</tr>
<tr>
<td></td>
<td>tonic</td>
<td>phsu</td>
<td>b21443038; illustrateddicti00gouluof</td>
</tr>
<tr>
<td></td>
<td>purulent,discharges</td>
<td>fndg</td>
<td>b21443038; illustrateddicti00gouluof</td>
</tr>
<tr>
<td>Pangium edule</td>
<td>anthelmintic</td>
<td>phsu</td>
<td>Malaypoisonscha00Giml</td>
</tr>
<tr>
<td>Terminalia catappa</td>
<td>purgative</td>
<td>phsu</td>
<td>illustrateddicti00gouluof</td>
</tr>
<tr>
<td></td>
<td>diarrhea</td>
<td>fndg</td>
<td>illustrateddicti00gouluof</td>
</tr>
<tr>
<td>BHL identifier</td>
<td>Title ID</td>
<td>Title</td>
<td>Publication details</td>
</tr>
<tr>
<td>----------------</td>
<td>---------</td>
<td>-------</td>
<td>---------------------</td>
</tr>
<tr>
<td>illustrateddicti00goul</td>
<td>31340</td>
<td>An illustrated dictionary of medicine, biology and allied sciences ... by George M. Gould. 5th ed., with additions and corrections</td>
<td>Philadelphia P. Blakiston's Son 1907</td>
</tr>
<tr>
<td>illustrateddicti00gouluoft</td>
<td>31340</td>
<td>An illustrated dictionary of medicine, biology and allied sciences ... by George M. Gould. 5th ed., with additions and corrections</td>
<td>Philadelphia P. Blakiston's Son 1907</td>
</tr>
<tr>
<td>veterinarymateri00wins</td>
<td>42778</td>
<td>Veterinary materia medica and therapeutics, by Kenelm Winslow</td>
<td>New York, W.R. Jenkins, [c1908]</td>
</tr>
<tr>
<td>veterinarymateri05wins</td>
<td>49567</td>
<td>Veterinary materia medica and therapeutics</td>
<td>New York, W.R. Jenkins Co. [c1913]</td>
</tr>
<tr>
<td>veterinarymaterie8wins</td>
<td>62470</td>
<td>Veterinary materia medica and therapeutics</td>
<td>Chicago, American Veterinary Publishing Co. [c1919]</td>
</tr>
</tbody>
</table>
use knowledge has been recognized, ethnobotanists have been generating data from indigenous cultures around the globe. The process of ethnobotany includes field surveys, interviews, collection of specimens, documentation, and analysis to elucidate plant use pattern within the context of human civilization. Search for existing records and literature related to a given identified plant species is an essential part of the process. In light of the growing amount of data being accrued from surveys and digitization of historical texts, cataloguing and indexing of plant use knowledge remains an essential task to support the needs of ethnobotanists and medical historians. Having scalability in implementation of automated pipelines for such tasks may result in ease of analysis of plant use patterns, which in turn may benefit in distilling bioprospecting studies. This study sought to test the feasibility of using biomedical NLP tools and terminological resources for automated extraction of plant related therapeutic indications from ethnobotanical and historical texts.

Table 5: Comparison of Therapeutic Indications from PHCMs and BHL. A BHL versus PHCMs comparative view of the therapeutic indications extracted from texts for the list of plants from PHCMs that had mention in BHL text.

<table>
<thead>
<tr>
<th>Plant</th>
<th>Therapeutic Indication(s) from BHL</th>
<th>Therapeutic Indication(s) from PHCMs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Allium cepa</td>
<td>digestion; stimulant; expectorant;</td>
<td>worms; intestinal worms; GI issues; diarrhea; Trichurus</td>
</tr>
<tr>
<td></td>
<td>Diuretic</td>
<td>muris nematode</td>
</tr>
<tr>
<td>Areca catechu</td>
<td>astringent; demulcent; anthelmintic</td>
<td>shaking sickness</td>
</tr>
<tr>
<td></td>
<td>vermifuge</td>
<td></td>
</tr>
<tr>
<td>Barringtonia racemosa</td>
<td>skin,diseases; diarrhea</td>
<td>heartburn; skin rash; antinoceceptive; analgesic; pain;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>antibacterial; back pain</td>
</tr>
<tr>
<td>Carica papaya</td>
<td>abortifacient; vermifuge; round, worms; freckles; heart,disease; emetic; digestive; abscess; tumors; malignant,growth; anthelmintic; Indigestion</td>
<td>stings of stonefish and scorpionfish; burning rash from fire coral; antioxidant; anti-inflammatoryy; local inflammation; genotoxicity; antimicrobials; guinea worm infection; anthelmintic</td>
</tr>
<tr>
<td>Curcuma longa</td>
<td>dropsy; intermittent,fever; ulcers; stimulant</td>
<td>skin rash; black spots; melasma; stretch marks; anti-inflammatoryy; skin conditions; anti-aging; dry and damaged skin; darkened skin areas; sting of sea urchin; antibacterial; antiviral; antifungal; antispasmodic; hepatoprotective; inflammatory diseases; mangrove sickness</td>
</tr>
<tr>
<td>Luffa cylindrica</td>
<td>diuretic; purgative; emetic</td>
<td>skin burns; antioxidant; antifungal; Mycosphaerella arachidicola; Fusarium oxysporum; fever; milk production in lactating women</td>
</tr>
<tr>
<td>Mangifera indica</td>
<td>astringent; tonic; catarrh of nasal passage; purulent,discharges from the vagina; skin,diseases</td>
<td>syphilis; antiviral; anti-inflammatoryy; antinoceceptive; analgesic; dental hygiene; immune enhancing; hypoglycemic activity; back pain; joint pain; stomachache</td>
</tr>
<tr>
<td>Musa paradisiaca</td>
<td>epizootic,disease; foot,and,mouth,disease</td>
<td>bites; stings; boil; skin burn; burn wound; cuts; wounds; syphilis; pregnant women who has stomachache</td>
</tr>
<tr>
<td>Pangium edule</td>
<td>anthelmintic</td>
<td>hemorrhoids</td>
</tr>
<tr>
<td>Terminalia catappa</td>
<td>purgative; diarrhea; catarrh</td>
<td>diarrhea; antibacterial; antifungal; antioxidant; hepatoprotective; chemopreventive</td>
</tr>
</tbody>
</table>

The development of automated pipelines for extraction of plant related medicinal use information from ethnobotanical and historical texts is fraught with challenges. A major challenge is the lack of annotated reference corpora for training and evaluating NLP approaches. A contribution of this study is the manual annotation of PHCMs (Palau and Pohnpei) that can be used to support the development and evaluation of NLP tools for ethnobotany. The annotation of therapeutic indications from these manuals reveal range of descriptions for describing therapeutic use of medicinal plants. For the purpose of evaluation and optimization of the pipeline developed in this study, the concepts identified by MetaMap were organized and combined according to the UMLS Semantic Network. The UMLS Semantic Network reduces the more than one million concepts in the UMLS Metathesaurus into 133 semantic types that are further grouped into 15 semantic groups. After applying the cutoff value used in this study, 25 and 22 different semantic types, respectively from Palau and Pohnpei PHCMs, were identified over which the relevant concepts were distributed. Such spread over different semantic type categories reflects the diverse nature of terms or phrases used to describe medicinal uses of plants within these texts. Owing to the diverse spread, an optimization approach was used in this study to filter noise while still retaining relevant therapeutic information. The efficacy of such filter is evident from the significant improvement in F-score.
over the default filter that only included the semantic group ‘Disorder’ (see Table 1).

Several interesting medicinal applications were identified from the BHL texts analyzed in this study (Tables 2 and 3). The year of publication for the identified titles ranged from 1874 for the oldest to 1919 for the most recent (see Table 4 for list of identified BHL titles). A comparative view of treatments from PHCMs and associated titles identified from BHL is presented in Table 5. There were few medicinal uses that were common: *Allium cepa* is described in BHL title to stimulate digestion and PHCM mentions its use for gastrointestinal issues; *Barringtonia racemosa* in BHL title has use in skin diseases and PHCM lists its use for skin rashes; the anthelmintic potential of *Carica papaya* is also common between BHL title and PHCM. Additional therapeutic use information for related species were also visible in the defined scope of text such as the use of *Terminalia chebula* for diarrhea, dysentery and in bilious disorders.

The scope of text analyzed was defined as a paragraph containing the mention of plant name. Although interesting therapeutic indications related to the PHCMs plant list were identified, future work is needed to further investigation of individual texts for designing custom templates specifying relevant scope of text (e.g., document, page, section, paragraph, or sentence). Also, use of co-reference resolution methods and techniques for analysis may potentially be beneficial. Additionally, correction of optical character recognition (OCR) output text will be essential for contemporary NLP systems to map concepts correctly. The plant species names can be used as an identifier for integrating information, but there may be challenges involving ambiguity and use of alternative name(s) (‘synonyms’). Future work may be more inclusive by using a collection of synonyms and vernacular names associated with a given plant for identifying relevant texts. Similarly, there may be challenges in terms of correlating words or phrases used in historical text to describe different disorders as well as their respective synonyms and variants into canonical medical concepts. Future work might therefore aim to include semantic libraries for improving the recall of relevant biomedical concepts, including the use of methods for automated lexicon generation. Finally, approaches for developing entity name reconciliation services may be used for better aligning name strings (e.g., from iPhylo) with collaborative knowledge bases such as Wikidata (or the now deprecated, but still accessible, Freebase). Such services may allow for the building of custom knowledge graphs for obtaining answers to questions of interest.

This study highlighted several challenges and opportunities in text mining historical data available in resources like the BHL for identification of potential therapeutic leads. The results and insights gained in this study serve as a foundation upon which a bridge between historical plant-based knowledge (which include those described in biodiversity texts) and contemporary biomedical application can be built. In particular, the success of this feasibility study demonstrates the potential for automated cataloguing of medicinal plant specific information from ethnobotanical and historical data sources such BHL using existing biomedical NLP resources. It is anticipated that the approach demonstrated here can be used on other relevant sources of historical medicinal plant knowledge, such as the collection of health related historical material maintained by National Library of Medicine’s History of Medicine Division.

**Conclusion**

The process of cataloguing medicinal plant use information by ethnobotanists has largely relied on manual searching of records from surveys and biomedical literature. Historical evidence about medicinal use of plants from archival sources may be a valuable addition to the existing pipeline of cataloguing information. Automated methods can support this process of distilling large amounts of text descriptions to extract meaningful information. This preliminary study demonstrated the potential of informatics methods for extraction of plant-associated medicinal use information from historical biodiversity literature archived in the Biodiversity Heritage Library. In addition to identification of historical medicinal use leads, several challenges were highlighted that provides insights for how to adapt biomedical terminology and approaches.

**Acknowledgement**

This study was funded by grant R01LM011963 from the National Institutes of Health.
References


24. Gurulingappa H, Mateen-Rajput A, Toldo L. Extraction of potential adverse drug events from medical case


The Effect of Neighborhood Disadvantage on Diabetes Prevalence

Lincoln Sheets, MD, PhD¹, Gregory F. Petroski, PhD¹, Julie Jaddoo, MS¹, Yan Barnett, MA¹, Chris Barnett, MA¹, Laura E. Henderson Kelley, MD, MPH, FAAP¹, Vaishnavi Raman, MBBS¹, Amy J. H. Kind, MD, PhD²,³, Jerry C. Parker, PhD¹

¹University of Missouri, Columbia, Missouri; ²University of Wisconsin School of Medicine and Public Health, Madison, Wisconsin; ³VA Geriatrics Research Education and Clinical Center (GRECC), Madison, Wisconsin

Abstract

Patient socioeconomic data is not usually included in medical records nor easily accessible to clinicians, yet socioeconomic disadvantage can be an important guide to disease management. This study evaluated the neighborhood-level Area Deprivation Index (ADI), a measure of neighborhood socioeconomic disadvantage, as a factor in diabetes mellitus prevalence. Electronic health records at an academic hospital system identified 4,770 Medicare beneficiaries. Logistic regression of diabetes diagnosis (ICD9=250.x) against ADI quintile, age, gender, and race/ethnicity found all these patient characteristics to be significantly associated. Diabetes prevalence was lowest in the least disadvantaged quintile of neighborhoods after adjusting for age, gender, and race/ethnicity. The positive non-linear association of diabetes prevalence with ADI demonstrates the power of this index to practically quantify socioeconomic disadvantage. The ADI may be suitable for clinical decision support, and for informing the policy changes which are needed to reduce socioeconomic disparities in diabetes prevalence and other health outcomes.

Introduction

Epidemiological significance of diabetes mellitus

One twelfth of the world’s population has diabetes mellitus, including 10.8% of Americans¹. In the United States (US) alone, diagnosed diabetes costs $176 billion annually in direct medical costs and another $69 billion annually in decreased productivity². Diabetes incidence rates are rising, with 1.9 million US adults newly diagnosed with diabetes in 2010³.

Education, occupation and income are independently associated with a 25-30% decreased risk of type 2 diabetes⁴. Lower subjective social status (the individual’s perception of his or her position in the social hierarchy) is also associated with increased risk of diabetes⁵, as are depression⁶ and lower self-efficacy and social support⁷.

Prevalence of the disease varies by ethnic group⁸, but disparities are more strongly associated with socioeconomic status than with race or ethnicity⁹. The least well-off suffer a disproportionate share of the burden of diabetes¹⁰, and children who live in poverty are more likely to develop type II diabetes as adults and more likely to die from it earlier¹¹,¹². The association between socioeconomic status and diabetes mellitus is worldwide, in low-income countries as well as high-income countries¹³.

Limitations of existing approaches

Clinicians need additional training and tools to assess and respond to social determinants of health at the point of care¹⁴. However, because patient income and education are often not part of the medical record, nor often discussed directly during care delivery, clinicians must infer socioeconomic status from available data, such as patient neighborhood of residence. The Area Deprivation Index (ADI), is a geographic, area-based measure of the socioeconomic deprivation to which census-defined neighborhoods (i.e. census block groups) are exposed¹⁵. Addition of this socioeconomic risk factor to more traditional individual-level clinical data such as age, gender, and race/ethnicity could improve the understanding of diabetes risks in individual patients and patient populations.

Objective of the current study

The objective of this study was to develop and evaluate a clinically useful tool for more practically identifying the social determinants which impact diabetes prevalence in a patient population.
Methods

Study population and setting

The study population comprised 4,770 primary-care patients at the University of Missouri who were 65 years or older and enrolled in Medicare (US federal health insurance for the elderly and disabled) on September 1, 2014. All patient diagnoses, demographics, and other clinical attributes were retrieved from University of Missouri Health System medical records.

Data collection

This study used an updated and refined 2013 ADI developed by the University of Wisconsin School of Medicine and Public Health (AJK). This 2013 ADI uses data from the 2009-2013 American Community Survey to construct a factor-based ADI score from 17 census block-group level markers of socioeconomic status for each neighborhood of interest (Table 1).

Table 1. Components of the Area Deprivation Index (ADI)

<table>
<thead>
<tr>
<th>Area Socioeconomic Disadvantage Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percent of the population aged 25 and older with less than 9 years of education</td>
</tr>
<tr>
<td>Percent of the population aged 25 and older with at least a high school diploma</td>
</tr>
<tr>
<td>Percent employed persons aged 16 and older in white-collar occupations</td>
</tr>
<tr>
<td>Median family income in US dollars</td>
</tr>
<tr>
<td>Income disparity</td>
</tr>
<tr>
<td>Median home value in US dollars</td>
</tr>
<tr>
<td>Median gross rent in US dollars</td>
</tr>
<tr>
<td>Median monthly mortgage in US dollars</td>
</tr>
<tr>
<td>Percent of owner-occupied housing units</td>
</tr>
<tr>
<td>Percent of civilian labor force population aged 16 years and older who are unemployed</td>
</tr>
<tr>
<td>Percent of families below federal poverty level</td>
</tr>
<tr>
<td>Percent of the population below 150% of the federal poverty threshold</td>
</tr>
<tr>
<td>Percent of single-parent households with children less than 18 years of age</td>
</tr>
<tr>
<td>Percent of households without a motor vehicle</td>
</tr>
<tr>
<td>Percent of households without a telephone</td>
</tr>
<tr>
<td>Percent of occupied housing units without complete plumbing</td>
</tr>
<tr>
<td>Percent of households with more than 1 person per room</td>
</tr>
</tbody>
</table>

Because the ADI is an index rather than a concrete measure such as income or percent of households, it is not associated with a defined unit of measure. Differences between index values must be interpreted in relation to the range of index values in the data set. For these reasons, the possible ADI scores for Missouri were calculated for each
Census block group and grouped into quintiles, with both the ADI score and the ADI quintile assigned to each member of the study population. Figure 1 shows the resulting map of ADI quintiles for the state of Missouri.

Figure 1. Missouri Area Deprivation Index (ADI) quintile by Census Block Group.

Statistical methodology

Diabetes mellitus, the outcome of interest, was defined as the presence of any 250.x ICD-9 code in a patient’s EHR record. Because that outcome is a binary categorical value (either present or absent), logistic regression was used to test the association between ADI and diabetes, adjusting for patient age, gender, and race. Diagnostic plots of the empirical log-odds of diabetes against ADI quintiles indicated that relationship is not linear, and so ADI was treated as a nominal scale variable. Integer age in years was treated as a continuous variable. Gender was treated as a categorical variable, with “female” and “male” being the only two values recorded for this study population. Race/ethnicity was also treated as a categorical variable with two values, “White/Non-Hispanic”, and “Other”. The data source included eight possible values for race/ethnicity, but because the “White/Non-Hispanic” category included 88% of the study population, all other categories (“American Indian or Alaskan Native”, “Asian”, “Black or African American”, “Hispanic”, “Native Hawaiian or other Pacific Islander”, “Other race”, and “Unknown race”) were reclassified as “Other”.

Results

All the variables tested were found to be statistically significant with \( p < 0.05 \), as shown in Table 2.
Table 2. Analysis of effects on diabetes prevalence in the multivariate logistic regression model

<table>
<thead>
<tr>
<th>Effect</th>
<th>Odds Ratio</th>
<th>Pr &gt; Chi-Square</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>--</td>
<td>0.0191*</td>
</tr>
<tr>
<td>ADI quintile 1 (least disadvantaged)</td>
<td>1.37</td>
<td>&lt;.0001***</td>
</tr>
<tr>
<td>Age (integer years)</td>
<td>0.99</td>
<td>0.0003***</td>
</tr>
<tr>
<td>Gender (female)</td>
<td>0.81</td>
<td>0.0011**</td>
</tr>
<tr>
<td>Race/ethnicity (White/Non-Hispanic)</td>
<td>0.44</td>
<td>&lt;.0001***</td>
</tr>
</tbody>
</table>

* p<0.05, ** p<0.01, *** p<0.001

A graphic visualization of the percent of patients with diabetes mellitus (DM) within each ADI quintile (see Figure 2) clearly shows a positive but non-linear association of increasing DM prevalence (ranging from less than 25% to more than 35% in the study population) with increasing ADI (i.e., increasing neighborhood deprivation.).

Figure 2 shows visible differences in diabetes prevalence between patients from the least disadvantage neighborhoods (quintile 1), from the most disadvantaged neighborhoods (quintile 5), and from neighborhoods in the middle (quintiles 2-4). However, pairwise comparison of the five quintiles found statistical differences only between quintile 1 and all other quintiles, as shown in the Table 3.
Table 3. Analysis of effects in the logistic regression model

| ADI Quintile Comparison | Odds Ratio | Pr > |t| |
|-------------------------|------------|------|---|
| 1 v 2                   | 1.3198     | 0.0013** |
| 1 v 3                   | 1.3461     | 0.0012** |
| 1 v 4                   | 1.2928     | 0.0171*  |
| 1 v 5                   | 1.3886     | 0.0261*  |
| 2 v 3                   | 0.9804     | 0.8307   |
| 2 v 4                   | 1.0208     | 0.8487   |
| 2 v 5                   | 0.9505     | 0.7304   |
| 3 v 4                   | 1.0412     | 0.7176   |
| 3 v 5                   | 0.9694     | 0.8355   |
| 4 v 5                   | 0.9311     | 0.6532   |

* p<0.05, ** p<0.01

The low statistical significance of the apparent increase in diabetes prevalence for quintile 5 may be partly due to the relatively small sample size of this quintile, comprising only 5.7% of the study population. This limits power significantly.

Discussion

Fulfillment of study objective

The positive non-linear association of diabetes mellitus prevalence with neighborhood disadvantage, as measured by the Area Deprivation Index (ADI), indicates that ADI could be a useful tool for converting patient address into a significant predictor of diabetes prevalence in a patient population.

Additional findings

In this study population, the prevalence of diabetes in the least disadvantaged neighborhoods (ADI quintile 1) is significantly lower than all other neighborhood quintiles after adjusting for age, gender, and race/ethnicity. Quintiles 2 through 5 were not statistically different. The statistical equivalence of quintile 5 with quintiles 2 through 4 may be due to relatively small size of quintile 5, or may show an approximate equivalence between diabetes prevalence at moderate and high levels of neighborhood socioeconomic disadvantage.

Limitations of the current study

Because the study was limited to a single academic health center in mid-Missouri, and to Medicare enrollees over 65 years of age, these results may not be generalizable to other populations. Because nearly 90% of the study population was non-Hispanic White, the important characteristics of race and ethnicity may not adequately tested by this study. Furthermore, as previously stated, relatively small sample size in the most disadvantaged neighborhood quintile limits our statistical power and our ability to draw conclusion for this group.

Directions for future research

Repeating this study with a larger patient population, and with a population sample more representative of the entire US, may add to the understanding on how neighborhood disadvantage by ADI could be harnessed in diabetes care and population health efforts.
Conclusion

This study demonstrates the power of the Area Deprivation Index (ADI) to quantify neighborhood socioeconomic deprivation into a single index for use in predictive analytics. The ADI may be suitable for clinical decision support, as well as for informing the policy changes which are needed to reduce socioeconomic disparities in diabetes prevalence and other health outcomes. In addition to confirming the existing literature in finding an association between neighborhood socioeconomic deprivation and the prevalence of diabetes mellitus, this application of the ADI adds new findings to our understanding of health disparities. Most strikingly, the association of neighborhood socioeconomics with good health was strong in this population only for those residing within the highest levels of privileged neighborhoods; middle-class neighborhood status granted no measurable protection from the diabetes health disparities that affected residents of the poorest and least privileged neighborhoods. Clinicians and population health managers armed with a simple and unobtrusive measure of neighborhood socioeconomic status, based not on private financial data but simply on patient address, may be able to more effectively screen for, prevent, and manage diabetes in individual patients and managed populations. This is an important idea worthy of future research and study.

References

18. For additional information on the ADI or to access ADI data sets, please contact Amy Kind, MD, PhD at ajk@medicine.wisc.edu at the University of Wisconsin School of Medicine and Public Health.
Acknowledgements

This work was supported by the National Institute on Minority Health and Health Disparities of the National Institutes of Health under Award Number R01MD010243 (PI: Amy Kind).
Leveraging Collaborative Filtering to Accelerate Rare Disease Diagnosis

Feichen Shen, Ph.D., Sijia Liu, M.S., Yanshan Wang, Ph.D., Liwei Wang, M.D., Ph.D., Naveed Afzal, Ph.D., Hongfang Liu, Ph.D.
Department of Health Sciences Research, Mayo Clinic, Rochester, MN, USA

Abstract

In the USA, rare diseases are defined as those affecting fewer than 200,000 patients at any given time. Patients with rare diseases are frequently misdiagnosed or undiagnosed which may due to the lack of knowledge and experience of care providers. We hypothesize that patients’ phenotypic information available in electronic medical records (EMR) can be leveraged to accelerate disease diagnosis based on the intuition that providers need to document associated phenotypic information to support the diagnosis decision, especially for rare diseases. In this study, we proposed a collaborative filtering system enriched with natural language processing and semantic techniques to assist rare disease diagnosis based on phenotypic characterization. Specifically, we leveraged four similarity measurements with two neighborhood algorithms on 2010-2015 Mayo Clinic unstructured large patient cohort and evaluated different approaches. Preliminary results demonstrated that the use of collaborative filtering with phenotypic information is able to stratify patients with relatively similar rare diseases.

Introduction

In the USA, rare diseases are defined as those affecting fewer than 200,000 patients at any given time\(^1\). According to research conducted by Rare Disease Day\(^2\) and the National Organization for Rare Disorders (NORD)\(^3\), to date, there are over 7,000 known rare diseases inflicting 30 million Americans (nearly 1 in 10), more than half of whom are children\(^2\). However, due to the lack of scientific knowledge and clinical experience for rare diseases, most patients with rare diseases are frequently misdiagnosed or undiagnosed. In addition, only 5% of those diseases have corresponding treatment plans\(^2\). Therefore, there is an urgent need to accelerate the diagnosis of rare diseases as well as explore the science behind them.

Almost 80% of rare diseases are genetic\(^2\) and the first step towards rare disease research and diagnosis is to identify patients with similar phenotypes, where a phenotype refers to a clinical observable sign or symptom. Multiple studies have reported on studying rare disease phenotyping. To facilitate the discovery of genotype-phenotype association, the Human Phenotype Ontology (HPO)\(^4\) was developed to capture human phenotype information and is one of the most representative efforts that conducts phenotype-oriented analysis for rare disease differential diagnosis. The current version of HPO contains more than 116,000 annotations for over 7,000 rare diseases. To understand genotypes for rare diseases, Phen-Gen\(^5\) provides a method to combine phenotypes and patients’ sequencing data with domain knowledge in order to locate genotypes for rare disorders. Other efforts, PhenomeNET\(^6\) and PheWAS\(^7\) leveraged phenotypic information to discover genotype-phenotype associations that can be applied to prioritize genes for rare diseases.

Meanwhile, in the era of e-commerce, collaborative filtering techniques\(^8\) are popularly applied to recommend products to a customer based on customers with similar purchase preferences or other interests. The problem of diagnosing a patient with a disease based on patients’ phenotypic information is very similar to recommending a product to a customer, and therefore it is natural to propose the use of collaborating filtering for disease diagnosis. For example, given the hypothesis that disease prediction is achievable through analyzing phenotype and disease history, CARE provided an individualized healthcare framework by analyzing ICD-9-CM codes in patient’s medical history\(^9\-11\). Similarly, Steinhaeuser and Chawla proposed a combined disease network and collaborative filtering to perform disease prediction on structured patient data\(^12\).

In this study, we proposed a patient based collaborative filtering system enriched with natural language processing (NLP) and semantic techniques to assist rare disease diagnosis based on phenotypic information extracted from free-text clinical notes. Specifically, we extracted the Unified Medical Language System (UMLS)\(^13\) concepts using MetaMap and used terms in the Human Phenotype Ontology (HPO)\(^4\-14\) and the Genetic and Rare Diseases Information Center (GARD)\(^15\) based on dictionary lookup to preprocess 2010-2015 clinical notes from Mayo Clinic Rochester campus. We then used a patient based collaborative filtering framework for rare disease diagnosis and compared four similarity measurements and two neighborhood algorithms.
In the following, we first introduce materials used in this study. Next, we describe the methods used to build the framework and conduct the evaluation. We then present the results followed by discussion. Lastly, we conclude and discuss potential future directions.

**Materials**

*Clinical Data Collection*

We collected all clinical notes during the years of 2010 to 2015 generated at Mayo Clinic Rochester campus. The resulting corpus contains 12.8 million clinical notes corresponding to 729,000 patients. We limited our annotation to sections containing problems and diagnoses.

*The Unified Medical Language System and MetaMap*

The Unified Medical Language System (UMLS), developed at National Library of Medicine (NLM), is an integrated knowledge base that involves key medical terminologies and related resources. There are three components in the UMLS, the Metathesaurus, Semantic Network, and Specialist Lexicon. The Metathesaurus lists all clinical concepts integrated from over a hundred of terminological resources. A unique identifier, Concept Unique Identifier (CUI), is assigned to each medical concept. Each concept can be associated with one or more semantic types defined in Semantic Network. MetaMap is a configurable application for mapping biomedical text to the Metathesaurus. Here, we applied MetaMap with the UMLS version 2015 to extract UMLS concepts from clinical notes as a preprocessing step.

*Human Phenotype Ontology*

The Human Phenotype Ontology (HPO) has been developed as a controlled vocabulary for phenotypes by mining and integrating phenotype knowledge from medical literature and ontologies. HPO also provides associations with other biomedical resources such as Gene Ontology. HPO contains four sub-ontologies focusing on different annotation areas: Phenotypic Abnormality, Mode of Inheritance, Clinical Modifier, and Mortality/Aging. HPO and the UMLS have been cross-referred and we limited phenotype concepts extracted from clinical notes only to UMLS concepts that are cross-referred with the HPO released on September 2016. We only considered concepts from HPO sub-ontology Phenotypic Abnormality (HP_0000118) and its descendants, consisting of 11,721 phenotypes.

*Genetic and Rare Diseases Information Center*

The Genetic and Rare Diseases (GARD) Information Center is a program initiated by the National Center for Advancing Translational Sciences (NCATS) and funded by the National Institutes of Health (NIH). GARD extracts information from NIH resources, medical textbooks/databases, literature and the Internet to aggregate knowledge of rare or genetic diseases and group them into 32 categories. In the current version, GARD contains 4,560 diseases. We limited rare diseases extracted from clinical notes only to UMLS concepts that can be mapped to GARD.

**Methods**

Figure 1 shows the overview of our system workflow which includes two modules: i) a preprocessing module leveraging NLP and semantic processing techniques to identify patients with rare diseases and collect their phenotypes; and ii) a collaborative filtering model to recommend similar patients and possible disease recommendations.

*Preprocessing Module*

We first extracted problems and diagnoses from clinical notes using MetaMap. We kept UMLS concepts from the following semantic types: Disease or Syndrome (dsyn), Neoplastic Process (neop), Mental or Behavioral Dysfunction (mobd), Anatomical Abnormality (anab), Congenital Abnormality (cgab), Injury or Poisoning (inpo), Finding (fndg) and Sign or Symptom (sosy). To limit our analysis to only patients’ phenotypes and rare diseases, we conducted another refinement round on preprocessed UMLS terms to only keep phenotype-related concepts in HPO and rare diseases in GARD. For patients with rare diseases, we collected their phenotypic information within the last twelve months of the first mention of rare diseases.

*Patient based Collaborative Filtering Module*

In traditional user-based collaborative filtering, user preference data with various features describe different angles of user interests. Similarly, in the clinical domain, we considered patients as users and leverage large clinical cohorts to extract phenotypes as items for each patient. In contrast to rating-based recommender systems in which
preference score matters, we considered patient profile with phenotypes as binary inputs, that is, the relationship between patient and phenotype is either yes or no.

This module was designed on top of a user based collaborative filtering engine consisting of similarity measurements and neighborhood algorithms. Specifically, in this study, we applied Tanimoto coefficient similarity (TANI)\textsuperscript{18}, Overlap coefficient similarity (OL)\textsuperscript{19}, Fager & McGowan coefficient similarity (FMG)\textsuperscript{20}, and Log likelihood ratio similarity (LLRS)\textsuperscript{21} as four similarity measurements for binary data to calculate patients’ similarity based on phenotypes. Each measurement is feasible for different recommendation tasks, and as we did not have any \textit{a priori} reason to prefer one measurement over another, we tested all four methods.

Let $|\text{Phe}_i|$ and $|\text{Phe}_j|$ be the total number of phenotypes, and $|\text{Phe}_i \cap \text{Phe}_j|$ be the number of common phenotypes between any two patients $i$ and $j$. Tanimoto, Overlap and Fager & McGowan coefficient similarity are defined as shown in Equations 1, 2, and 3, respectively.

\[
\text{TANI}(i,j) = \frac{|\text{Phe}_i \cap \text{Phe}_j|}{|\text{Phe}_i| + |\text{Phe}_j| - |\text{Phe}_i \cap \text{Phe}_j|} \quad (Eq \ 1)
\]

\[
\text{OL}(i,j) = \frac{|\text{Phe}_i \cap \text{Phe}_j|}{\min (|\text{Phe}_i|, |\text{Phe}_j|)} \quad (Eq \ 2)
\]

\[
\text{FMG}(i,j) = \frac{|\text{Phe}_i \cap \text{Phe}_j|}{\sqrt{|\text{Phe}_i| + |\text{Phe}_j|}} - \frac{1}{2 \sqrt{|\text{Phe}_i| + |\text{Phe}_j|}} \quad (Eq \ 3)
\]

Log likelihood ratio similarity is built based on Shannon entropy\textsuperscript{22}. Given an event $X$ with its possible value $x_i$, Shannon entropy is defined in Equation 4.

\[
H(X) = - \sum_{i=1}^{|X|} P(X = x_i) \log P(X = x_i) \quad (Eq \ 4)
\]

In Shannon entropy, let $p_{ij}$, $p_\eta$, $p_\gamma$, $p_\Pi$ denote probabilities of common phenotypes shared by both patient $i$ and $j$, patient $j$ but not $i$, patient $i$ but not $j$, and neither patient $i$ nor $j$, respectively. The Log likelihood ratio (LLR) between any two patients $i$ and $j$ can be defined as shown in Equation 5.

\[
\text{LLR}(i,j) = 2 \times (H(p_{ij} + p_\eta + p_\gamma) + H(p_{ij} + p_\eta + p_\Pi) - H(p_{ij} + p_\eta + p_\gamma + p_\Pi)) \quad (Eq \ 5)
\]

As a result, Log likelihood ratio similarity (LLRS) between patient $i$ and $j$ is defined in Equation 6.

\[
\text{LLRS}(i,j) = 1 - \frac{1}{1 + \text{LLR}(i,j)} \quad (Eq \ 6)
\]
We also applied two commonly used neighborhood algorithms in collaborative filtering for given similarity metrics to recommend patients. One is K Nearest Neighbor (KNN), where the neighbors included are the k nearest neighbors. The other is Threshold Patient Neighbor (TPN), where a similarity threshold t is used to select neighbors.

Evaluation

This proposed system was implemented in Java with Eclipse Standard/SDK version Luna 4.4.0 running on 64 bit Linux CentOS 6.8 servers hosted by Mayo Clinic. We used the Apache Mahout framework to establish the environment for implementing the collaborative filtering framework. We evaluated eight different experimental groups formed as: 1) TANI with KNN; 2) LLRS with KNN; 3) OL with KNN; 4) FMG KNN; 5) TANI with TPN; 6) LLRS with TPN; 7) OL with TPN; 8) FMG with TPN.

To determine the best k and t, we first conducted a 10-fold cross validation for each experimental group. For each round, we randomly selected 90% data for training and the rest 10% for testing. To deal with binary preference value (i.e. either yes or no to a phenotype), for a specific patient, the Apache Mahout framework considered phenotype preference values as the summation of all his/her neighbors who also have the same phenotype. To determine the best k for KNN and t for TPN, we compared those estimated preference values with patients’ actual phenotype binary values (0 or 1). Specifically, for each predicted value \( p'_i \) and its corresponding actual value \( p_i \), we applied root-mean-square error (RMSE) on the total n results across all patients as shown in Equation 7.

\[
RMSE = \sqrt{\frac{\sum_{i=1}^{n}(p'_i - p_i)^2}{n}} \quad (Eq 7)
\]

We then computed second derivative for each RMSE function to select the optimal k and t at the biggest second derivative value. For function \( y=f(x) \), second derivative \( sd \) can be used to detect the concavity of a graph as shown in Equation 8.

\[
sd = \frac{d^2y}{dx^2} \quad (Eq 8)
\]

After the optimal k and t for each experiment was confirmed, the system returned a number of neighbors for each patient ranked in descending order by similarity scores.

To evaluate the performance of patient recommendation, we used information retrieval techniques to evaluate all ranked recommendations. We selected k recommended patients for KNN and all patients with similarity higher than t for TPN. We considered each patient as a query and their neighbors as a group of ranked recommendations. In addition, we used patients’ diseases as a gold standard to validate their similarity and considered a recommendation as an optimal one as long as the recommended patient affected similar rare disease(s). The confusion matrix in Table 1 generally depicts how to evaluate the system performance. According to Table 1, precision, recall and F measure can be computed as shown in Equations 9, 10 and 11. To analyze the performance and observe the trade-off between precision and recall, starting from top ranked patients, we divided them into ten portions and plotted precision-recall curves for each experiment. In addition, we calculated precision-recall area under curve (PRAUC) for each case. Moreover, we computed mean average precision (MAP) for recommendations of all patients. As shown in Equation 12, for each query \( q \), we computed average precision \( AveP \) over all relevant answers and calculated summation of \( AveP \) for all queries then divided by the number of total queries \(|Q|\) to compute the MAP.

To further evaluate performance for individual rare disease prediction, we assigned similarity scores of recommended patients to their corresponding rare diseases and accumulated all similarity scores for each disease. We considered the rare disease with the highest accumulated similarity score as the final diagnosis to evaluate performance for rare disease prediction. We used the optimal algorithm to conduct the evaluation. Precision, recall and F measure were computed for each rare disease as evaluation outputs.

Here we applied three level matching criteria to determine if two rare diseases are similar or not. The first one is string matching. That is, we compared two diseases directly by checking their exact names. Considering each physician might use different terms or concepts to make a diagnosis, the use of strict string matching would probably miss some semantically similar diseases. The Systematized Nomenclature of Medicine – Clinical Terms (SNOMED-CT) groups comprehensive medical terminologies with a semantic hierarchy in a standard manner. Therefore, as a second level matching, we mapped diseases to SNOMED-CT and considered two diseases to be
related if they contributed to a common ancestor node within 3 hierarchical generations\textsuperscript{31}. In addition, to have broader rare disease similarity checking in terms of categorization, we used the GARD dictionary and considered two diseases have a close relationship if they were listed in the same rare disease category. Therefore, for different matching criteria, definitions of relevant or not in confusion matrix are different.

Table 1. Confusion matrix for system performance evaluation.

<table>
<thead>
<tr>
<th></th>
<th>Recommended Rare Disease</th>
<th>Not Recommended Rare Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relevant Rare Disease</td>
<td>True Positive (TP)</td>
<td>False Negative (FN)</td>
</tr>
<tr>
<td>Not Relevant Rare Disease</td>
<td>False Positive (FP)</td>
<td>True Negative (TN)</td>
</tr>
</tbody>
</table>

\[
\text{Precision} = \frac{TP}{TP + FP} \quad (Eq \ 9)
\]

\[
\text{ Recall} = \frac{TP}{TP + FN} \quad (Eq \ 10)
\]

\[
F \text{ measure} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (Eq \ 11)
\]

\[
\text{MAP} = \frac{\sum_{q=1}^{Q} \text{AveP}(q)}{|Q|} \quad (Eq \ 12)
\]

Results

After pre-processing, there are about 31,000 patients with at least one rare disease diagnosis. After removing rare diseases affecting only one patient, the final data set includes about 29,000 patients with 437 rare diseases (29 out of 32 GARD categories) and 2,400 phenotypes. In addition, 24,000 patients were diagnosed with only 1 rare disease and 5,000 patients were diagnosed with 2 rare diseases.

Figure 2 shows statistics of rare diseases in our clinical notes with GARD categories. Unique count indicates the number of unique rare diseases in each GARD category, and category coverage shows the percentile of rare diseases out of the total number of rare disease in each GARD category. The category with the biggest unique count is Rare Cancer consisting of 155 diseases, for instance, lentigo maligna melanoma and papillary thyroid carcinoma. In terms of category coverage, Autoimmune Autoinflammatory Diseases is the highest one with 45.45% (5 of 11), which includes autoimmune hepatitis, cretin syndrome, addison’s disease and so on.

Figure 2. Statistics of rare disease in clinical notes with GARD categories.
Root-Mean-Square Error for Optimal Threshold Selection

For different similarity measurements with KNN, Figure 3(a) plots the relationship between RMSE and the increasing number of neighbors from 2 to 30. Since KNN selected the closest \( k \) nodes incrementally, more neighbors will be included with \( k \) increased even though they do not locate within an absolutely close distance, which should increase the error. We found that with \( k \) increased, RMSE for LLRS had a fastest increase rate after optimal \( k \), indicating that LLRS is too sensitive to the change of \( k \). RMSE didn’t change obviously for OL and FMG, which shows that these two similarities with KNN are not capable of differentiating neighbors and non-neighbors. RMSE for TANI slightly increased RMSE after its optimal \( k \), suggesting that TANI+KNN is able to group neighbors in a moderate way.

Similarly, Figure 3(b) depicts the relationship between RMSE and patient similarity threshold varying from 0.01 to 0.99 for different similarity measurements with TPN. Generally, RMSE decreased as we increased patient similarity. The reason is that criteria for neighbor selection became higher when we increased the user similarity, which resulted in more precise neighbor detection and thus decreased the error. We found that RMSE for LLRS stayed stable until similarity threshold \( t \) became relatively higher, which indicates that LLRS is less sensitive to similarity threshold. We also found that curves for TANI, OL, and FMG had the similar trends but OL started with the highest RMSE and FMG started with the lowest RMSE. TANI had the moderate RMSE at the beginning and reached its optimal threshold \( t \) before OL and after FMG.

As a result, the best \( k \) and \( t \) for each combination are given in Table 2.

![Figure 3. RMSE evaluation for KNN (a) and TPN (b) with four similarity measurements.](image)

**Table 2.** The optimal \( k \) and \( t \) for different experiments.

<table>
<thead>
<tr>
<th></th>
<th>TANI</th>
<th>OL</th>
<th>FMG</th>
<th>LLRS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Best ( k ) for KNN</td>
<td>8</td>
<td>11</td>
<td>10</td>
<td>4</td>
</tr>
<tr>
<td>Best ( t ) for TPN</td>
<td>0.21</td>
<td>0.34</td>
<td>0.17</td>
<td>0.84</td>
</tr>
</tbody>
</table>

Performance for Patient Recommendation

Figure 4 shows precision-recall curves for 8 experiments and PRAUC for each matching criterion. We found that GARD matching yielded the best performance overall, and SNOMED-CT semantic matching always performed better than string matching. There is no significant difference between TANI+KNN and TANI+TPN for each of the three matching criteria. Specifically, TANI+TPN contributed to the best PRAUC for string matching while TANI+KNN led to the best PRAUC for SNOMED-CT and GARD matching. LLRS was suboptimal and LLRS+KNN outperformed LLRS+TPN with all three matching criteria. In contrast, PRAUC indicated that FMG+TPN performed better than FMG+KNN. OL contributed to relatively lower PRAUC than other three similarity measurements, OL+KNN performed slightly better for SNOMED-CT and GARD matching while OL+TPN was more suitable for string matching.

Table 3 shows mean average precision for patients with all their recommendations. MAP scores showed consistent performance with what PRAUC evaluated. But the only difference is that MAP scores indicated that TANI with
KNN performed slightly better for string matching while TANI with TPN was slightly better for SNOMED-CT and GARD matching.

![Figure 4. Precision-Recall curves for 8 experiments with 3 matching criteria (number in bracket indicates PRAUC).](image)

**Table 3.** Mean average precision for 8 experiments with 3 matching criteria (highest value in bold).

<table>
<thead>
<tr>
<th></th>
<th>TANI</th>
<th>LLRS</th>
<th>OL</th>
<th>FMG</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>KNN</td>
<td>TPN</td>
<td>KNN</td>
<td>TPN</td>
</tr>
<tr>
<td>String</td>
<td>0.4229</td>
<td>0.4228</td>
<td>0.3736</td>
<td>0.3672</td>
</tr>
<tr>
<td>SNOMED</td>
<td>0.4727</td>
<td><strong>0.4729</strong></td>
<td>0.4297</td>
<td>0.4215</td>
</tr>
<tr>
<td>GARD</td>
<td>0.7525</td>
<td><strong>0.7530</strong></td>
<td>0.7116</td>
<td>0.7084</td>
</tr>
</tbody>
</table>

**Performance for Rare Disease Prediction**

Here we chose TANI with KNN as the optimal algorithm and applied it on 24,000 patients with only one rare disease. Scatter plots in Figure 5 describe prediction performance for different matching criteria. There are 417 rare diseases in total. For these diseases, string, SNOMED-CT and GARD matching found 49.4%, 56.6%, and 92.1% correct predicted rare diseases, and weighted micro-average F measure for them are 0.4, 0.44, and 0.71 respectively.

Table 4 shows the top five diseases with the best F measures for each matching criterion. To protect patients’ privacy, any rare disease with affected cases less than 10 were marked as <10. We found for some diseases, the unique features affecting very few patients may contribute to a high prediction performance. For example, there were less than 10 patients with *ichthyosis bullosa of siemens*, but the prediction F measure with string matching is 0.8. All these patients have symptom *ichthyosis*, some of them have *ichthyosis* and *hyperkeratosis*, and some others have *ichthyosis* and *congenital bulous ichthysiform erythroderma*. These phenotypes grouped as unique combinations that can help diagnose *ichthyosis bullosa of siemens*. Meanwhile, with 1,380 affected patients, *abdominal aortic aneurysm* had a high F measure with string matching. However, not all rare diseases with a relatively large number of affected patients yielded the same performance. For instance, *meningioma* had 1,476 affected patients, but prediction F measures for it is only 0.44.

For string matching, some rare diseases didn’t get any correct predictions, such as *cadasil* (<10 patients) and *myotonic dystrophy* (25 patients), the reason is that those diseases not only had a small group of affected patients but also didn’t show unique groups of symptoms and signs. Due to the semantic hierarchical processing, SNOMED-CT matching had a slightly better performance than string matching. It was able to predict some diseases that string matching considered as non-relevant, such as *acquired von willebrand syndrome* (<10 patients) and *acute disseminated encephalomyelitis* (<10 patients). What is more, GARD gave predictions based on category and all top 5 predictions are 100% predicted. This indicates that phenotypes can help to infer similar type of rare diseases. Although such prediction cannot identify the exact rare diseases, similar rare diseases within the same category can still give clues for physicians to make diagnoses.
Discussion

Our study utilized similarity measurements that do not take individual preference scores, therefore, other methods provided by the Mahout collaborative filtering engine that either consider preference values or preference rankings were not suitable for our preprocessed clinical notes (e.g., Euclidean distance similarity, Pearson correlation similarity, Uncentered cosine similarity, and Spearman correlation similarity). TANI is similar to LLRS, with the only difference that the latter gives more weight to dissimilar patients if they have common phenotypes and assigns less weight to similar patients even they share exactly the same phenotypes. This weighting difference slightly depressed LLRS’s performance relative to TANI. Similar to LLRS, FMG gives weight to similarity, but not as much as LLRS. In addition, OL gives too much weight to patients’ similarities even with few shared phenotypes, which lacks the ability to stratify patients well. What is more, FMG is the only one that is sensitive to the selection of KNN or TPN, and significantly outperformed TPN. The reason is that setting threshold as 0.17 is more suitable than selecting 10 neighbors according to the neighborhood density formed by FMG in our clinical notes. Therefore, making a good balance between KNN and TPN has a potential ability to optimize the overall performance with idealized neighbors and similarity at the same time.

Figure 5. Scatter plot of precision-recall for rare disease prediction (circle size is proportional to the number of affected patients).

Table 4. Recommendation performance for selected rare diseases.

<table>
<thead>
<tr>
<th>Approaches</th>
<th>Top Diseases</th>
<th>Number of Affections</th>
<th>Precision</th>
<th>Recall</th>
<th>F Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tani+KNN with String Matching</td>
<td>osteochondritis dissecans</td>
<td>158</td>
<td>0.94</td>
<td>0.9</td>
<td>0.92</td>
</tr>
<tr>
<td></td>
<td>frontotemporal dementia</td>
<td>221</td>
<td>0.81</td>
<td>0.91</td>
<td>0.85</td>
</tr>
<tr>
<td></td>
<td>spasmodic dysphonia</td>
<td>173</td>
<td>0.8</td>
<td>0.9</td>
<td>0.84</td>
</tr>
<tr>
<td></td>
<td>abdominal aortic aneurysm</td>
<td>1,380</td>
<td>0.71</td>
<td>0.92</td>
<td>0.8</td>
</tr>
<tr>
<td></td>
<td>ichthyosis bullosa of siemens</td>
<td>&lt;10</td>
<td>0.8</td>
<td>0.8</td>
<td>0.8</td>
</tr>
<tr>
<td>Tani+KNN with SNOMED Matching</td>
<td>acute myelomonocytic leukemia</td>
<td>11</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>osteochondritis dissecans</td>
<td>158</td>
<td>0.94</td>
<td>0.9</td>
<td>0.92</td>
</tr>
<tr>
<td></td>
<td>frontotemporal dementia</td>
<td>221</td>
<td>0.81</td>
<td>0.91</td>
<td>0.85</td>
</tr>
<tr>
<td></td>
<td>spasmodic dysphonia</td>
<td>173</td>
<td>0.8</td>
<td>0.89</td>
<td>0.84</td>
</tr>
<tr>
<td></td>
<td>abdominal aortic aneurysm</td>
<td>1,380</td>
<td>0.71</td>
<td>0.92</td>
<td>0.8</td>
</tr>
<tr>
<td>Tani+KNN with GARD Matching</td>
<td>acute myelomonocytic leukemia</td>
<td>11</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>angioimmunoblastic t-cell lymphoma</td>
<td>18</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>ataxia telangiectasia</td>
<td>&lt;10</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>chronic myelomonocytic leukemia</td>
<td>&lt;10</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>congenital heart block</td>
<td>&lt;10</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
For overall performance across all recommendations, considering physicians have different descriptions of specific
diagnosis, string matching may overlook some associations, so we enriched disease matching with semantic
similarity. There exist some ontologies that describe diseases, such as Disease Ontology (DO)\textsuperscript{32} and Orphanet Rare Disease Ontology (ORDO)\textsuperscript{33}. However, for the detected rare diseases from clinical notes, only 56% and 52% are
covered by DO and ORDO, respectively. Therefore, we used SNOMED-CT, a more comprehensive ontology that
comprises 60% of our extracted rare diseases and maintains a more complicated hierarchical semantic relationship
for discovering disease associations. Nevertheless, due to the limited coverage, semantic disease checking didn’t
significantly improve the performance. To address this, some associations among rare diseases can be mined from
literature for evaluation.

For rare disease prediction with string matching, some phenotypes cannot uniquely characterize a certain type of
rare disease in our clinical notes. For example, hypopituitarism has 51 affected patients and prediction F measure is
0. Top frequent phenotypes for hypopituitarism are hypothyroidism, neoplasm, hypertension, apnea and
hyperlipidemia, which are very common symptoms shared with other rare diseases and are not useful to make
correct diagnose. Therefore, common comorbidities may create noises for decision making. In this preliminary
study, we only focused on patients with rare diseases and filtered out phenotypes that didn’t happen within 12
months of their rare disease diagnosis encounter time. To better characterize rare diseases and comorbidities, in the
future, we will pass all patients’ data to our system and target on giving recommendations for misdiagnosed and
undiagnosed cases. In addition, it would be interesting to investigate cross-institutional rare diseases to acquire
diagnosis experiences and intelligence from different hospitals and healthcare systems to build a more generic rare
disease diagnosis system.

Conclusion and Future Work

In this study, we have investigated patient based collaborative filtering with NLP and semantic techniques on large
patient cohort to assist rare disease diagnosis. We demonstrated its potential in facilitating rare disease prediction.

In the future, we plan to incorporate HPO phenotypic information content\textsuperscript{14}, topic modeling\textsuperscript{34}, word embedding\textsuperscript{35}
and deep learning temporal sequence analysis\textsuperscript{36} to give more degrees of similarity measurement for improving
prediction performance. In addition to phenotype based analysis, we plan to involve rare disease genotype from
PheWAS\textsuperscript{7} and literature to assist rare disease diagnose as well.

Acknowledgements

The authors gratefully acknowledge the support from the National Institute of Health (NIH) grants R01GM102282,
R01LM011934, R01EB19403, and OT3TR002019.

References

1. The orphan drug act implementation and impact. Available: https://oig.hhs.gov/oei/reports/oei-09-00-
4. Robinson PN, Köhler S, Bauer S, Seelow D, Horn D, Mundlos S. The Human Phenotype Ontology: a tool
17;83(5):610-5.
methods. 2014 Sep 1;11(9):935-7.
DM, Crawford DC. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene–disease
Microsoft Corporation One Microsoft Way Redmond, WA. 1998 May;98052.
9. Davis DA, Chawla NV, Blumm N, Christakis N, Barabási AL. Predicting individual disease risk based on
Oct 26 (pp. 769-778). ACM.
10. Davis DA, Chawla NV, Christakis NA, Barabási AL. Time to CARE: a collaborative engine for practical
20. Fager EW, McGowan JA. Zooplankton Species Groups in the North Pacific: Co-occurrences of species can be used to derive groups whose members react similarly to water-mass types. Science. 1963 May 3;140(3566):453-60.
35. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems 2013 (pp. 3111-3119).
Utilizing Smartphone-Based Machine Learning in Medical Monitor Data Collection: Seven Segment Digit Recognition

Varun N. Shenoy¹, Oliver O. Aalami, M.D.²
¹Cupertino High School, Cupertino, California, ²Stanford University, Palo Alto, California

Abstract

Biometric measurements captured from medical devices, such as blood pressure gauges, glucose monitors, and weighing scales, are essential to tracking a patient's health. Trends in these measurements can accurately track diabetes, cardiovascular issues, and assist medication management for patients. Currently, patients record their results and date of measurement in a physical notebook. It may be weeks before a doctor sees a patient's records and can assess the health of the patient. With a predicted 6.8 billion smartphones in the world by 2022¹, health monitoring platforms, such as Apple's HealthKit², can be leveraged to provide the right care at the right time. This research presents a mobile application that enables users to capture medical monitor data and send it to their doctor swiftly. A key contribution of this paper is a robust engine that can recognize digits from medical monitors with an accuracy of 98.2%.

Introduction

The goal of this research is to develop and evaluate the accuracy of a smartphone based system to automatically recognize and record biometric monitor results to Apple's HealthKit (Apple Inc., Cupertino, Ca). As the internet becomes more embedded into medical monitors through Wi-Fi and Bluetooth technologies, people who cannot afford to upgrade to these expensive machines will fail to receive the benefits of rapid medical attention. Our solution relies on a common smartphone to do the data capture, processing and communication. We use computer vision based feature extraction alongside machine learning to decipher a reading of a medical monitor captured by a smartphone. The medical monitors generally use seven-segment displays (where each digit is made up of a combination of segments). The model proposed in this paper can accurately read the monitor 98.2% of the time.

Prior Work

Whilst seven-segment digit recognition technology has been created in the past, prior work has not accounted for conditions affected by variable lighting and positioning. One proposed solution was ssocr³, an acronym for Seven Segment Optical Character Recognition. Such an approach was unsuitable to actively adapt to dynamic orientations and lighting of a smartphone image. Tesseract⁴, Google's popular optical character recognition model, was also another choice. However, we learned that while Tesseract is strong at reading regular text on a page, it had a difficult time accurately reading seven segment displays. Prior work in the field of medical informatics focussed on the recognition of hand written characters⁵.

Implementation

We delineate implementation into three phases. The first part will cover the use model of the app, the second part will cover the front end implementation and user interface. Finally, we will discuss the backend digit recognition approach.

Application Use Model

The use model is described as follows. Consider the case of measuring a biomarker such as blood glucose. After the glucose monitor displays a reading, the patient opens the smartphone application and selects the “Glucose” option. Next, the patient scales a bounding rectangle to fit the digits on the display. These bounds are saved for the next time a measurement of glucose will be taken. After taking a picture, the image within the bounds is sent to a backend server using wireless communications via the internet. A trained machine learning Random Forest⁶ classifier with image processing based feature extraction divides the image into individual digits and classifies each of them. These
numbers are sent back to the user’s device. The user is presented with the server’s prediction for the digits in the image. The user can easily correct the predicted numbers by using a scroll picker for each of the digits. This data is then saved into Apple HealthKit and is readily available for the physician.

**Front-end Implementation and User Interface**

The front end implementation of the application is focused on an easy and intuitive use model. The target audience is mainly the elderly who lack the dexterity required to accurately use mobile keyboards. As such, it is imperative that the font sizes throughout the mobile application are large and bold to aid the user’s eyes in reading. We also implemented a voice assistant to guide the user through the steps and in reading the result. Ultimately, the goal was to make the final user flow (Figure 1) of the app succinct and user-friendly. Figure 1 shows the steps involved in measuring systolic blood pressure through our application.

![Figure 1. The User Interface Flow for the Mobile Application](image)

On the front end side of the application, the Swift 3.0.x programming language was used for developing the iOS app and interfacing with HealthKit in the Xcode 8.3.x integrated development environment. Alamofire, a networking framework, was used to simplify the process of sending the image from the phone to the server via Hyper Text Transfer Protocol (HTTP) in a multipart form. Since our model was not embedded into the device, and lives on a server, computational requirements for the client are minimal. Our application will work as long as the device can make HTTP calls and access the internet. Apple’s AVSpeechSynthesizer class was used in the app to provide an active voice assistant in the app. The voice tells the patient exactly how to take a picture and reads out the resulting digits when it is saved to HealthKit. All components of the app were optimized to work on any Apple iPhone running iOS 8 and above.

**Backend Digit Recognition Approach**

A basic block diagram (Figure 2) shows the backend methodology for the server side digit recognition algorithm. The model is first trained on a training set of images and then tested on a new set of images it has never seen before.

We chose to implement our algorithmic model in Python 2.7.x, since there exist a diverse set of libraries for machine learning and computer vision. We used the scikit-learn package to implement a Random Forest classifier and model persistence. OpenCV for Python was used for feature extraction and image normalization due to its versatility and high computation with a simple programming interface. The Matplotlib graphics library was used to create data visualizations for this research.
Figure 2. Image Processing and Machine Learning Methodology.

A shared aspect of the training and testing steps is the ability to separate an image into disparate digit components. This is called feature extraction. The input image is divided into individual digits so the classifier only has to work with ten possible classes (0-9). Image processing is used to reduce the complexity of input data and normalize the images for the machine learning model. OpenCV’s built-in Otsu’s thresholding algorithm is used to reduce the given image to just black and white pixels. Otsu’s method dynamically separates the foreground and background of the input image and maximizes the variance between the two. After thresholding, we segment the image into individual digits using a vertical projection algorithm. A graph for the average pixel color in each column of the image is created (Figure 3). Peaks on the graph where the average pixel value is white (255) should be disregarded and gives us a location for digit segmentation. Each column is scanned and a splice is created wherever the average color is close to white. All sections of pure white are deleted and each individual digit is separated to be classified in the machine learning model. All of the digits are scaled to 15x45 pixels to minimize variance between images with the same digit in them.

Figure 3. Feature Extraction Process on an Example Image.

A Random Forest is a machine learning classifier that is composed of an ensemble of decision tree estimators. Each decision tree is a tree-like model of decisions and their possible consequences, ultimately leading to a final conclusion (leaf node). While decision trees on their own are weak classifiers, Random Forests are fast, accurate, and not prone to overfitting. In this research, 100 trees were used in the forest for rapid training and testing. Trees in the forest are grown using bootstrap aggregating, or bagging. Each tree is constructed on a random subset of the input data, so they produce different models that can be averaged. This ensures that the model cannot be overfitted by the training data. When a new, unlabeled, piece of data is shown to the random forest, each tree assesses it individually. The forest makes a decision through majority voting by the trees. The Random Forest classifier machine learning model was chosen over others, since it is very fast and accurate. The fully trained Random Forest ensemble for this research only takes up 847 kB of space and takes under 3 seconds to train on a 2.9GHz processor for laptops. Calls to the server take 2-3 seconds to process and report data back to the client device.
Experimental Results

The past records of a patient’s heart pressure monitor and a weighing scale were used to cultivate training and testing data (Figure 4). We sampled the historical data stored within these devices to recover 108 images for the testing set. These images had 1-3 seven segment digits, which our algorithm had to segment and classify. The software was trained on a small additional subset of representative seven segment displays (25 images). Screenshots of digits typed using a seven-segment font were used as training data as well. Images in the training set were not used in the testing set.

![Example Images Used to Train and Test the Model.](image)

Furthermore, we observed a non-uniform distribution of digits in the collected samples (Figure 5). E.g. there were 85 images of “1”'s, significantly more than the rest. This was expected as systolic blood pressure values in healthy adults have a leading “1”. The same can be said for weight measurements as leading “1”’s and “2”’s are the most common.

![The Distribution of Digits throughout the Entire Dataset.](image)

The machine learning model successfully recognized all of the digits except in two cases, achieving an accuracy of 98.2%. We examined three dominant features using Principal Component Analysis (PCA) on the testing and training set. This helps us observe the strength of our algorithm by visualizing clustering patterns determined by the 3-component PCA (Figure 6). In these three-dimensional plots, digits of the same class are observed to be well clustered. However, certain digits are clumped closer together than others. For example, the digits “0” and “8” are clustered together. This means that visually, digits that are “8”’s and “0”’s are similar, and therefore, we require more data so the model can learn concrete differences between the two. Also, since we reduced every vector to three dimensions from 675, certain information that can be crucial to differentiating “8”’s and “0”’s may have been lost.

![Distribution of Digits in Dataset](image)
Figure 6. Different Orientations of PCA-Reduced Data in Three-Dimensions.
We analyzed the output of our model on our testing data and we found that it performed well across all 10 classes (Figure 7). Across the 262 individual digits in the testing set (post image segmentation), we achieved an overall accuracy of 98.2% and an F1-score of 0.978 with our Random Forest model. An F1-score is a metric for classification which is best at 1 and worst at 0. It is the harmonic mean of precision and recall.

![Confusion Matrix of Our Model on All 10 Classes.](image)

Figure 7. Confusion Matrix of Our Model on All 10 Classes.

Other machine learning approaches from the scikit-learn library were used as well to verify the strength of our proposed model (Table 1). Ultimately, the Random Forest model achieved the highest accuracy of 98.2% while the General Purpose Multi-layer Perceptron was only accurate 61.1% of the time. Furthermore, a single decision tree (a voting member of a random forest model trained on the entire training set) attained an accuracy of 64.8%. Random Forests did the best since it is very difficult to overfit on the testing data. Other models are more prone to overfitting on the training data and as such, fail to recognize testing data correctly.

Table 1. Other Tested Machine Learning Models and Their Respective Accuracies from scikit-learn

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random Forest</td>
<td>98.2%</td>
</tr>
<tr>
<td>Linear Support Vector Machine</td>
<td>94.4%</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>64.8%</td>
</tr>
<tr>
<td>Naive Bayes</td>
<td>86.1%</td>
</tr>
<tr>
<td>K-Nearest Neighbors (neighbors = 5)</td>
<td>92.6%</td>
</tr>
<tr>
<td>General Purpose Multi-layer Perceptron</td>
<td>61.1%</td>
</tr>
</tbody>
</table>
Conclusion

In this research, we presented the design and implementation of a smartphone application that can capture medical data and communicate it to physicians. Using our tool, physicians can now catch early trends and diagnose critical issues before the onset of the medical vulnerability. We described a versatile seven segment digit recognition algorithm that can be used to rapidly read any digital display via computer vision and machine learning. Special attention was given to the user interface to serve the ill and elderly.

Since the submission of the paper for review, we have extended our backend infrastructure by uploading it to a cloud server (Amazon Web Services Lambda).

Further Work

There are many ways to further improve our algorithm. For one, sharp images guarantee better results. We would like to consider blur detection prior to analyzing our image. If the image is too blurry, we can send a message back to the user, requesting a clearer picture. There are many well known techniques to accurately measure blur within an image.

On a larger scale, it is necessary to gather more images for both training and testing. Creating a robust corpus of images will enable us to improve the accuracy of our method. We plan to test the software with real patients to validate the user interface in a medical environment.

We would also like to look into embedding our model into mobile devices directly without the need for a server. This will drastically increase speed for users and enable them to use the app in locations without access to the internet. State-of-the-art research is currently focussed on mobile embedded machine learning16.

References

Leveraging Clinical Time-Series Data for Prediction: A Cautionary Tale

Eli Sherman¹, Hitinder Gurm, MD², Ulysses Balis, MD³, Scott Owens, MD³, Jenna Wiens¹
¹University of Michigan Computer Science and Engineering, Ann Arbor, MI; ²Michigan Medicine Cardiology, Ann Arbor, MI; ³Michigan Medicine Pathology, Ann Arbor, MI;

Abstract In healthcare, patient risk stratification models are often learned using time-series data extracted from electronic health records. When extracting data for a clinical prediction task, several formulations exist, depending on how one chooses the time of prediction and the prediction horizon. In this paper, we show how the formulation can greatly impact both model performance and clinical utility. Leveraging a publicly available ICU dataset, we consider two clinical prediction tasks: in-hospital mortality, and hypokalemia. Through these case studies, we demonstrate the necessity of evaluating models using an outcome-independent reference point, since choosing the time of prediction relative to the event can result in unrealistic performance. Further, an outcome-independent scheme outperforms an outcome-dependent scheme on both tasks (In-Hospital Mortality AUROC .882 vs. .831; Serum Potassium: AUROC .829 vs. .740) when evaluated on test sets that mimic real-world use.

1 Introduction

The widespread adoption of electronic health records (EHR) for collecting and analyzing patient data presents a promising avenue for improving patient care¹. These retrospective data have been used in traditional statistical analyses to identify relationships between patient data and patient outcomes. For example, extensive work has been done to identify risk factors for sepsis², hospital readmission³, and heart failure-induced mortality⁴, among other conditions. In such analyses, the primary goal is to increase our understanding of the relationship between covariates and the outcome of interest. In contrast, the increasing accessibility of machine learning (ML) approaches has triggered a shift in focus to the development (and deployment) of predictive models, where the primary goal is good predictive performance⁵, ⁶.

This shift in focus necessitates a similar shift in methodology. For example, traditional analyses are often limited to a small number of variables, based on expert-driven clinical hypotheses or observations. Because the risk of over-fitting is low, researchers in this setting rarely consider a held-out test set. In a high-dimensional setting a held-out test set is imperative. More clinical researchers are recognizing this need and adapting their analysis appropriately, but this alone is not enough. The precise problem statement also requires an important, but often overlooked, reformulation. When the goal is prediction, the choice of method by which one chooses to extract one’s data is critical, for both training and testing purposes.

In particular, one must be careful to evaluate candidate models in a way that accurately estimates how they will perform when applied in clinical practice. This stands in contrast to previous analyses, in which the models may not have been intended for predictions. In a predictive setting, when a model will be applied, i.e., when a prediction will be made, is important. Two common methods for defining the time of prediction include: 1) indexing relative to a clinical event of interest e.g., onset of infection, and 2) indexing relative to an outcome-independent fiducial marker e.g., time of admission. The first approach is most common in a traditional retrospective analysis. In such settings, examples are typically derived by considering the event of interest and looking backwards to extract data collected prior to the event⁷,⁸. The second approach must be used a predictive analysis setting. Here, examples are derived based on a fiducial marker that precedes the clinical event of interest but is temporally independent of that event e.g., a particular operation/procedure.

In this paper, through two case studies, we illustrate some of the subtleties that surround prediction tasks and demonstrate how using an evaluation method that does not reflect clinical practice can result in misleading results. Leveraging a large publicly available dataset of ICU admissions, we apply several indexing techniques to two prediction tasks. In one prediction task, we aim to predict in-hospital mortality. In the other prediction task, we aim to predict hypokalemia, low serum potassium. We consider these two tasks since they are both important from a clinical perspective, but also because they present a contrast in terms of the number of events per admission. I.e., for the first task we may observe several test results, but in the second task there is only a single endpoint - in-hospital mortality. Through these
experiments we illustrate the importance of carefully defining one’s problem setup and demonstrate how evaluation performance can vary across settings.

In the following sections, we will: 1) discuss previous work in the development of predictive tools including examples of several indexing techniques, 2) formalize different indexing techniques/problem setups and outline how we will apply them to our case studies, 3) present results on the two real-world prediction tasks, and finally 4) reflect on the implications our experiments have on choosing an indexing method for extracting training and evaluation sets for a prediction model.

2 Related Work

Traditional low-dimensional statistical analyses are common in the clinical literature. Such studies typically focus on modeling relationships between covariates and outcomes\(^7\)–\(^\text{11}\). While the focus of these statistical analyses is often on testing hypotheses about relationships among variables, such approaches have been applied to prediction tasks. The popular mortality prediction scoring systems APACHE III\(^\text{12}\) and SAPS II\(^\text{13}\) serve as illustrative examples of this approach. In both models, a relatively small number of candidate covariates are hand-selected by experts and a mapping between covariates and outcomes is learned using standard statistical methods (e.g., logistic regression). This approach is limited in that data are collected at a specific point in the admission (24 hours after the time of admission, and at the time of ICU transfer for APACHE and SAPS respectively) and thus the models are only designed to make a single prediction.

In recent years, the focus has shifted to developing prediction models with high-dimensional feature spaces using machine learning techniques. These models are applied throughout the admission, providing updated predictions. As such, the question of how to index and extract time-series data from the EHR is now critically important. Numerous examples of indexing from the time of admission (or a related fiducial point) exist in both the healthcare and machine learning literature, including applications to patient status estimation\(^\text{14}\), diagnosis\(^\text{15}\), and sepsis\(^\text{16}\).

Despite the number of examples in which data are indexed based on an outcome-independent fiducial marker, the approach of indexing based on the event of interest still creeps into analyses today (perhaps since it was so common in a traditional, non-prediction centric, setting). Oftentimes, identifying the use of the event-based method is subtle and it is necessary to carefully consider the features included in the proposed model to confirm its use. For example, in work on learning a model for Clostridium difficile infection (CDI) risk prediction, it is not explicitly stated how examples are extracted\(^\text{17}\). However, in previous CDI work event-based indexing is clearly used\(^\text{18}\). Another paper developed a risk stratification tool for predicting sepsis risk in ICU patients in which the model used length of stay as a feature\(^\text{19}\). These types of features serve as a clear indication of backward-looking example selection. In another example, a paper proposing a heart failure prediction system generated covariates by looking backwards, testing different prediction intervals, from the diagnosis\(^\text{20}\). Collectively, these examples present an issue of data leakage, allowing extra information into the training process through problematic indexing schemes.

Others have published work on the potential pitfalls of working with EHR data\(^\text{6}\), and developed tutorial-style overviews of the process of developing and validating a clinical prediction tool\(^\text{21}\). These papers provide a fairly comprehensive discussion of how to undertake careful data analysis to create useful prediction tools. In contrast to previous work, we focus entirely on how training and test data are indexed when developing clinical prediction models.

3 Methods

In this section, we present and contrast common approaches for extracting and indexing EHR data. We begin by describing the general framework for data extraction, highlighting key choices one must make when extracting data. Then, we present the two case studies in which we aim to predict laboratory test results (specifically hypokalemia, i.e., low serum potassium) and in-hospital mortality.

3.1 Problem Setup and Notation

We limit our discussion to clinical prediction tasks during a hospital admission. For an admission, we have a set of \(d\) irregularly sampled, timestamped features. Additionally, we have a set of irregularly sampled, timestamped outcomes
\((y_j, t_j)_{j=1}^k\) where \(y_j\) is recorded at time point \(t_j\). Here, we restrict \(y \in \{0, 1\}\). Note that it may be possible for an outcome to occur multiple times, only a single time (e.g., death) or never at all within a given admission (i.e., \(k = 0\)). In the following sections, we will describe procedures for mapping these features and outcomes for an admission \(i\) into feature vector-outcome pairs of the form \((x, y)\) where feature vector \(x \in \mathbb{R}^d\) represents the covariates used to predict the outcome variable \(y\). Since our focus is prediction, \(x\) represents data recorded prior to observing outcome \(y\).

### 3.2 Indexing Longitudinal Data for Prediction

Considering the raw data for each admission as a collection of clinical data, we now outline careful considerations for extracting \((x, y)\) pairs. Figure 1 serves as a guide for the following descriptions.

- **Time of Prediction** The time of prediction, \(t_p\), is the time point corresponding to when a predictive model is applied to the data. The time of prediction affects covariate extraction and outcome/label extraction (more on this below), and is determined relative to some reference point \(t_0\). Given a model that predicts a particular outcome, one may choose to apply it at a single time point (as in the first time series in Figure 1), resulting in a single prediction per admission. If such a setup is desired then one test example should be extracted for each admission based on an outcome-independent fiducial marker (e.g., time of admission). Alternatively, one may choose to apply the model multiple times throughout the admission, updating predictions as new data become available (as in the second and third time series’ in Figure 1). To mimic this setup, examples can be extracted on a rolling basis, again starting from an outcome-independent reference point \(t_0\). More formally, starting from \(t_0\), one may extract either a single example \(x_p\) representing the admission from the time of admission to the time of prediction \(t_p\), or multiple examples \((x_1, t_1), (x_2, t_2), \ldots, (x_m, t_m)\) (at perhaps regularly spaced intervals). In the latter, each feature vector is updated based on data available at the time of prediction. The corresponding outcome variable \(y_p\) depends on the prediction horizon and the availability of ground truth, discussed next.

- **Prediction Horizon** The prediction horizon \(h\) (or window) is defined as the period for which a prediction applies and aids in defining \(y_p\). It begins at the time of prediction \(t_p\). In some settings, the prediction horizon is fixed and remains constant across examples, (e.g., when predicting 30-day mortality, or the value of a laboratory test in the next 12 hours; see the third time series in Figure 1). In many cases, the event of interest occurs at most once during the prediction horizon (e.g., 24-hr mortality), but is observed continuously (i.e., at every time point we know whether or not the patient is alive). In such settings ground truth is readily available and the corresponding \(y_p\) can simply be set to the observed value. In other settings, however, ground truth is only available at specific time points (e.g., when a laboratory test was ordered and returned). Here, it may be necessary to impute the corresponding outcome value or label. In the case study described below, we use a “copy-and-hold” approach: if there is no ground truth at the target time, we use the most recently observed \(y\) value that precedes the target time. Another way around this issue is to simply adjust the prediction horizon to be near zero and the prediction time to be driven the laboratory test order. In many settings, the time of the laboratory test order is still an outcome independent reference point, but depending on the task this may have limited clinical utility. Finally, in some settings, the prediction horizon may vary across admissions. E.g., many researchers aim to predict in-hospital mortality. Here, the prediction horizon varies with the length of stay, but \(y_p\) can be extracted similarly as if it were fixed.

Once the time of prediction and prediction horizon are defined, one can extract a feature vector \(x_p\) based on data available at the time of prediction, and \(y_p\) based on observations made during the prediction horizon. This results in either a single \((x, y)\) pair per admission in the test set, or \(m_i\) pairs for each admission \(i\).

To ensure evaluation accurately reflects how a model will perform in a practice, test data must be extracted in a way that mimics the clinical use case. In particular, when extracting test data it is necessary to use an outcome-independent reference point since outcomes are not available at test time. However, when training, we may want to make use of this additional information. In particular, one may choose to define the time of prediction based on the time of the event (e.g., work backwards from time of death) when extracting examples. When making multiple predictions for each admission, one extracts test examples regularly throughout each admission. But, applied to the training set, this procedure could introduce bias, since patients with longer stays will appear more often. To mitigate this issue, one
3.3 Dataset & Prediction Tasks

To measure the effects of the different data extraction approaches on the predictive performance of a model, we considered two prediction tasks applied to the same ICU dataset. In particular, we leverage the MIMIC-III Database\textsuperscript{22}. This dataset consists of 58,976 ICU admissions collected at a Beth Israel-Deaconess Medical Center in Boston, MA over the course of 12 years. The median length of stay is 6.9 days. We consider variables related to patient demographics, medications, laboratory tests, vital signs, and fluid inputs and outputs that are present in at least 5% of admissions. We encode each patient’s demographics as binary features. For the other variables, we generate summary statistics (e.g., min. value, max. value, mean value based on the several hours preceding prediction time) to encode the features for a given example. Using these data, we consider two clinical prediction tasks:

Predicting In-hospital Mortality In-hospital mortality prediction is a well studied problem in the healthcare-related literature. It is frequently used to benchmark new methods. For this task, we include all adult hospital admissions, resulting in 49,909 admissions. Here, we aim to predict whether or not the patient will expire over the course of the remainder of the admission (i.e., a variable prediction horizon).

Predicting Hypokalemia Accurate predictions of whether or not a particular laboratory test will return a hypokalemic result could help clinicians make better informed decisions regarding whether or not to order that test. This could in turn lead to a reduction in the number of unnecessary laboratory test orders. Here, we aim to predict hypokalemia. We focus on serum potassium (as opposed to other laboratory tests), since it is one of the most high volume tests in U.S. healthcare\textsuperscript{23,24}. From the MIMIC database, we include all adult patients with at least one serum potassium test. This results in a final study population of 49,354 admissions.

Using these data, we aim to predict whether or not the serum potassium laboratory test will return a hypokalemic
result, defined as serum potassium < 3.5, within the next 12 hours (i.e., our prediction horizon is 12 hrs.). This threshold was derived from a standard reference range. The prediction horizon was chosen based on the fact that many laboratory tests are ordered on a standard 24-hour schedule and discretionary testing occurs outside of this cycle. A 12-hour prediction window provides information that would ordinarily correspond to a discretionary test and therefore has the potential for eliminating the need for an extra discretionary test. Later, we consider another setup in which we generate a prediction of the current hypokalemia risk every time a potassium test is ordered. We describe this setup in more detail in Section 4.3.

3.4 Experimental Setup

Across all experiments, we build our training and test sets by repeatedly (100 times) randomly assigning 25% of admissions to the test set, and the remaining 75% to the training set. We are careful to not allow any admission to be represented in both the training set and test set since this would allow our learning algorithm to “memorize” the admission in question and subsequently perform better on the examples it had already seen. We use 5-fold cross validation on the training set to select hyper-parameters and retrain on the entire training set using the optimized parameters.

We extract examples according to the particulars of each experiment. Each example is a feature vector with 1, 222 features that encode summary information about the patient’s status over the 12 hours preceding prediction time. For instance, the patient’s glucose measurements over the 12 hours preceding the prediction time are encoded into variables describing the minimum, mean, and maximum values over the 12 hours. This summarization step allows us to account for the fact that different patients might have a different number of tests over the course of those 12 hours. We pick 12 hours so as to focus on the most recent clinical data while capturing temporal trends.

Using the training data, we train a classification model. Since the focus is on the problem setup, rather than the overall classification performance, we use a simple linear approach, specifically the liblinear implementation of Linear SVMs. For each experiment, we report mean and standard deviation statistics for area under the receiver operating curve (AUROC) scores across all 100 train-test splits. For models in which we make multiple predictions, each with a variable horizon as in the in-hospital mortality prediction task, we include each patient only once in the AUROC calculation. For each patient, we have a series of predictions, each in the form of a real valued number, the output of our classifier’s decision function. We sweep the decision threshold and if any of a patient’s predictions exceed that threshold we classify that patient as positive, and negative otherwise. This procedure is necessary since we are making predictions with a variable prediction horizon.

4 Experiments & Results

In this section, we describe and present results for several experiments using various problem setups. We analyze how different reference points and different numbers of predictions per admission impact model performance.

4.1 Event-dependent Training and Testing Yields Misleadingly Good Performance

We first illustrate the importance of selecting an appropriate outcome-independent reference point when extracting test data. For each task, we consider two different reference points: 1) time of admission and 2) the event of interest.

We first consider a single prediction setup for each task. For the in-hospital task, we use two different reference points to extract data. We 1) use the time of admission as our reference point, making a prediction 12 hours after admission, and 2) use the time of death (or time of discharge in the case of negative examples) as our reference point, making a prediction 24 hours prior to the end of the admission. When extracting examples for the laboratory testing task we 1) use time of admission as our reference, making predictions every 12 hours and 2) use the time of each potassium test as our reference, with a time of prediction 12 hours prior to each serum potassium test.

For these and all experiments in this paper, when we use the time of admission as our reference point, our first prediction is made 12 hours after the time of admission.

In order to isolate the impact that using an admission-based reference point has on this problem (rather than the number...
of predictions), we also consider a multiple prediction formulation. We create multiple-prediction training sets for the in-hospital mortality task using both event-independent and event-dependent indexing. For the event-independent training set we make predictions regularly – every 24 hours – throughout the admission. For the event-dependent training set, we make predictions in 24 hour increments, looking backwards from the end of the admission all the way to the beginning of the admission. The AUROC is calculated as described previously, where each patient contributes equally.

<table>
<thead>
<tr>
<th>Task</th>
<th>Reference Point</th>
<th>N Examples</th>
<th>N Pos. Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>In-Hosp. Mort.</td>
<td>Admission</td>
<td>49,909</td>
<td>5,244</td>
</tr>
<tr>
<td>In-Hosp. Mort.</td>
<td>Event</td>
<td>49,909</td>
<td>5,244</td>
</tr>
<tr>
<td>Hypokalemia</td>
<td>Admission</td>
<td>2,831,268</td>
<td>256,891</td>
</tr>
<tr>
<td>Hypokalemia</td>
<td>Event</td>
<td>644,371</td>
<td>69,234</td>
</tr>
</tbody>
</table>

Table 1: Number of extracted examples and the number of positive examples for the single mortality and hypokalemia prediction tasks. Note that the two in-hospital mortality sets have the same number of examples as they represent making a single prediction per admission, 12 hours after admission and 24 hours before death/discharge respectively.

<table>
<thead>
<tr>
<th>Task</th>
<th>Reference Point</th>
<th>N Examples</th>
<th>N Pos. Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>In-Hosp. Mort.</td>
<td>Admission</td>
<td>504,147</td>
<td>58,984</td>
</tr>
<tr>
<td>In-Hosp. Mort.</td>
<td>Event</td>
<td>476,170</td>
<td>54,338</td>
</tr>
</tbody>
</table>

Table 2: Number of examples extracted and the number of positive examples for each multiple prediction mortality set.

The resulting number of examples for each task and reference point are given in Tables 1 (single prediction tasks) and 2 (multiple prediction tasks). Admission-based reference points typically result in more examples compared to event-based reference points. The effects of these differences are explored later in this section.

The results for these experiments are presented in Figure 2. The three sections of this figure give results for the single mortality (a), potassium (b), and multiple mortality (c) tasks respectively. The red bars correspond to admission-based training sets applied to admission-based test sets; the blue bars correspond to event-based training sets applied to event-based test sets; and the green bars correspond to event-based training sets applied to admission-based test sets.
The first thing to note is the extremely good performance apparent in (a) and (b) when both training and test data are extracted in an outcome-dependent manner (AUROC .963 and .897 for in-hospital mortality and serum potassium respectively). However, when such a model is applied to outcome-independent test data (green), performance is significantly worse. Using time of event as a reference point for evaluation is not reflective of a realistic clinical use case. This conclusion is critically important and motivates our use of outcome-independent evaluation sets for the experiments that follow.

For the multiple prediction model (c), the event-dependent model performs relatively poorly when applied to both the event-dependent test set and to the event-independent training set. We suspect that this model does not perform as well as the single prediction model (applied to the event-dependent test sets) because the single prediction model is tested on only the easiest cases, those immediately preceding the end of the admission. The multiple prediction model, on the other hand, has a wider variety of examples (for both training and testing). While the point-wise performance of this model on the event-dependent test set is lower than the performance on the event-independent test, this difference is not statistically significant.

These results demonstrate the importance of admission-based indexing. When testing, admission-based indexing more accurately reflects the desired clinical use case. Moreover, in this more realistic test setting, a model trained using admission-based indexing outperforms a model developed using event-based indexing.

4.2 Training Data Must Be Chosen Carefully to Avoid Biased Models

When we used time of admission as the reference point for in-hospital mortality prediction above, we made daily predictions of in-hospital mortality. This approach is potentially problematic since there are patients who are represented in the training set several times more than average due to longer hospital stays. Here, we consider how sub-sampling the training data can affect performance.

We explore this issue by comparing the performance of the multiple predictions admission-based training set to a sub-sampled training set derived from the admission-based set. We randomly sample 6 times (with replacement) from each patient’s full set of training examples. This number was chosen because it corresponds to the median number of days in an admission and we are building a daily prediction model. We round down from 6.9 because our first prediction is not generated until after 12 hours have passed in the admission, as described above. This ensures that each patient is equally represented in the training set.

<table>
<thead>
<tr>
<th>Sampling Scheme</th>
<th>Mean (STD) AUROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Median Sample</td>
<td>.900 (.003)</td>
</tr>
<tr>
<td>All Samples</td>
<td>.882 (.005)</td>
</tr>
</tbody>
</table>

Table 3: AUROC scores for the admission-based sub-sampling experiment. All Samples corresponds to a multiple prediction admission-based model. Median sample is a sub-sampling (based on the median number of days in an admission) of All Samples. Both are evaluated on a test set that uses time of admission as the reference point.

The results for this experiment are presented in Table 3. From these results we can see that sub-sampling helps performance, improving AUROC from .882 using the original training set, to .900 using the re-sampled set. Sub-sampling reduces the impact of outliers (i.e., patients with extended hospital admissions) on the model, and thus, we are able to learn a more generally applicable model.

4.3 The Use Case Can Significantly Affect Performance

As discussed above, evaluation should mimic the intended clinical use case. Different clinical use cases can result in different models and performance can vary greatly. To highlight this point we consider two additional experiments described below.

Many of the most widely used ICU risk scoring systems, including APACHE and SAPS, are defined based on a single time of prediction. While useful for making care decisions in the early stages of the admission, it is not clear how these scores generalize to multiple predictions (updated based on newly available data). To explore this issue, we
compare a model for predicting in-hospital mortality that makes only a single prediction at 12 hours to one that makes a prediction every 12 hours. While the model that makes multiple predictions has the advantage of additional data, it also has more opportunity to get the prediction incorrect. To compare these two approaches, we test both models identically in a multiple prediction setting (variable prediction horizon).

<table>
<thead>
<tr>
<th>Model Type</th>
<th>Evaluation</th>
<th>Mean (STD) AUROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single Prediction</td>
<td>Rolling</td>
<td>.821 (.009)</td>
</tr>
<tr>
<td>Multiple Prediction</td>
<td>Rolling</td>
<td>.882 (.005)</td>
</tr>
</tbody>
</table>

Table 4: Single and multiple mortality prediction AUROC scores. Single prediction is trained using data available 12 hours into the admission. Multiple Prediction is the previously described multiple prediction admission-based mortality model. Both models are tested on an admission-based test set with predictions made on a rolling basis throughout the admission.

We observe that a training set that makes predictions throughout the admission outperforms a training set that only makes a single prediction at the beginning of the admission (Table 4). A model trained on a single prediction made near the time of admission cannot necessarily be applied to the remainder of the admission. The model benefits from additional examples collected throughout the admission.

For the problem of serum potassium testing, we considered a use case in which one aims to predict future laboratory values every 12 hours. One could imagine a different use case in which one aims to predict the “current” laboratory test value. In this setting, a prediction is made every time a serum potassium test is ordered and the prediction horizon is immediate or essentially zero. To measure performance in this setting, we extract the training and test sets identically: we use each observed serum potassium result in a patient’s admission as the time of prediction and make a prediction using the covariates available prior to this order. The observed result serves as ground truth. In total we extract 649,949 examples, 69,723 of which are positive.

This extraction scheme is not the same as using an outcome-dependent reference point since the clinical use case is one of “on-demand” estimation rather than advance prediction. Note that this model is not directly comparable to an outcome-independent approach. Nevertheless, we can still draw some conclusions about the performance of such an on-demand model (Table 5).

<table>
<thead>
<tr>
<th>Model</th>
<th>Mean (STD) AUROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>On-Demand</td>
<td>.738 (.003)</td>
</tr>
<tr>
<td>Time of Admission</td>
<td>.829 (.002)</td>
</tr>
</tbody>
</table>

Table 5: AUROC scores for the on-demand and time of admission reference point models on the serum potassium task. Note that these problems are very different and their performance isn’t necessarily directly comparable. From the results, we note that the on-demand task is more difficult, despite the shorter prediction horizon. This is somewhat expected since we only make predictions when a laboratory test is ordered, versus every 12 hours. The fact that a test was ordered suggests the clinician suspects the value to differ from that of previous values. Again, while the tasks are similar (“predict potassium values”) the resulting performance is not comparable since test data are extracted in different ways.

While, the on-demand setting does not require imputation (since ground truth is always available), it could fail to mimic a real clinical use case. Once made available, clinicians may query the model more often (e.g., every 12 hours). This would shift the underlying distributions, resulting in a difference in performance.

5 Summary & Conclusion

In this work, we described methods for extracting examples from longitudinal clinical data for prediction tasks. These methods vary in terms of when predictions are made (e.g., in reference to what, and how often), and for how long each prediction applies (e.g., for the remainder of the admission or for the next 24 hours). These decisions should be largely dictated by the clinical use case. We illustrate how these decisions affect extracted data and in turn model performance using two different prediction tasks: hypokalemia prediction and in-hospital mortality prediction.
In both cases, we showed that using an outcome-dependent point of reference yields misleadingly good performance. Creating a test set from such a point of reference amounts to picking the easier examples and does not accurately reflect how the model would perform in practice. For example, a patient about to be discharged from the hospital will almost certainly be characterized by a healthier set of covariates just prior to discharge than a patient who is about to expire. For evaluation purposes, it is imperative that test examples be extracted independently from the label/outcome. Still, while predictions generated by models trained on outcome-dependent examples are biased, they can help uncover relationships between covariates and outcomes and shed light on our understanding of disease processes.

This work serves as a cautionary tale of the importance of carefully extracting examples from clinical data for the purposes of building a predictive model or clinical decision support tool. When training and testing prediction models using retrospective data, careful attention to the problem setup such that it accurately reflects the intended real-world use is critical. Neglecting to do so could result in a dangerous misinterpretation of the model’s clinical value.

Acknowledgements

This research program is supported by the Frankel Cardiovascular Center and the Department of Pathology at the University of Michigan, and the National Science Foundation (NSF award numbers IIS-1553146). The views and conclusions in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of NSF.

References


Secondary Use of Patients’ Electronic Records (SUPER): An Approach for Meeting Specific Data Needs of Clinical and Translational Researchers

Evan T. Sholle, MS¹, Joseph Kabariti, BS¹, Stephen B. Johnson, PhD²,³, John P. Leonard, MD⁴, Jyotishman Pathak, PhD², Vinay I. Varughese, BS¹, Curtis L. Cole, MD¹,²,⁴, Thomas R. Campion, Jr., PhD¹,²,³,⁵

¹Information Technologies and Services Department, Weill Cornell Medicine, New York, NY; ²Department of Healthcare Policy and Research, Weill Cornell Medicine, New York, NY; ³Clinical and Translational Science Center, Weill Cornell Medicine, New York, NY; ⁴Department of Medicine, Weill Cornell Medicine, New York, NY; ⁵Department of Pediatrics, Weill Cornell Medicine, New York, NY

Abstract

Academic medical centers commonly approach secondary use of electronic health record (EHR) data by implementing centralized clinical data warehouses (CDWs). However, CDWs require extensive resources to model data dimensions and harmonize clinical terminology, which can hinder effective support of the specific and varied data needs of investigators. We hypothesized that an approach that aggregates raw data from source systems, ignores initial modeling typical of CDWs, and transforms raw data for specific research purposes would meet investigator needs. The approach has successfully enabled multiple tools that provide utility to the institutional research enterprise. To our knowledge, this is the first complete description of a methodology for electronic patient data acquisition and provisioning that ignores data harmonization at the time of initial storage in favor of downstream transformation to address specific research questions and applications.

Introduction

Secondary use of electronic health record (EHR) data to support biomedical research is challenging. A common approach to secondary use of EHR data is a centralized clinical data warehouse (CDW)¹-⁵ involving dimensional modeling and terminological harmonization of source system data as a precondition for storage and use. To provide CDWs, institutions require extensive resources and may experience slow delivery of solutions for investigators due to maintenance of centralized data curation methods⁶. Additionally, the centralized modeling of data may limit the ability to meet researchers’ specific needs¹,⁶. Optimal approaches to secondary use of EHR data for meeting investigator needs are unknown.

Lacking a centralized clinical data warehouse, our institution needed to meet investigator needs for secondary use of EHR data. We hypothesized that an approach that aggregates raw data from source systems, ignores initial modeling typical of clinical data warehouses, and transforms raw data for specific research purposes would meet investigator needs. To our knowledge, literature describing such an approach to secondary use of EHR data is limited. The objective of this case report is to describe our approach to inform efforts at other institutions.

Methods

Setting

Weill Cornell Medicine (WCM), located on the Upper East Side of Manhattan in New York City, is the clinical research facility and medical college of Cornell University. WCM supports an academic staff of 1,049 physicians and educators and trains over 950 resident physicians and 450 medical students yearly. Faculty clinicians see patients through the Weill Cornell Physician Organization, a multispecialty group practice with more than 900 physicians and more than 20 sites across New York City. WCM faculty physicians have admitting privileges to NewYork-Presbyterian Hospital (NYP), a 2,478-bed hospital with multiple facilities. As long-time clinical affiliates yet separate legal entities, WCM and NYPH frequently collaborate in regards to clinical and translational research. In order to strengthen this collaboration, WCM and NYPH established the Joint Clinical Trials Office (JCTO) in 2013 to grow research activities between the two institutions. The JCTO regularly collaborates with the Clinical and Translational...
Science Center (CTSC) and the information technology (IT) departments at WCM and NYPH for research integrity and resource access purposes.

WCM and NYPH make use of multiple EHR systems from different vendors. In the outpatient setting, WCM Physician Organization Information Services (POIS) provides Epic Ambulatory. In the inpatient and emergency setting, NYPH uses Allscripts Sunrise Clinical Manager (SCM). Additional ancillary EHR systems cover various specialty areas, including perioperative documentation and imaging. Across WCM and NYPH, patients have a shared medical record number (MRN). Multiple interfaces between WCM and NYPH enable sharing of data.

The Information Technology and Services (ITS) department at WCM provides IT infrastructure, management, and service for the WCM community. Included within ITS is the Research Informatics division (RI), which, with financial support from the JCTO and CTSC, administers the Architecture for Research Computing in Healthcare (ARCH) program, a suite of tools and services for investigators to obtain electronic patient data.

System Description

To support the research enterprise with patient electronic data, RI has implemented a methodology and technical infrastructure termed Secondary Use of Patient Electronic Records (SUPER). SUPER undergirds all of RI’s efforts to provide investigators with patient data processes, and comprises multiple components, from data acquisition and manipulation, to feeding other clinical and research tools, to the underlying technical infrastructure. Figure 1 illustrates the workflow that powers SUPER activities.

![Figure 1. Model of flow from source systems to SUPER infrastructure](image)

Data sources

SUPER stores data from multiple clinical and research information systems, comprising the entirety of data gathered to support clinical, billing, and research activities across WCM and NYP. Data enter SUPER and are maintained in the same format in which they exist in their source systems.

For outpatient clinical and billing data, SUPER receives copies of Epic Clarity database tables which is a relational subset of the original transactional data from the EpicCare EHR. These relational data enter the SUPER infrastructure directly without any additional transformation. From the inpatient EHR, SUPER obtains a regular data feed from
NYP’s Clinical Data Warehouse (CDW), which similarly aggregates inpatient billing data in the form of HL7 messages from NYP’s Eagle patient accounting system, and the Jupiter analytics database. Jupiter contains tables copied directly from the Allscripts SCM inpatient EHR, as well as data from ancillary systems, including CompuRecord, the inpatient perioperative EHR for anesthesiology administration, OR Manager provides operative data, and Provation provides documentation from endoscopic procedures. SUPER imports all data from NYP’s CDW and Jupiter database without any form of transformation or harmonization; tables are replicated exactly as they exist in the source systems.

For biospecimen data, SUPER obtains data from Profiler, a locally developed system used by one of the larger tissue banks, and RURO FreezerPro, a commercial specimen management system used to track inventory for other research programs. For tumor registry data, SUPER receives registry data from a third-party vendor responsible for abstracting charts according to North American Association of Central Cancer Registries (NAACCR) standards. WCM’s clinical trials management system feeds directly to SUPER, providing records of patient enrollment in research protocols. For next generation sequence (NGS) data, SUPER receives data from the Standard Molecular genomic information system (GIS) for molecular tests, including the EXaCT whole exome test performed at WCM, and from Genoptix, a third-party laboratory. For case report form data, SUPER obtains data from WCM’s instance of REDCap.

Data acquisition process
As described above, the SUPER infrastructure and methodology depends on multiple partner organizations and teams, all of whom maintain their own clinical and research data systems. Obtaining data from these sources systems requires careful coordination with the teams responsible for their maintenance and care – the data systems feeding SUPER also send data to other systems and, in some cases, support user-facing activities. Given the significant demands imposed by regular data transfers, flexibility is crucial in order to avoid interference with clinical operations and other activities. To facilitate a flexible and modular approach, a series of extract, transform, load (ETL) database operations, application program interface (API) calls, and flat file imports provide data to SUPER. While SUPER pulls data from most source systems, some external systems push data to SUPER. Most source-to-SUPER data transmissions occur monthly while some occur weekly and daily to support specific projects.

The primary intake of data from the outpatient EHR occurs monthly over a one-week period, with each night of the seven-day period handling one section of a large ETL. Generally, each night’s job covers some of the significantly larger tables, along with related smaller lookup and metadata tables. The modular nature of this process allows for one night’s worth of data to be available immediately for use: in the event that a subsequent transfer fails, not all data is useless. Due to the size of the data set and underlying infrastructure limitations, the refresh is limited to 5 to 7 nights a week per source system. To ensure each night’s job proceeds as quickly as possible and ensure the capture of any retroactively imposed changes, WCM POIS will wipe any relevant data from the servers before initiating the transfer of the data set. A delta load approach, while appealing, would render the process overly time-consuming.

For data intake from the inpatient EHR via NYP’s CDW and Jupiter, the acquisition process differs. To allow for fine-grained control, we automate the process through the combination of Microsoft SQL Server Integration Services (SSIS) and SQL Server Agent. Using the proper driver – i.e., a DB2 driver to power the transfer from a DB2 server – we configure an encrypted SSIS package to make a connection to the relevant source tables, run a SQL query pulling out that data, and transfer the data to the SUPER landing server. The SSIS package is configured to ensure the preservation of the source schema. Overall these ETLs span a seven-day period, with each day encompassing several table transfers.

Other sources require different processes. To obtain NGS data from Genoptix, we rely on a cron job run from a virtual machine to pull data from a CSV file located on an FTP server into the SUPER infrastructure. To integrate data from FreezerPro, a Python script polls FreezerPro’s API and pulls data into SUPER. Intake of data from NAACCR’s tumor registry requires a SAS dictionary invoked from the command line, in conjunction with a Python script to transform the data and load it into SUPER. SUPER obtains data from small-scale systems, including the clinical trials management system and REDCap, more frequently, as the load is significantly smaller than other incoming data sources.

ETL code management
All ETL scripts are crucial to SUPER workflows. To ensure their security, currency, and homogeneity, SUPER relies on Subversion (SVN), a version control system that allows developers to save and compare their ETL scripts across
multiple domains and develop the versions of the process over time. With SVN, one change to an ETL script will propagate across multiple systems at once, ensuring that all iterations of a data model are identical.

To ensure that no individual ETL usurps all available resources, thereby affecting other jobs, the RI schedule places refreshes of the larger data sources at the last and first week of the month, and the smaller sources over a weekend. Individual data mart requests and ad hoc reporting take place during the two to three weeks in between each larger refresh, with bigger jobs occurring over a weekend. To avoid interference with day-to-day operations, weeknights and weekends are dedicated for resource-intensive ETL jobs.

ETLs are inherently ambiguous: while they inform the what and where for the transfer of data, they do not inform the how. Paver, a task manager developed in Python, manages priority levels and the order of task execution. With the use of SVN and a collaborative development process, ETL scripts are automatable and reusable. Paver renders them modular, affording the ability to configure the order of a set of scripts and have them run piecewise, thereby automating those reusable scripts in different contexts.

Indexing

SUPER stores several terabytes of source data, with some tables holding billions of rows of data. Without an indexing process, it is impossible to expect any ETL, or even query, to finish within a practical timeframe. However, indexing is time-consuming and must take place after every refresh cycle. To optimize the indexing process and ensure it takes place as quickly as possible, a cron job powered by Paver syncs index generation code with the latest revision and runs the index refresh script every six hours unless an ETL is currently running, signaled by a system flag created as part of the ETL script. The indexing script uses dynamic SQL to drop and recreate stored procedures. These stored procedures reduce active indexing time through the use of a procedural log, rather than a transactional log, and through previewing and optimization of the index processing time. Indexing scripts first check for presence of the index, with its presence signaling the job to move on to the next index. Since indexes are recreated with every refresh, there is no possibility of long term index usage and bug appearances - therefore, the method of ignoring an already created index should be sufficient.

Terminology management

SUPER does not conduct terminology management or data harmonization on sources accessioned to the landing server, instead relying on efforts already put in place by POIS and NYP IS to harmonize data. While not detailed here, a great deal of effort goes into standardizing data as it enters the EHR using an internally developed terminology server, TruData©, that normalizes data against reference and interface terminologies.

RI pulls data from a variety of sources, each of which is organized according to its own internal schema. Some data sources have already been subject to a data harmonization process, leveraging proprietary data models or other schemas and consistently using reference terminologies. Examples include Epic Clarity, which transforms Epic Chronicles data, and the NYP CDW, which has rules for data representation. Others consist of relatively raw transactional data. While internal data dictionaries for individual source systems afford developers and analysts the ability to understand the relations between tables within source systems, exploration of data is necessary to understand table relationships within and across source systems. Therefore, rather than maintain a common data dictionary for the various systems, SUPER relies on source system use of reference terminologies and engineering expertise to query data within and across applications.

Documentation and workflow management

To ensure a streamlined process for maintaining SUPER, analysts and engineers use multiple off-the-shelf work management tools, including ServiceNow, Jira, Subversion, Sharepoint, Box, Outlook, and Slack. RI has implemented a system of workflow management that leverages these tools to ensure all team members know what to do, when to do it, and how to ask for help. ServiceNow, WCM ITS’s internal ticketing system, allows for the triage of data requests and inquiries from investigators and research personnel, as well as infrastructure change requests and access control issues. To track work, including acquisition and integration of new data sources, data quality efforts, and the development of user-facing informatics tools, Jira allows for ticket-level tracking of tasks, priorities, and day-to-day work logging. As previously discussed, SVN allows developers to maintain and share code.

The use of SharePoint and Box allows team members to collaborate on external-facing reports and presentations to the research community, as well to aggregate and track documents gathered from and produced for investigators. A shared Outlook calendar makes all team members aware of ongoing ETL jobs or other pertinent events. Confluence stores wiki documentation written for the reusable code available to all team members, as well as ongoing issues
relating to individual tables, data sources, or user-facing tools. After a researcher submits a request via ServiceNow and engineers use Jira to work toward a solution, team members use Confluence to represent the knowledge distilled through the development process.

A weekly code review session allows software engineers to circulate and present code to their colleagues, exposing problems or issues, requesting help, or working together to develop and implement novel solutions. WCM ITS’s departmental Slack instance allows all team members to stay in constant contact, notifying each other of pertinent issues, as well as facilitating contact with other ITS business units to address issues relating to infrastructure, identity management, or project management.

Hardware and software infrastructure

SUPER consists of four Microsoft SQL Server 2014 database servers and five Linux virtual machines. WCM ITS provides clusters of virtual servers, which allow a more robust and flexible system of maintenance and deployment. The primary server is the landing server\textsuperscript{11}, which uses 40 gigabytes of random access memory (RAM), 21 terabytes of hard disk drive storage, and 20 virtual CPU cores. The landing server requires extensive system resources to ensure that data flows unhindered from source systems. In addition to enabling queries for the most recent data, the landing server provides data to development and production servers.

The SUPER development server is populated with data a week behind the regular refresh schedule of the landing server so as not to perturb data pulls; it also provides a backup should the landing server fail. Configured with 16 gigabytes random access memory, approximately 12 terabytes of hard disk drive storage, and 13 virtual CPU cores, the development server is not as powerful as the landing server, but allows for the development and trials of small ETLs, data mining queries, and relational data exploring. RI endeavors to keep the server as closely in line as possible with the landing server to ensure backup capabilities as well as utility for query transferring.

For production purposes, SUPER has two separate servers. For identified data available to researchers who access RI services, the production servers use 16 gigabytes of RAM, approximately 5 terabytes of solid state drive space, and 9 virtual CPU cores. Production servers are intended to handle multiple concurrent users.

Results

As shown in Figure 2, SUPER has enabled ARCH to support WCM researchers with multiple tools for using electronic patient data. For cohort discovery, i2b2 enables WCM investigators to query de-identified data from EHR and research systems\textsuperscript{12}. For EHR analytics, the Observational Medical Outcomes Partnership common data model (OMOP CDM)\textsuperscript{13} allows researchers to perform robust queries using standardized vocabularies. Alternately, for investigators with specific needs, developers and analysts work together to generate custom data marts from SUPER tables. Using REDCap in conjunction with the REDCap dynamic data pull (DDP) plugin and generalizable middleware to support REDCap DDP, investigators can populate REDCap fields with EHR data stored in SUPER\textsuperscript{14}. For multi-institutional data sharing, the New York City Clinical Data Research Network (NYC-CDRN), which receives funding from the Patient-Centered Outcomes Research Initiative (PCORI), enables queries of citywide EHR data; the SUPER infrastructure enables WCM to contribute data to the effort. Use of SUPER to support additional multi-institutional data sharing initiatives for clinical trial recruitment, including TriNetX\textsuperscript{15} and the National Center for the Advancement
of Translational Science’s Accrual for Clinical Trials program (NCATS ACT)\textsuperscript{16}, is underway. SUPER also previously enabled use of RexDB from Prometheus Research\textsuperscript{17}.

**Figure 2.** Model of data flow from sources to WCM SUPER staging environment.

**Discussion**

To our knowledge, this is the first complete description of a methodology for electronic patient data acquisition and provisioning that ignores data harmonization at the time of initial storage in favor of downstream transformation to address specific research questions and applications. While other institutions have outlined approaches that integrate data from different sources in their raw format and transform them on an ad hoc basis\textsuperscript{18}, the SUPER approach expands previous domain-specific work into the complex realm of health informatics. SUPER has successfully enabled multiple tools that provide utility to the WCM research enterprise, including cohort discovery, electronic data capture, and robust querying. In contrast to centralized data warehouses, the SUPER approach can enable institutions to tailor solutions that meet investigator needs rather than focus on dimensional modeling and other technology-oriented activities.

In characterizing SUPER’s methodology and infrastructure, the question naturally emerged: what is SUPER? Popular terms include data warehouse, data mart, and data lake. A data warehouse aims to render organizational data easily accessible in a credible and consistent fashion, adapt to shifting methodologies and use cases, secure sensitive data elements, drive enhancements in decision making, and serve as a usable and accessible resource to end-users\textsuperscript{19}. The Mayo Clinic’s Enterprise Data Trust (EDT), a “top-down, subject-oriented, integrated, time-variant, and non-volatile collection of data in support of Mayo Clinic’s analytic and decision-making processes,” serves as an example of the application of the data warehouse model within the healthcare setting. The EDT seeks to normalize data from disparate sources, implementing “consistent information models” and shared vocabulary to enable queries that touch data elements enterprise-wide\textsuperscript{1}. Other institutions have also pursued centralized data warehouses with controlled terminology and modeling rules\textsuperscript{3,4,6,7}. While the clinical data warehouse is defined by efforts to conduct semantic harmonization and data structuring during the ETL process from source systems, SUPER does not conduct any data harmonization, instead relying on efforts already put in place by teams maintaining source data.

An alternate approach to the data warehouse is the data mart – a specifically tailored data set specific to one research question, scientific workflow, or operational/quality improvement requirement\textsuperscript{20}. Often extracted from enterprise-wide data repositories, data marts offer an agile approach that is tailored directly to a specific use case. Purpose-built, they offer greater potential utility at the cost of flexibility and effort – each data mart often requires iterative definitional work with a business analyst to determine the specific requirements and often cannot be reused to support research questions outside of its bailiwick. Data marts also come with a particular suite of risks, including redundancy of effort of aggregation, inconsistency between marts, and difficulty connecting results with larger data sets\textsuperscript{20}. Although SUPER enables the creation of data marts, it also constitutes a more holistic effort to aggregate data across the entire institution and is not confined to individual research questions.

A third and relatively novel dimension for considering approaches to enterprise data aggregation and integration is that of the data lake. Coined by James Dixon, the data lake refers to an approach by which institutions aggregate
enterprise data in their native formats and store them unaltered. Analysts and developers then configure analytical components and export data to respond to individual queries and use cases as they emerge, working with users on an ad hoc basis. The data lake is not mutually exclusive with the data mart—indeed, they are intrinsically related, as the data lake relies on the data mart for the analysis and dissemination of the data that feeds its “aquifer.” SUPER seems to most closely fit the definition of a data lake, as it is characterized by its avoidance of data harmonization efforts during the intake process. However, its reliance on terminology management and semantic restructuring by groups maintaining source data feeds distinguishes it from a data lake in the strictest sense of the word. SUPER is, perhaps, comparable to a “data kitchen,” which regularly receives shipments of ingredients from vendors—some processed and some in their raw state—and uses these ingredients to “cook” an array of meals, some simple and designed to satisfy large numbers of people, and some complex, prepared for a specialized audience with particular tastes.

Our implementation of SUPER has yielded a number of pertinent lessons both technical and organizational. From a technical standpoint, SUPER’s ETL jobs are notable for their exhaustive resource use and lengthy duration. Our experience shows the value of developing large ETL jobs in a modular fashion and running them piecewise. While a large transaction log can have substantial utility, this approach offers greater benefit to the overall workflows the system facilitates and enables rapid recovery in the event of an individual component’s failure. Maximizing preprocessing is also crucial, with indexes configured to help as much of the process as possible. Additionally, retaining previous iterations of data sets on a development server creates a constant backup for operational use, obviating the need to ensure that the landing server is constantly available. While teams responsible for maintaining clinical and billing source systems may be initially reluctant to establish a workflow for transferring raw clinical tables, the SUPER approach offloads much of the iterative workflow in determining data flow and conducting tailored exports from source system teams onto RI. Investigators desirous of using clinical data for research may also be unprepared for the ramifications of its nature—underlining that data are accessioned from multiple sources and may not always be in agreement is crucial in setting expectations for the utility and quality of data delivered, especially for research use cases where data cleanliness is paramount.

Some limitations apply to this analysis. The study was conducted at a single site and may not generalize to other institutions. However, it may be potentially useful to sites with existing clinical data warehouses or sites considering their adaptation. Furthermore, the lack of terminology management may limit our approach: however, in a post-meaningful use healthcare setting, this may not be as applicable.

To our knowledge, this is the first comprehensive description of a methodology and infrastructure for secondary use of patient electronic data for clinical research that stores data from source systems in their raw format to serve specific investigator needs. We feel that this methodology possesses unique advantages and that it may be of use to comparable organizations considering potential approaches for secondary use of patient electronic data for research.

Acknowledgements

This study received support from NewYork-Presbyterian Hospital (NYPH) and Weill Cornell Medical College (WCMC), including the Clinical and Translational Science Center (CTSC) (UL1 TR000457) and Joint Clinical Trials Office (JCTO). We thank Rajesh Bollapragada for help in conceptualizing the institutional setting, and Monika Ahuja for developing and describing some of the methodologies that support SUPER. We also thank the members of the Research Informatics team—Brant Lai, Julian Schwartz, Sean Pompea, David Kraemer, Marcos Davila, Jacob Weiser, Prakash Adekkanattu, and Steven Flores—as well as Project Management Office colleagues Cindy Chen and Anthony DiFazio.

References


Disease-Specific Integration of Omics Data to Guide Functional Validation of Genetic Associations

Maya Shumyatcher, BA1, Rui Hong, BS1, Jessica Levin, MS1, Blanca E. Himes, PhD1
1Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, Philadelphia, PA, USA

Abstract

Unbiased genetic association studies, including genome-wide association and whole-genome sequencing studies, have uncovered many novel disease-associated variants. Relatively few of these associated regions, however, have led to insights that are biologically mechanistic or clinically actionable due in part to the difficulty in designing appropriate functional validation studies to understand how variants contribute to disease. Asthma is a complex inflammatory lung disease for which many genetic associations have been identified. Using asthma as a disease model, we designed Reducing Associations by Linking Genes And transcriptomic Results (REALGAR), an app that facilitates the design of functional validation studies by integrating cell- and tissue-specific results of disease-relevant gene expression and other omics studies. Via specific examples, we demonstrate how integrated gene-centric and disease-specific information leads to asthma insights, and more broadly, can help understand complex diseases.

Introduction

The results of genome-wide association studies (GWAS) for various complex diseases have contributed to the growing credibility of using unbiased genetic association approaches for uncovering novel disease variants. In most cases, however, findings from GWAS have not yet led to significant progress toward understanding how identified genomic variants are functionally and clinically relevant. Recent exome sequencing and whole-genome sequencing (WGS) studies are finding an even greater number of variants associated with complex diseases, most of which have not yet been related functionally to outcomes. Among the factors that contribute to the slow translation of genetic association results into functional insights are: (1) functional studies are time-consuming, as each follow-up experiment has to be tailored to a particular complex disease phenotype and type of polymorphism in a genomic region, (2) in order to test genes and variants for function, complex diseases have to be simplified into assays that may not capture the cell-specific, developmental or environmental context necessary for functional elucidation of the gene, and (3) unlike older candidate gene studies, GWAS and WGS studies have identified loci in gene deserts and in genes with no annotated function, making the design of functional experiments even more difficult. In silico approaches that screen genes and variants for potential function are needed to guide the efficient experimental validation of top hits.

Functional validation studies of gene associations often begin with searches for what is known about a specific locus, including what genes are nearby and whether associated variants are expression quantitative trait loci (eQTL) and/or regulators of transcription of these genes. Although this information serves as a starting point, it does not offer phenotype-specific mechanistic clues about how the genes in question modify relevant biological pathways. To obtain such information, researchers often search public databases to find out the tissue(s) where identified genes are expressed and under what disease and treatment conditions they are differentially expressed. Public gene expression data from sources such as the Gene Expression Omnibus (GEO) are a primary resource for answering these questions, but many wet-lab and clinical researchers do not have the proper expertise or dedicated computational resources necessary to obtain and integrate gene expression microarray, RNA-Seq, and other omics results. Even researchers who do have such resources often repeat similar analytical tasks every time a new association locus is found. Having an integrated resource of tissue-specific results from expression and other omics studies that is guided by disease-specific knowledge would facilitate prioritization and rational design of experiments that can provide clinically actionable insights.

Asthma is an episodic inflammatory lung disease characterized by variable airflow limitation and airway hyperreactiveness that affects over 25 million Americans. One of the most common pharmacologic treatments of asthma consists of glucocorticoid medications, given in inhaler form as maintenance therapy or oral form to alleviate exacerbations or severe disease. Glucocorticoids, which are also used to treat various inflammatory diseases, act by modulating transcription of genes in a tissue-dependent fashion. The genetics of asthma has been studied for over 20 years, and consortium GWAS carried out in Europeans and diverse North American populations have
identified robust associations at loci such as the **GSDMB-ORMDL3-ZPBP2 17q21 locus, HLA-DQ, IL1RL1, IL18RL1, IL33, TSLP, SLC22A5, SMAD3, and RORA**. Most strikingly, the 17q21 locus association has been replicated in over 15 independent diverse populations and has been linked to childhood-onset asthma. How associations in this region are functionally related to asthma remains unclear. Because asthma features are heterogeneous and genetic variation also contributes to asthma-related phenotypes such as glucocorticoid response, there is difficulty in determining what specific traits are modulated by measured statistical associations with asthma.

Here, we present REducing Associations by Linking Genes And transcriptomic Results (REALGAR), an app that integrates results from gene expression, RNA-Seq and other omics studies to help guide validation experiments for gene associations. We use asthma as a disease model to show how gene-centric information that is gathered in a disease-specific fashion can be useful to better understand associations.

**Methods**

**Data Sources**

**Gene Expression.** A search for all human gene expression microarray datasets matching search terms related to “asthma” and “glucocorticoid” was performed in GEO. Raw signal intensity and phenotype files from relevant datasets were downloaded. The R package arrayQualityMetrics was used to perform quality control analysis on each dataset. Samples that did not pass quality control measures within datasets were excluded from subsequent analyses, and any provided covariates that were deemed to influence outcomes and/or produce batch effects via cluster analyses were noted. Pre-processing of raw signal intensities was performed with Robust Multi-Array Average (RMA). Three asthma-related RNA-Seq datasets were obtained from GEO.

**Gene Association.** Publicly available single nucleotide polymorphism (SNP) association results from asthma GWAS were obtained from three sources: (1) the Genome-Wide Repository of Associations Between SNPs and Phenotypes (GRASP), a catalog hosted by the National Heart, Lung and Blood Institute that contains all publicly available SNPs with reported p-values <0.05, (2) the GABRIEL Consortium, and (3) the EVE Consortium. GABRIEL results were filtered to select SNPs with p-values <0.05, and the EVE results, which contain distinct p-values for different racial/ethnic groups, were filtered to select SNPs that had p-value <0.05 in at least one group. The UCSC liftOver executable and chain files were used to convert the GABRIEL genome coordinates from the hg17 genome build into the hg19 genome build, which is the version used by GRASP and EVE. SNP locations were mapped to genes (within 10kb on either end) using BEDTools. ChIP-Seq results from this dataset were obtained via the UCSC genome browser, where they are reported according to the hg19 genome build, and mapped to genes (within 10kb on either end) using BEDTools.

**Glucocorticoid Receptor DNA Binding Sites.** ChIP-Seq results for a variety of tissues and transcription factors have identified protein-DNA binding events that may modulate transcription. Of relevance to asthma, one study identified regions of the genome where glucocorticoid receptors bind after 1 hour of treatment with a glucocorticoid (100nM dexamethasone) in A549 cells, an adenocarcinomic human alveolar basal epithelial cell line. ChIP-Seq results from this dataset were obtained via the UCSC genome browser, where they are reported according to the hg19 genome build, and mapped to genes (within 10kb on either end) using BEDTools.

**Gene Expression Analysis**

For microarray data, differential expression was quantified using the R package limma. For datasets with noticeable batch effects, the R package SVA was used to perform differential expression analyses while adjusting for batch effect. Fold changes were computed from the log2 fold changes output by limma or SVA, and standard deviations were computed as log2 fold change divided by the moderated t-statistic output by limma or SVA. Results from 27 asthma and glucocorticoid response datasets were saved for inclusion in the app.

RNA-Seq data was analyzed with Kallisto to quantify transcript abundance according to reads mapped to the hg38 reference genome and Sleuth to compute transcript-level differential expression results. Gene-level results displayed in REALGAR correspond to those of the transcript with the lowest differential expression p-value, in cases where genes had more than one transcript with reported results. The fold change and standard deviation reported in the app correspond to the beta and standard error of beta measures that are output by Sleuth.

For both microarray and RNA-Seq, differential expression results were computed for the test condition of interest relative to within-study controls (i.e. asthma vs. non-asthma samples, or glucocorticoid treatment vs. vehicle). Results of analyzed datasets were stored for use with the app.
To obtain an illustrative hypothesis related to our datasets, we conducted a meta-analysis of all available glucocorticoid response microarray datasets (Table 1) using the R package GeneMeta\textsuperscript{27}. Microarray probes were matched to official gene symbols using Bioconductor database packages for corresponding microarray platforms, and gene-level intensities were computed by obtaining the mean of probe intensities for all probes matching to an official gene symbol. Matrix files for overlapping genes of each study were used by GeneMeta to compute combined two-sided z-scores.

### Table 1. Glucocorticoid response datasets used for meta-analysis

<table>
<thead>
<tr>
<th>GEO Accession Number</th>
<th>Microarray Platform</th>
<th>Study Description*</th>
</tr>
</thead>
<tbody>
<tr>
<td>GSE55876</td>
<td>Affymetrix Human Gene 1.0 ST Array</td>
<td>Lymphoblastic leukemia cells; (N=6) 3 dexamethasone (100nM; 6h), 3 vehicle</td>
</tr>
<tr>
<td>GSE55877</td>
<td>Affymetrix Human Gene 1.0 ST Array</td>
<td>Lymphoblastic leukemia cells; (N=6) 3 dexamethasone (100nM; 6h), 3 vehicle</td>
</tr>
<tr>
<td>GSE4917</td>
<td>Affymetrix Human Genome U133A Array</td>
<td>MCF10A-Myc cells; (N=6) 3 dexamethasone (1uM; 24h), 3 vehicle</td>
</tr>
<tr>
<td>GSE22152</td>
<td>Affymetrix Human Genome U133 Plus 2.0 Array</td>
<td>Lymphoblastic leukemia cells; (N=12) 6 dexamethasone (100nM; 6h), 6 vehicle</td>
</tr>
<tr>
<td>GSE22779</td>
<td>Affymetrix Human Genome U133 Plus 2.0 Array</td>
<td>Peripheral blood mononuclear cells; (N=6) 3 dexamethasone (10uM; 6h), 3 vehicle</td>
</tr>
<tr>
<td>GSE61880</td>
<td>Affymetrix HT HG-U133+ PM Array Plate</td>
<td>Monocyte derived macrophages; (N=6) 3 dexamethasone (100nM; 10h), 3 vehicle</td>
</tr>
</tbody>
</table>

*Each entry contains: cell type; total sample size; sample size and treatment condition for cases, sample size of vehicle controls

### REALGAR Design and Features

The RStudio package Shiny, which enables creation of web applications that users can easily engage without needing a background in programming\textsuperscript{28}, was used to create REALGAR, as its end users include wet-lab researchers with little computing experience. The app code, results of analyzed datasets and GENCODE hg19 reference genome files were saved on a DigitalOcean droplet containing a Shiny server that includes various R packages to retrieve and display results (i.e., forestplot, lattice, and Gviz). The app is displayed at [http://himeslab.org/realgar](http://himeslab.org/realgar), and the full code is available at [https://github.com/HimesGroup/realgar](https://github.com/HimesGroup/realgar).

To use the app, an official gene symbol or SNP ID is provided by a user (Figure 1). Users can select (1) tissue and (2) asthma types to be included by clicking on corresponding checkboxes, (3) what asthma-related drug treatment gene expression results to display (e.g., glucocorticoid, beta-agonist or vitamin D treatment), and (4) which of the GWAS results (GRASP, EVE and/or GABRIEL) to include. If EVE is selected, an additional option box allows the user to choose results from specific populations (i.e., African American, Latino, European American, or all subjects). As selections are made, the analyzed GEO datasets that match the selected criteria appear in a table in a separate “Datasets loaded” tab that provides: (1) GEO accession numbers that link directly to GEO entries, (2) PMIDs for papers, when available, that link directly to PubMed entries, and (3) a brief description of each dataset.

Study-specific results are visualized as a forest plot in which the selected gene’s fold changes for all selected tissue and asthma type comparisons are shown as computed relative to within-study controls (Figure 2A). If a SNP ID, rather than a gene symbol, is entered, results for the gene closest to that SNP are displayed. If one or more of the treatment options are selected, a second forest plot displays the fold changes of this gene under the treatment(s) selected, relative to within-study controls, for all selected tissue types with available data. Forest plot rows are colored according to the negative log10 of the corresponding false-discovery rate adjusted p-values (i.e., q-values), and an adjoining table displays study information, including tissue type and endotype or treatment, as applicable.

Below the expression results, an ideogram of the selected gene’s chromosome with its location indicated by a red line is shown, along with a genome track displaying the gene’s transcripts (Figure 2B). If there are SNPs with GRASP, EVE or GABRIEL asthma association results and/or glucocorticoid receptor (GR) binding sites within
10kb of the gene’s boundaries, then association and/or GR binding site tracks will also be present. GR binding sites are colored according to the ENCODE-provided binding scores, with higher scores corresponding to brighter colors. SNP association results are colored according to negative log10 of p-value, so that lower p-values correspond to brighter colors. Individual plots can be downloaded in .png format, and tables of expression results and SNPs can be downloaded in .csv format. GWAS results displayed can be downloaded, allowing users to obtain quantitative results of the associations across populations.

Figure 2. REALGAR output. A) Forest plots show gene fold-changes of differential expression, with colors corresponding to q-values. Example shown for SRC indicates that its mRNA is increased in CD8+ cells of severe asthma subjects and decreased with glucocorticoid treatment in macrophages. B) Genome tracks show gene transcripts, GR binding sites and SNPs with GWAS results. Example shown for SRC reveals that asthma-associated SNPs are within/near GR binding sites.
Results

Generating Hypotheses That Link Gene Associations to Function with REALGAR. The tissue-specific results available in REALGAR offer insights about the role of prominent asthma-associated genes. Looking up the 17q21 loci genes GSDMB, ORMDL3 and ZPBP2 in the app reveals that glucocorticoid treatment does not alter the expression of these genes, nor are they differentially expressed among subjects with asthma in any of the available tissue types. Thus, these genes are not involved in glucocorticoid or immune responses, nor are they useful as asthma biomarkers, according to the evidence provided. In contrast, REALGAR shows that IL1R1 is highly glucocorticoid responsive in macrophages, where dexamethasone treatment induces a 3.41 (95% CI: 2.65-4.39) increased fold-change vs. control (q-value 3.13x10^-6). Additionally, IL1R1 mRNA has decreased expression in nasal epithelium of asthma subjects with allergic rhinitis vs. non-asthma controls, exhibiting a fold-change of 0.60 (95% CI: 0.49-0.73; q-value 0.020). These tissue-specific results help suggest experiments that may clarify the role of IL1R1 in asthma.

Using REALGAR to Re-Prioritize Disease-Associated Genes for Functional Validation Studies. By combining GR-binding site results with SNP results from GRASP, EVE and GABRIEL, 511 unique asthma-associated SNPs that were within GR-binding sites were identified. Among these was rs6090585, within an intron of the SRC Proto-Oncogene, Non-Receptor Tyrosine Kinase (SRC) gene, which had a nominally significant association with asthma in the GABRIEL GWAS consortium study (p-value 0.010; OR 1.13 [1.03-1.24]). Although this SNP would not be a preferred candidate for further study based on the GWAS results alone, the integrated data in REALGAR provides further evidence that the locus may be relevant to asthma: (1) it is located in a putative GR binding site, (2) SRC mRNA expression is increased in CD8+ cells in severe asthma (fold change 1.33 [1.17-1.5] vs. non-asthma controls; q-value 0.012), and (3) SRC mRNA expression is decreased in macrophages treated with the glucocorticoid dexamethasone (fold change 0.47 [0.35, 0.64] vs. control; q-value 0.0038) (Figure 2). Another asthma-associated SNP within a GR-binding site was rs3759324, a SNP located within introns of two genes: Lymphotoxin Beta Receptor (LTBR) and Sodium Channel Epithelial 1 Alpha Subunit (SCNN1A). This SNP was nominally associated with asthma in the GRASP study (p-value 0.0044; OR 1.07 [1.02-1.12]), and thus would not be prioritized for further study based on GWAS data. Considering other data in REALGAR, however, increases support for the hypothesis that this locus is related to asthma: (1) rs3759324 is located in a putative GR binding site with the highest possible ENCODE binding score (1000), (2) the GR-binding site in question occurs fewer than 30kb upstream of several SCNN1A transcripts, suggesting mRNA levels of this gene are modulated directly by glucocorticoids, (3) SCNN1A mRNA expression is increased in CD8+ cells in severe asthma (fold change 1.53 [1.22-1.92] vs. non-asthma controls; q-value 0.027), and (4) SCNN1A mRNA expression is increased in lens epithelial cells treated with dexamethasone (fold change 14.98 [9.34, 24.01] vs. control; q-value 0.023). In contrast, no significant changes were observed for LTBR mRNA levels in any disease state or tissue type tested, although the GR-binding site containing the SNP of interest lies upstream of several LTBR transcripts. According to these results, SCNN1A should be prioritized for study as an asthma gene, and specifically as a modulator of glucocorticoid and/or immune response. These examples highlight how REALGAR’s integrated data can aid in re-prioritizing associations for functional validation studies.

Insights Provided By Meta-Analysis of Single-Modality Data in REALGAR. In addition to using REALGAR to understand genetic associations, integration of data types in the app provides novel insights about the role of genes in disease. For example, a meta-analysis of six of the glucocorticoid-response expression microarray datasets in REALGAR found that Suppressor of Cytokine Signaling 1 (SOCS1) gene ranked as the “top” gene according to magnitude of two-sided z-score (6.2; two-sided z-score false discovery rate <10^-5) (Figure 3). While SOCS1 was highly responsive to glucocorticoids in most tissues, structural cells were an exception: SOCS1 was not responsive in epithelium, and results in airway smooth muscle were mixed, with a nominally significant decrease observed in one study (fold change 0.74 [0.65, 0.85]; p-value 6.1x10^-3; q-value 0.089) and an increase in another (fold change 1.81 [1.30, 2.52]; p-value 3.58x10^-3; q-value 0.020). Previous studies have identified SOCS1 as a target via which glucocorticoids regulate inflammation and a mediator of rhinovirus-induced asthma exacerbations. REALGAR results provide complementary information: (1) a GR-binding site near the gene suggests that SOCS1 is directly activated by glucocorticoids, and (2) the results of 17 microarray studies and RNA-Seq show that SOCS1 mRNA levels are not significantly different between asthma and non-asthma subjects, except for in bronchial epithelial cells, where they were decreased in asthma subjects (fold change 0.83 [0.76, 0.91]; q-value 0.021). These observations suggest that glucocorticoids may not be able to rectify differences in SOCS1 levels that are present in bronchial epithelial cells of asthma subjects. Thus, aggregating single-modality data allows for detection of patterns
not apparent in single datasets, while complementary information available in REALGAR allows for quick insights into disease-specific gene function.

**Figure 3.** *SOCS1* was the most significant differentially expressed gene in response to glucocorticoids across all tissues. Its levels increased highly in immune cells and were not changed in structural cells, with the exception of airway smooth muscle, which had mixed results between datasets.

**Discussion**

Integration of omics data is challenging, as it requires a thorough understanding of heterogeneous data types, as well as computational resources and expertise to store and analyze large datasets\(^{31}\). Various public resources host omics data, but much work remains in maximizing this data’s ability to inform human pathobiology by presenting results derived from omics data integration to wet-lab researchers, who can validate hypotheses. The development of novel computational tools that integrate and visualize omics datasets is vital in bridging the gap between streams of raw data and clinically relevant findings. Several publicly available apps that are helpful in understanding GWAS data exist. For example, HaploReg is a widely used tool that allows users to look up information about a SNP, including others in linkage disequilibrium to it, chromatin state and protein binding annotation from ENCODE, sequence conservation, its effect on regulatory motifs, and associated eQTLs\(^{32}\). A visualization-centered app that provides similar information is LocusExplorer, which is built using the shiny R framework to provide high resolution graphics of omics data for a genomic region\(^ {33}\). Despite their utility, these tools’ insights are largely not tissue-specific and hence are limited in their ability to provide insights into disease and treatment mechanisms, which are tissue specific. Apps that focus on tissue-specific analysis include NetWAS, an efficient and visually appealing app that integrates GWAS results with tissue-specific networks derived from gene expression data to re-prioritize disease-gene associations\(^ {34}\). While NetWAS includes many tissue-specific results, it does not use disease-specific knowledge to guide integration and presentation of results to wet-lab researchers, who answer specific, hypothesis-driven questions. We built REALGAR as a tissue-specific and disease-focused app to ensure its results are immediately helpful to wet-lab biologists, as well as to researchers who seek to carry out unbiased data analyses.

A well-known challenge in understanding complex diseases such as asthma is that disease-associated polymorphisms vary by race and ethnicity. This factor will be important to address in bridging the gap from omics results to clinically relevant insights. REALGAR allows users to view GWAS results at a population-specific level, and functional validation studies that harness the granular nature of these population-level results can lead to improved understanding of the clinical significance of disease-specific loci. A detailed understanding of the
correspondence between genetic variants and clinical outcomes will be an important stepping stone towards more personalized medical care.

In ongoing work, we are improving REALGAR by including datasets for more phenotypes, and while we continue to be focused on pulmonary disease phenotypes, future versions will include other complex diseases. In addition to incorporating a broader set of phenotypes, we are adding additional data types, namely eQTL and additional RNA-Seq results. To complement the population-level GWAS results currently included in the app, we will integrate population-specific information for the gene expression results, as this information is available for some of the datasets. Rather than duplicate the compilation of SNP data that exists in other tools, we seek to focus on providing novel integration of disease-focused omics datasets. While scoring metrics across integrated datasets remain highly biased and arbitrary from a biological perspective, future versions of REALGAR will include a scoring metric that weights user-selected datasets to provide an overall ranking that is based on ranks within datasets. As with all \textit{in silico} tools, the goal of such scores will be to provide evidence-based hypotheses for experimental studies.

Conclusion

We created REALGAR, an app that integrates gene- and disease-specific data to present insights that are helpful for designing functional validation studies in support of genetic associations. REALGAR has an intuitive, user-friendly interface and, while it harnesses the computational power of R, requires no programming background for its use. We provide several specific hypotheses about asthma-related genes based on REALGAR results that may lead to biological insights about asthma and its treatment.

References

Data Visualization of Home Care Admission Nurses' Decision-Making

Paulina S. Sockolow DrPH MS MBA1, Yushi Yang PhD2, Ellen J. Bass PhD1,2, Kathryn H Bowles RN PhD3,4, Annika Holmberg BSN RN3, Sheryl Potashnik PhD1

1Drexel University, College of Nursing and Health Professions, Philadelphia PA; 2Drexel University, College of Computing and Informatics; 3University of Pennsylvania School of Nursing, Philadelphia, PA; 4Visiting Nurse Service of New York

Abstract

The hospital to home care admission process is when nurses make important decisions about the post-transition episode, including medication reconciliation, plan of care, future visit patterns, and the inclusion of other disciplines. It is not clear how nurses get and use information to support decision-making. We conducted a focus group case study with six admitting home health nurses at a rural agency in Pennsylvania. We analyzed the data using thematic analysis and using our enhanced custom high level node-link diagram that highlights the relationships between decisions, tasks and information themes and sub-themes. The visualizations will be evaluated via review with home care subject matter experts.

Introduction

Home health agencies admit 11 million patients per year to home care1 with demand growing due to an aging population living longer with multiple chronic diseases and patient preferences for home care over facility-based care.2,3 Home care patients tend to be vulnerable older adults; 30% experience a readmission to the hospital within 60 days and many are readmitted to the hospital within the first two weeks of discharge.1 The hospital to home care admission process, when nurses make critical care planning decisions about the post transition episode, is therefore critical to quality healthcare and to reducing hospital readmissions.3,4 However, nurses’ information needs and how nurses obtain and use information for decision making during home care admission are currently poorly understood. What is known is that information relevant to a safe transition and developing the plan of care is often lacking at the start of home care. In fact, home care nurses receive only half of the recommended data.5 Missing, incomplete, or incorrect information is likely to impact the quality of the care plan which in turn may effect quality of care and patient outcomes.

This paper presents focus group findings of home care nurses’ tasks, information needs and decisions during the admission process. These findings will be integrated with those from observational studies and case reviews that are in progress in order to develop recommendations to enhance the home care admission process and to inform health information technology standards for home health agency electronic health record (EHR) systems. In addition to leading to home care domain related EHR requirements, this work introduces methods to support the analysts who seek to understand how nurses make decisions in relation to the information and tasks, and to facilitate the communication of findings to stakeholders including the home health agency nurses and other subject matter experts. Time with subject matter experts is precious and thus having supporting methods such as data visualization to represent the key issues for discussion is helpful.

A focus group was conducted at a rural Pennsylvania home care agency that uses a point of care EHR system. With respect to the home care domain, the focus group was designed to explore four main clinical decisions at the home care admission:

1. Medication reconciliation (process of creating an accurate list of medications that the patient should take and comparing that list against the physician's discharge orders and the medications in the home);
2. Plan of care (orders for assessment, education, and performance for the problems that the nurse has decided need to be addressed in the care episode);
3. Visit timing and frequency (the next skilled nursing visits following the admission and the total number of follow-up visits); and
4. Inclusion of other disciplines (“Services”) in addition to skilled nursing (additional agency services needed, e.g., physical therapy, social work) in the care plan.

To understand the associated tasks and information needed to make the four clinical decisions, we developed a custom node-link diagram visualization. The short term goal was to support the review of focus group findings with subject matter experts. The longer term goal was to inform health information technology standards. A node-link diagram, sometimes called a network diagram, is a visual representation that shows relationships using links (also
called edges) between nodes (also called vertices). Nodes are generally represented as shapes with or without text labels. Links are represented as lines between the nodes. According to Knuth, versions of node-link diagrams have been used since the 1200s. These diagrams have proliferated in use across many domains and with many types of representations, including health informatics. The links can be undirected showing association, or they can be directed showing the direction of the relationship. Arrows are typically displayed on the links to indicate direction.

This manuscript presents a focus group case study executed to identify the relationships between admission nurses’ decisions, tasks, and information gathering and use. It presents our enhanced node-link diagram concept developed to represent the qualitative findings from the focus group. The concept was informed by collaboration between experts in the clinical decision making process and human factors professionals. This manuscript introduces the features in the enhanced node-link diagram and presents visualization guidelines for how to instantiate each feature in order to highlight findings considered significant by clinical experts. The focus group data are presented with prototype visualizations that meet the guidelines. The manuscript presents a discussion of the use of the designed representations and highlights future work.

Methods

The Drexel University Institutional Review Board approved this study.

Participants

All participants provided informed consent. The study took place with six admitting registered nurses working in a free-standing, small, rural agency that serves a majority white population with lower socio-economic status. When conducting admissions, the clinicians at this agency at the time of the focus group used a laptop-based commercial EHR from Allegheny Software Publisher that is widely used in rural agencies in the region.

Focus Group Data Collection Procedure

The participating agency limited the focus group data collection period to 75 minutes. The researchers designed the focus protocol with this limitation in mind.

The researchers developed a demographic data collection instrument and a set of open-ended focus group questions. The demographic instrument was modeled based on the demographic data collection section of the EHR Nurse Satisfaction Survey which assessed clinician satisfaction with EHR impact on clinical process. Respondents documented their age, gender, ethnicity, race, job title, prior experience with EHRs, years of experience in health care, and self-assessment of computer knowledge.

The open-ended questions were developed based on the team’s experience. The questions focused on the admission decision making process and how nurses use the EHR and other information to document the care plan and medication reconciliation. The care plan defined the problem list or priorities that the agency will address, the timing of the next visit and the total number of visits, and inclusion of the other disciplines in addition to skilled nursing. Nurses were specifically asked about the start of care visit:

1. What tasks they needed to complete for each of the clinical decisions (i.e., medication reconciliation, visit pattern and frequency, care planning, services needed); and
2. Whether the task generated information they needed for the related clinical decision.

KB, an experienced focus group moderator and home care researcher, led the focus group. She elicited detailed responses from participants following the initial responses. All responses were recorded, including responses not related to the questions posed. As participants responded to the moderator-posed questions, responses were written on flip charts that were adhered to the meeting room walls for ease of reference. Using the participant input, researchers used colored circles to code tasks and decisions by topic (see an example in Figure 1). Visual display of the documented responses provided opportunity for member-checking and respondent feedback. Researchers also documented in field notes. In addition, the focus group was recorded using digital voice recorders.

A 30 minute follow-up phone call with two focus group members addressed member-checking and elicited responses for questions that were not addressed during the 75-minute focus group. For the follow-up call, the team selected focus group members who had contributed the most responses during the focus group and who were demographically representative of the agency’s nurses.
Figure 1. Example flip chart page for the “Start of Care tasks” with activities and color-coded circles representing the associated four clinical decisions (e.g., blue represents plan of care decision, green represents visit timing and frequency decision).

Data Visualization

Table 1 identifies the features to be represented in the visualization.

Table 1. Visualization Features

<table>
<thead>
<tr>
<th>Features</th>
<th>Representation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decisions that nurses make during the start of care</td>
<td>Rectangle with decision name</td>
</tr>
<tr>
<td>Tasks that nurses complete to make each decision</td>
<td>Rectangle with task name</td>
</tr>
<tr>
<td>Information that nurses use to make each decision</td>
<td>Rectangle with information element name</td>
</tr>
<tr>
<td>Association between tasks and decisions</td>
<td>Link</td>
</tr>
<tr>
<td>Association between tasks and information</td>
<td>Link</td>
</tr>
<tr>
<td>Number of tasks associated with each decision</td>
<td>Decision node rectangle height</td>
</tr>
<tr>
<td>Number of decisions associated with each task</td>
<td>Task node height (left side of isosceles trapezoid/triangle)</td>
</tr>
<tr>
<td>Number of information themes associated with each task</td>
<td>Quadrilateral height (right side of isosceles trapezoid/triangle)</td>
</tr>
<tr>
<td>Number of tasks associated with an information element</td>
<td>Information node height</td>
</tr>
<tr>
<td>Directional flow between task and information</td>
<td>Arrows</td>
</tr>
<tr>
<td>Indication that a theme can be decomposed into sub-themes</td>
<td>Filled triangle</td>
</tr>
<tr>
<td>Number of sub-themes that compose an information theme</td>
<td>A number associated with information theme</td>
</tr>
<tr>
<td>Number of sub-themes of information theme that are associated with a task</td>
<td>Link line style thickness (the more sub-themes, the thicker the line style)</td>
</tr>
<tr>
<td>The sub-themes that compose an information theme</td>
<td>List</td>
</tr>
</tbody>
</table>

The main nodes are decisions, tasks, and information gathered and/or used. In the diagram, nodes are grouped by category (decisions, tasks, and information) and listed in vertical columns. Each category is color-coded. This organization supports the analyst to see the list of decisions, tasks, and information elements. The redundant coding (use of columns and color coding of the node shape) help to differentiate one category from the other.

Links are used to represent associations between nodes. The vertical edge of the node is proportional to the number of links adjacent to that edge. Directed links indicate whether information is generated or used. For example, if the directed link is from a task to information, then the task generates the information.

In qualitative analysis, a concept may be decomposed into lower level concepts. For example, an information node
theme such as “Patient” may have many sub-themes. Representing all of the sub-themes can be useful in some cases but in other cases, it can make a visualization very cluttered. In order to leverage the design concept of overview plus detail,² node themes that have sub-themes have indicators. Optionally, the number of sub-themes as well as the sub-themes themselves can be displayed.

Nodes and links were identified based on the thematic analysis data. To produce the diagrams, the nodes defined from the qualitative analysis were entered into an Excel Spreadsheet (Microsoft Inc.). Each node was associated with a text description for display purposes and related identifiers. Columns identified the links between nodes.

To generate the data visualization, we used Lucidchart (Lucid Software) to represent the nodes (i.e., decisions, tasks, information) and links (relationships). After the nodes and links were displayed, annotations related to the sub-themes were added. Some changes (e.g., spacing of nodes, path of links) were made to the images to improved readability.

Data Analysis

We analyzed the focus group products (i.e., transcripts, visually recorded documentation) with thematic content analysis and presented the analysis using the developed data visualization. All data in response to the open-ended focus group questions was included in the analysis. We presented data from different perspectives to yield synergistic insights.

Thematic analysis is used to inductively analyze data to identify descriptive or topical categories. It involves familiarization of the data to identify themes that emerge, grouping of similar concepts, and consistent coding of all such occurrences. Using principles of thematic content analysis, AH, a nurse and experienced qualitative coder, analyzed inductively the transcribed audio recordings of the focus group sessions to identify descriptive or topical categories. Each theme was labeled as per the framework: as a decision, task, or information. AH created a codebook detailing the rules and definitions of the nodes that emerged. Three authors (SP, AH, PS) participated in the coding process of the focus group transcript. SP used the codebook to assign themes to 50% of the transcript. Two authors (SP, AH) discussed differences in individual coding and updated the codebook and coding until 80% inter-rater reliability was established. AH then updated the second half of the transcript according to the updated codebook. PS joined the coders to oversee the updated theme assignment of the second half of the transcript and codebook.

NVivo (QSR International), a qualitative data analysis software, was used to organize themes that emerged from the focus group transcript. AH and PS identified similar concepts in the transcript, and combined them under assigned aliases under the related theme. Using NVivo, concepts were categorized under the appropriate sub-node (child theme) of the associated larger concepts(s) (parent theme). Parent themes categorized under a clinical decision or a task were to be represented in the node-link diagram as linked concepts.

The team met to review the qualitative analyses and results. The team discussed all themes that emerged and the diagram showing all themes related to all of the clinical decisions and also the start of care process. The team grouped similar themes within each category (tasks, information) based on the underlying concept (e.g., patient characteristics, caregiver characteristics, home characteristics). Team members (AH, YY, PS) grouped information themes into aggregated themes which reflected the underlying concept. To improve presentation and readability, additional diagrams were generated showing the themes related to each separate clinical decision. Information sub-themes were listed in versions of diagrams for the clinical decisions. KB led the team in review of the clinical relevance and validity of the node-link diagram analysis findings.

Results

Participants

The focus group nurses were RNs and care managers, white, young (median age of 32 years) and majority female. They had some experience with EHRs (median of 3 years) and in health care (median of 6 years).

Thematic analysis

The thematic analysis was clustered around the four clinical decisions as well as the start of care process. The focus group identified 28 tasks (Figure 2). For the clinical decisions, the Visit Timing and Frequency decision was associated with 11 related tasks, the Plan of Care decision with 14 tasks, Medication Reconciliation with four, and the Other Services with six. The Start of Care was associated with 22 tasks.

The focus group identified seven themes of information generated from tasks to make clinical decisions:
1. Patient (e.g., vital signs);
2. Resources external to the home (e.g., insurance coverage);
3. Home environment (e.g., clutter);
4. Medication reconciliation;
5. Medication self-administration;
6. Caregiver; and
7. Home care agency resources (e.g., services, policies and procedures).

These information themes were distributed among clinical decisions, with some information types occurring in more than one clinical decision. The seven information themes contained 51 sub-themes. Of these sub-themes, 38 were distributed in the seven information themes as follows (details appear in Figure 2): 22 in Patient Information, 3 Medication Reconciliation, 5 Medication Self-Management, 2 Home Care Agency Resources, 4 External Resources, 1 Home Environment, and 1 Caregiver. The remaining 13 sub-themes were not related to a task and were not shown in the visualizations. Almost all these sub-themes were patient sub-themes (e.g., allergies, baseline function, admission indication), with one external resource theme (i.e., community resources needed). With respect to Medication Reconciliation, the focus group participants clarified that this clinical decision has two main aspects: 1) comparison of the patient’s medications to the discharge documentation medication list (drug name, dosage, frequency, and route) and 2) medication self-management (e.g., the patient understanding and ability to self-manage the medication (e.g., self-administration)).

**Node-link analysis**

The node-link diagram shown in Figure 2 with the abbreviations in Table 2 characterizes the relationships between tasks, information, and clinical decisions. Figure 2 was designed to show the high level findings from the focus group. Figures 3 and 4 were designed to highlight specific design features related to the decomposition of information into sub-themes. Each diagram has three columns: nodes in the same category are organized by vertical position. The left column shows the clinical decisions of interest in the study. The middle column contains the tasks. The right column shows the nodes of information grouped into the aggregated themes. The color of the node indicates node type taxonomy (e.g., clinical decision nodes are blue). The horizontal links between nodes indicate associations between tasks, information, and decisions. The vertical edge of the node is proportional to the number of links adjacent to that edge and can indicate prominent tasks and information. Directed links were included when relevant. For example, if the flow is from task to information, then the task generates the information. An example of the opposite direction is a cluttered home (information) which is a reason for the nurse to document in the EHR outside the home (task).

**Table 2. Abbreviations in Figures**

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>HIPAA</td>
<td>Health Insurance Portability and Accountability Act</td>
</tr>
<tr>
<td>HC</td>
<td>Health Care</td>
</tr>
<tr>
<td>POC</td>
<td>Plan of Care</td>
</tr>
<tr>
<td>OASIS</td>
<td>Outcome and Assessment Information Set</td>
</tr>
<tr>
<td>EHR</td>
<td>Electronic Health Record</td>
</tr>
<tr>
<td>DC</td>
<td>Discharge</td>
</tr>
<tr>
<td>IV</td>
<td>Intravenous therapy</td>
</tr>
<tr>
<td>Med Rec</td>
<td>Medication Reconciliation</td>
</tr>
<tr>
<td>Med Self-Mgmt</td>
<td>Medication Self-management</td>
</tr>
</tbody>
</table>

Figure 2, as a high level summary, illustrates a number of design features. Regarding the clinical decisions, the Start of Care, Plan of Care, and Visit nodes have longer vertical sides (indicating more links) than the Medication Reconciliation and Services nodes. Among the 28 tasks, the majority have nodes with sides of shorter height (nodes with few links) meaning that there is a small set of information types associated with the tasks. The few task nodes with sides having larger height are: Medication Reconciliation; Disease Management Teaching; Obtain Health History; Conduct Review of Services; the Physical Assessment; and Document in the EHR. These tasks thus are associated with many different types of information. Eleven task nodes do not link to an information node and this lack of associations is apparent in the figure. During the focus group no associated information was mentioned for these tasks: participants did not mention generating or using information types.
Figure 2. Enhanced node-link diagram of four clinical decisions and start of care. Nodes are organized by type (decision, task, and information). Node side height indicates the number of links. Direction links indicate information.
Among the seven information themes, the Patient theme dominates with respect to the number of tasks that generate information. This is as expected as the purpose of the admission is to assess the patient’s situation and needs in order to develop the plan of care.

To support analysis for each clinical decision, we created a separate node-link diagram. Plan of Care (Figure 3) includes 13 tasks and 5 information themes (27 sub-themes). The Conduct Medication Reconciliation task node has the most links to information types and thus its longer left side is salient in the figure. Five tasks did not connect to an information theme and thus come to point on the left side. These pointed sides are also salient and may indicate where further discussion with subject matter experts is warranted. The information group with the most associations was Patient, as expected.

**Figure 3.** Enhanced node-link diagram of plan of care clinical decision information. The total number of information sub-themes of an information theme is indicated with the number next to the triangle in the information theme node rectangle. For a task-information theme pair, the number of shared information sub-themes is indicated by the line style thickness of the associated link.

Links between task and information nodes have different line thickness to indicate the number of associated sub-themes. For example, the task, Establish the Goal of Care, was associated with 5 information sub-themes (Diagnosis, Other Therapy Needed, Patient’s Desired Outcomes, Pre-Hospitalization Baseline, OASIS, and Start Of Care Assessment). Having an ability to be able to drill down into the sub-themes is the focus of the visualization in Figure 4.

Visit Frequency and Timing had 11 tasks, six of which linked to an information theme (Figure 4). This clinical decision had five information themes and 21 information sub-themes. The information theme, Caregiver, appeared in only this diagram. This situation is clear in Figure 2 as there is only one link to this node. However, in Figure 2 it is not clear which decision is relevant. Of special interest were task nodes with fewer links (and thus shorter sides) than expected. In addition, there were task nodes with no links to information nodes. Team members knowledgeable of home care admissions (KB, PS, EB) identified cases where the visualization results did not match their observa-
tions of admissions or their knowledge of the admission process. One would expect, for example, that the Assess Home Safety task would link to the information theme Home Environment. The team was cognizant that sparse or absent links may have been due to data collection limitations (75-minute period).

Figure 4. Visualization of visit timing and frequency decision showing information sub-themes

As in Figure 3, Figure 4 includes the number of links between tasks and sub-themes, as indicated by line width. Figure 4 also includes the specific information sub-themes associated with the tasks in the information nodes. While somewhat cluttered, this representation showed, for example, that information sub-themes such as Health History and Diagnosis can be generated from multiple tasks. This type of analysis would be more difficult to conduct without the sub-theme data.

The diagram for Services (not included as a separate figure herein for space reasons) includes very few nodes, having only six tasks and no information themes. This finding related to Services was unexpected, as in our admission nurse observations, we documented the nurse receiving an intake form with specific information pertaining to the services ordered by the physician. For example, we expected to see the Services decision linked to the task, Review the Home Care Chart, which was linked to the information item, Home Care Agency Resources.

The medication reconciliation diagram (not included as a separate figure for space) included four tasks and three information themes. Two tasks (Educate the Use of IV and Medication Reconciliation) did not link to information. However, we would have expected the task of Educate the Use of IV to link to the Patient information node, to ascertain the patient’s medication self-management capability.

Start of Care (SOC) (not included for space) contained the most tasks, which was expected as the SOC is the main purpose of the nurse’s visit. However, 11 tasks linked exclusively to SOC did not link to an information group. The diagram indicated that the SOC involved many tasks to gather a plethora of information.
Discussion

We investigated home care nurses’ information needs during the admission with qualitative data derived from a focus group. We developed a custom node-link diagram visualization of the themes that emerged related to admission nurses’ tasks and information related to four clinical decisions. The node-link analysis revealed the relationships between tasks, information, and decisions as well as highlighted cases where more data may need to be collected.

From an analyst’s perspective, the visualizations help to answer a number of types of questions. The first are exploratory questions. Information emerges from the node-link diagrams which the analyst may not have noticed otherwise. For example, information about the home environment was sparse (Figure 2). In our observations of home care admissions, nurses often identified the need for occupational therapy. Those therapists would assess the safety of the home in a separate visit. This lack of a connection in the node-link diagram, for example, highlights the need to ensure that the nurse assumes that another clinician (e.g., occupational therapist) is responsible for surveying the home environment. This assumption is based on the nurse’s expectation that the therapist would contribute to an assessment of patient safety.

The second type of question is related to our interpretation of the data, and checking back with the subject matter experts to verify what the data suggest. As indicated in Figure 1, there are many ways links between categories of information can be represented. In Figure 1 relationships were identified using colored circles. During the focus group, the nurses were reminded what the colors represented. The data visualization graphics may be more informative to the nurses as compared to a flip chart, spreadsheet or table. Also it allows patterns to be identified using perceptual abilities. For example, the Visit Timing and Frequency diagram (Figure 4) indicates the Caregiver information theme is linked to Conduct Medication Reconciliation. However, there is no link from Conduct Physical Assessment to Caregiver, contrary to our observations during nurse assessment of patient functional status related to activities of daily living (e.g., ability to bath, toilet). Here, the nurse could be shown the gap (the lack of a link) and asked if the physical assessment task was conducted to garner information about the caregiver situation.

Another type of question is related to admission workflow: showing where the nurse was expending effort to do something that may better be achieved by another entity or process. For example, the nurse obtains eligibility criteria information (e.g., insurance coverage, homebound status) from the patient during the start of care (Figure 2). Obtaining this information does not require clinical expertise, nor does the task generate information used by the nurse. After member-checking, this finding might indicate a change in workflow or a need for interoperability which could be incorporated in a requirements document. Similarly, the diagrams indicate where the nurse conducted many tasks (e.g., start of care, plan of care) which did not produce information needed for the admission.

The fourth type of question is relevant to the methods used to collect information. In this case, did our focus group data collection method provide a consistent set of information? For example, the Medication Reconciliation task is associated with decisions and information, and occurs among all clinical decisions and Start of Care (see five links going into the node in Figure 2). Member checking would inform the team whether the occurrence of different information related to medication reconciliation distributed among the four clinical decisions was meaningful or was an artifact of how the question was asked. Another method question is whether the numerous tasks and the sparse Services diagram indicate that the 75-minute data collection may not have been enough time. The data visualization does suggest where the gaps in data are, so that during the next data collection the team can probe further.

Our custom visualization includes design features to support the production of unique node-link diagrams (Table 1). In this case study, we presented different combinations of the design features in Figures 3 and 4. In future work, we will evaluate the efficacy of the visualizations in check backs with the home care agency. Future work should investigate the implementation of the design guidelines in Table 1 with respect to usability as well as extending the set of features included. We can develop additional representation guidelines related to the links, such as line color or line style, and related to the nodes, such as symbol shape, to represent additional features.

Conclusion

We developed a custom information visualization to characterize the decision-making process of home care nurses during admission. This visualization has advantages, compared to other methods such as a flow chart, information flow diagram, or decision ladder\(^1\) that respectively show detailed process information for an individual activity, how information is communicated, and decompositions of decision sequences and short cuts that can be taken in the process. The node-link diagram can be synergistic with those techniques and be helpful in answering other types of questions. An advantage of the information visualization method was that it provided a high-level view of themes and relationships of clinical decisions, tasks, and information. This high-level representation enabled us to have a
systematic understanding of the decision-making process, i.e., how information is used (and not used) in multiple tasks and decisions. Detailed diagrams with information about sub-themes helped to highlight information used across tasks. We anticipate that the data visualization method will be useful outside of healthcare in domains that have limited numbers of decisions, tasks, and information elements, for those seeking to map information-related decision-making.

Acknowledgements

This research is supported by the Agency for Healthcare Research and Quality (AHRQ) Grant # R01 HS024537. The authors thank Wei Quan for encoding data visualizations that were used in an earlier version of this manuscript. They also thank the home care agency and the participants. And the authors thank the anonymous reviewers who encouraged the authors to improve upon the originally submitted data visualizations.

References

Current State of Electronic Consent Processes in Behavioral Health:
Outcomes from an Observational Study

Hiral Soni, MS1, Adela Grando, PhD1, Anita Murcko, MD1, Mike Bayuk, BS1, Pramod Chandrashekar, MS1, Madhumita Mukundan, B.Tech. 1, Meredith Abrams, MPH1, Marcela P. Aliste, PhD1, Megan Hiestand1, Julia Varkey, BS1, Wentao Zhou, ME1, Caroline Horro, BS3, Michael Saks, PhD2, Richard Sharp, PhD3, Mary Jo Whitfield, MSW4, Mark Callesen, MD4, ChristyDye5, Darwyn Chern, MD5

1 Department of Biomedical Informatics, Arizona State University, Scottsdale, Arizona
2 Sandra Day O’Connor College of Law, Arizona State University, Tempe, Arizona
3 Biomedical Ethics Research Program, Mayo Clinic, Rochester, Minnesota
4 Jewish Family and Children’s Services, Phoenix, Arizona
5 Partners in Recovery, Phoenix, Arizona

Abstract
An integral element of value-based care is care team access to both physical and behavioral health data. Data release processes in both environments are governed by federal and state statutes. The requirements for obtaining consent are complex and often confusing. Little is known about the consent processes and practices in the behavioral health setting, specifically how patients and surrogates engage in the process and their interactions with electronic consent tools. This study analyzes the consent processes from the patient perspective at two community behavioral health clinics. Outcomes include description of the processes using electronic consent, workflows and consenter-provider interactions. Conclusions include need to streamline and standardize consent technologies and improve consenter engagement. This study supports the development of an electronic consent tool, My Data Choices (MDC), funded by the National Institute of Mental Health, that offers individuals with behavioral health conditions more control over their medical records.

Introduction
Behavioral health conditions affect nearly 46 million American adults, a quarter of whom have a serious mental illness1. Substance Abuse and Mental Health Services Administration (SAMHSA) defines behavioral health as the mental/emotional well-being and/or actions that affect wellness. Behavioral health problems include substance use disorders; alcohol and drug addiction; and serious psychological distress, suicide, and mental disorders2. Caring for the total person is important. Patients with behavioral health conditions are more likely to have a chronic medical condition, with nearly 70% of adults with a behavioral health condition having at least one medical condition. Individuals suffering from a serious mental illness (SMI) have higher emergency room, primary care and specialist visit rates1. Patients receiving both behavioral health and physical health treatment often see multiple providers and thereby benefit from integrated care amongst a variety of providers and organizations1. Sharing health data between providers requires obtaining informed consent from the patient or legal guardian of the patient, when the patient is physically, mentally or legally not capable of providing consent3. Informed consent is the most widely recognized ethical safeguard in research and clinical care used to protect patient rights3,4. Patients use informed consent to decide what health information to share and with whom.

Patients with behavioral health conditions often suffer social stigma, fear of losing employment, insurance discrimination or legal concerns5. Such fears might influence their care and related data-sharing decisions5. Studies have shown that patients want more control over their data and the Office for the National Coordinator for Health Information Technology recommends that patients be given more choices at a granular level7,8. Granular control could afford patients more control over how personal health information is shared, with whom, and why. The consent decisions pertaining to sharing health data for care can affect treatment and care coordination, compelling the need for meaningful consent. Meaningful consent involves key components such as patient engagement, education and ensuring alignment with laws and regulations6. Understanding the quality of provider-consenter interaction and patient and legal guardian comprehension and involvement is vital. Studies rating the quality of informed consent and related tools exist to guide the design of
consent processes for clinical research. Our review of the state of the art indicates lack of studies about the quality and effectiveness of informed consents for treatment, and studies focusing on patients with behavioral health conditions and legal guardians.

The conventional means to obtain the informed consents has been paper based hard copy forms. Nevertheless, these forms could be long, verbose and monotonous and retrieving information could be time consuming and cumbersome. With advancements and increased adoption of electronic health records (EHR), over the last decade, studies documenting the introduction and development of electronic consent technologies to support the transition from paper-based to electronic consents have been published. Many universities and organizations have contributed to developing multiplatform open-source consent systems for research and treatment. Apple’s ResearchKit provides an open source framework to create visual consent templates. Similarly, in its early stages, ‘Consent2Share’ is an open source web-based consent tool developed by SAMHSA which supports a patient facing interface with granular and dynamic consent choices. Electronic consents could be more convenient, less time consuming, more informed and interoperable with EHR systems. Studies and pilots have noted improved comprehension, better patient awareness, reduced provider workload, improved patient and provider satisfaction and more effective consent management with the transition to electronic consent systems.

The main objective of this paper is to observe, capture and report on the current state of consent tools and processes for treatment and care at two community behavioral health clinics in Phoenix, Arizona to better understand behavioral health consent workflows and challenges and to identify improvement opportunities related to electronic consent technologies. The outcomes presented in this paper will contribute to the broader goal of deploying a patient-centered electronic consent tool, My Data Choices (MDC), that supports more granular data access for patients with behavioral health conditions and surrogates based on interoperability standards.

Methods

Study Participants

Following Institutional Review Board (IRB) approval, subjects were recruited from two outpatient behavioral health clinics in Phoenix, Arizona. Patients and legal guardians of patients diagnosed with general mental health illness or SMI and providers involved in the consent process, were enrolled. Patients and legal guardians were required to be 21 years old or older and English speakers. Participants were identified by facility staff during routine clinical visits and referred to recruiters. Participating providers at each facility were 21 or older, certified behavioral health professionals with experience as case managers and/or behavioral therapists. Patients, legal guardians and providers to be observed were consented by recruiters before the observation sessions. Participating patients and legal guardians were compensated for their participation with gift cards.

Study Sites

Our study sites include one clinic (Site 1) which offers general behavioral health and social services to children, families and adults of all ages. Licensed clinicians are responsible for treating and consenting the patients at Site 1. Study site 2 offers a range of recovery-focused services to adult patients with SMI. Providers consenting the patients at Site 2 are case managers. Both the study sites use a similar proprietary EHR system with embedded consent management and e-signature systems. The EHR system used in both facilities is widely used in the US, including customizable behavioral health modules. Use of the same EHR system in these two different care environments aided in the capture of comparable and contrastable workflow processes at both sites.

The Network: Arizona’s statewide physical and behavioral health information exchange, also known as ‘The Network’, follows an opt-out consent model where patient data is automatically shared unless explicitly declined to share. About 6.7 million Arizona patients receiving care from 246 participating providers are enrolled in the state. Physical and behavioral health information is shared automatically, except the alcohol and drug abuse records protected under 42 CFR Part 2 regulations, accessible only if the explicit opt-in consent is obtained which requires a patient consent to share personal health information.

Consent Observations

For this non-participatory qualitative study, seven graduate students and one undergraduate student, all from the Department of Biomedical Informatics at Arizona State University (ASU) completed IRB and facility requirements, including Human Subject Protection and HIPAA training, completion of required documents from the study sites and compliance with Level 1 fingerprint security clearance.
Prior to observations, students reviewed hard copies of electronic consent forms used at each site. A semi-structured Observation Outline (Table 1) was developed by ASU researchers to guide observers and was piloted during the consent sessions. The goal of the outline was to capture qualitative and quantitative information such as demographics of the process, formats of consent, quality of patient/legal guardian-provider interactions, and patient consent choices. The Observation Outline was divided into five sections:

1. **Header**: captured information about the observer, study site, date of observation and length of the observation in terms of start and end times.
2. **Demographics of the process**: six questions (Questions 1-6) related to consenter (patient or legal guardian) observed, accompanying individuals, purpose of the visit, types of consent reviewed and discussed, and format of consent (paper, electronic or both).
3. **Quality of provider-patient/legal guardian interaction**: eleven questions divided in four sections (Questions 7-10) assessed the quality of the provider’s consent explanation, questions or clarifications asked by the consenter and provider’s responses, and moods of patient/guardian and providers.
4. **Patient choices**: five questions in one section (Question 11) to record consent choices of patient/guardian, and related discussions.
5. **Additional notes**: allowed observers to document comments that they deemed important during the observation.

The sessions were carried out in providers’ personal offices or facility meeting rooms. Each session was observed by only one student observer. Observers documented processes using the Observation Outline form, but were not permitted to ask questions or otherwise participate in or interfere with the session. Though the providers and patients/legal guardians signed a consent to be observed, they were unaware of the objectives of the observation or the content of Observation Outline.

We assessed the readability of the facility consent forms per Flesch-Kincaid Grade Level (FKGL) test in Microsoft Word software. FKGL scores range from a minimum of grade level 1 but do not have a maximum grade level calculation bound\(^2\). Even though FKGL scale was designed to measure readability of school texts, it is widely used to measure readability of informed consents, patient education materials and medical literature\(^1\). It is anticipated for scores beyond grade level 12 to bear less practical validity.

**Table 1. Observation Outline, created to standardize observations**

| Name of Observer (last name, first name) __________________________ |
| Location (circle) Site 1 or Site 2 |
| Date of observation (month, date, and year) ________________________ |

<table>
<thead>
<tr>
<th>1</th>
<th>Are you observing the consent provided by <strong>client (patient) OR legal representative</strong> (legal guardian or surrogate)?</th>
<th>[ ] Client/Patient  [ ] Legal guardian/ Surrogate</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>In addition to the staff provider and client/legal representative, if applicable, were any other individuals present?</td>
<td>[ ] Yes  [ ] No</td>
</tr>
<tr>
<td>3</td>
<td>What was the <strong>purpose</strong> of the visit?</td>
<td>[ ] New patient  [ ] Referral  [ ] Transfer  [ ] Reassessment  [ ] Other (write in) ________________________</td>
</tr>
<tr>
<td>4</td>
<td>Which <strong>facility consent forms</strong> were reviewed and discussed? (List)</td>
<td>________________________________________________</td>
</tr>
<tr>
<td>5</td>
<td>What <strong>format</strong> did the <strong>provider</strong> use when reviewing the consent?</td>
<td>[ ] Electronic  [ ] Paper  [ ] Both</td>
</tr>
<tr>
<td>6</td>
<td>What <strong>format</strong> did the <strong>client / legal representative</strong> use when reviewing the consent?</td>
<td>[ ] Electronic  [ ] Paper  [ ] Both</td>
</tr>
<tr>
<td>7a</td>
<td>Did the provider explain the consent(s)?</td>
<td>[ ] Yes  [ ] No</td>
</tr>
<tr>
<td>7b</td>
<td>How would you rate the consent(s) explanation <strong>overall</strong>?</td>
<td>[ ] Excellent (Detailed explanation)  [ ] Good (High level explanation)  [ ] Poor (Minimal explanation)</td>
</tr>
</tbody>
</table>
8a Did the **provider** ask the client/legal representative if he/she had **questions or needed clarification**?

8b Did the **client/legal representative** ask questions about consent?

*If answered YES to above question:

8c Did the client ask questions related to what **specific health information would be shared or with whom**?

8d Did the client/legal representative have questions related to how consent would affect his/her care?

8e How would you **rate provider’s answers** to these questions?

8f **Was the client receptive** to the consent explanation provided by the provider in response to the questions?

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>9a</td>
<td><strong>What was the mood of the client</strong> during the consent process? (Choose all that apply)</td>
<td>[ ] Angry/Frustrated [ ] Sad [ ] Uninterested/bored [ ] Overwhelmed [ ] Other: _________________________________</td>
</tr>
<tr>
<td>9b</td>
<td><em>If a legal representative was present, what was the mood of the legal representative</em> during the consent process? (Choose all that apply)</td>
<td>[ ] Not applicable [ ] Angry/Frustrated [ ] Sad [ ] Uninterested/bored [ ] Overwhelmed [ ] Other: _________________________________</td>
</tr>
<tr>
<td>10</td>
<td><strong>What was the mood of the provider</strong> during consent process? (Choose all that apply)</td>
<td>[ ] Angry/Frustrated [ ] Sad [ ] Uninterested/bored [ ] Overwhelmed [ ] Other: _________________________________</td>
</tr>
<tr>
<td>11a</td>
<td><strong>Was consent given</strong>? (Did the client/legal representative give permission for information to be shared?)</td>
<td>[ ] Yes, agreed to share ALL information [ ] Yes, agreed to share SOME information and to restrict access other information [ ] No, refused to consent share ALL information</td>
</tr>
<tr>
<td>11b</td>
<td>If the client/legal representative <strong>initially refused to share</strong> specific information, did the provider then provide a discussion or explanation?</td>
<td>[ ] Yes [ ] No</td>
</tr>
<tr>
<td>11c</td>
<td>If YES, how would you <strong>rate the discussion</strong> by provider?</td>
<td>[ ] Excellent (Thorough discussion) [ ] Good (High level discussion) [ ] Poor (Minimal discussion)</td>
</tr>
<tr>
<td>11d</td>
<td>Was the consent provided after the above discussion?</td>
<td>[ ] Yes [ ] No</td>
</tr>
</tbody>
</table>

**Workflow Analysis**

After completing the observations, ASU students created mental models of the consent workflows at Sites 1 and 2. Discrepancies were resolved through consensus. The purpose of this process was to identify similarities and
differences of consent processes between the facilities. A generalized workflow was created using a free, web-based workflow modeling tool, LucidChart23.

Results
Consent Observations
Of the 20 observed consenters, 19 (95%) were patients and one (5%) was a legal guardian. Participants at Site 1 included 10 patients with general mental health conditions and 7 providers. At Site 2, participants included 9 patients with SMI, 1 legal guardian of a patient with SMI, and 7 providers. In one case, other individuals (such as family, caretakers, etc.) who were not legal guardians accompanied patients during the session.

The average length of sessions at Site 1 was 10 minutes (range 5 - 20). An average of 3 forms (range 1-5), were discussed during the process. At Site 1, the readability of the forms ranged between grade level 10.6 to 12.1. At Site 2, the average length of session was 23.1 (range 3-36) minutes. An average of 6.1 (range 1-11) forms, were discussed. Readability of the consent forms at Site 2 ranged between grade level 7.0 and 15.0.

At both sites, observers identified a variety of consent forms covered during respective processes. Site 1 used electronic consents exclusively, while Site 2 providers used both paper and electronic consents based on patient preference. In all sessions, providers referenced their computer (desktops or laptops) screens to explain electronic consents. In 20% of sessions, providers referred also to paper-based materials to explain consents. Observers rated the explanations provided by the providers at both sites while categorizing the explanations into three categories: ‘Excellent’ (40%), ‘Good’ (40%) and ‘Poor’ (20%).

The number and type of consent forms discussed per session were based on the purpose of the appointment. Of the 10 appointments at Site 1, 9 (90%) appointments were for new patients, followed by 1 (10%) appointment for yearly reassessment. On average, new patients completed 2.6 (range 1-5) forms and the reassessment patient completed 3 forms. At Site 2, 5 (50%) appointments were for new patients and 5 (50%) for transfer patients. The new patients at Site 2 completed an average of 6.4 (range 2-10) forms, whereas the transfer patients completed 5.8 (range 1-11) forms.

Each site had unique consent forms. Table 2 provides a summary of the types of consent forms observed at both sites. Most of the forms, except forms 3a-3e, supported broad consent models. Forms 3a-3e provided consenters with granular consent opportunities to share or restrict specific health information (such as medications, assessments, diagnosis, substance abuse and HIV related information, treatment and discharge records and psychotherapy notes) and to direct the information to and from specific entities such as patients’ primary care provider (PCP) or other specialty providers, family members, living facilities (such as group homes) and any federal or state government agencies (e.g. Department of Child Safety or data protected under 42 CFR Part 2 regulations), as necessary.

Table 2. Types of consent forms observed at both sites

<table>
<thead>
<tr>
<th>#</th>
<th>Type of Consent</th>
<th>Purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Consent to Treatment</td>
<td>To request permission to provide the patient care</td>
</tr>
<tr>
<td>2</td>
<td>Advance Directive</td>
<td>To request patient to appoint someone else for making health care or mental health decisions in the event the patient becomes incapable of making decisions related to treatment</td>
</tr>
<tr>
<td>3a</td>
<td>Release of Information</td>
<td>a. Facility to PCP</td>
</tr>
<tr>
<td>3b</td>
<td></td>
<td>b. PCP to Facility</td>
</tr>
<tr>
<td>3c</td>
<td></td>
<td>c. To Family</td>
</tr>
<tr>
<td>3d</td>
<td></td>
<td>d. To Care Facility</td>
</tr>
<tr>
<td>3e</td>
<td></td>
<td>e. To Government Agencies</td>
</tr>
<tr>
<td>4</td>
<td>Communication via Email/ Voicemail</td>
<td>To request permission for provider to leave a voicemail or send email pertaining to treatment</td>
</tr>
</tbody>
</table>

In most sessions (75%), providers asked consenters if they had any questions. Overall, among all the sessions, only two (10%) of the consenters had any questions. One of the two consenters asked questions pertaining to what type of personal health information to share and with whom. No consenters had any questions about how consents would
affect their care. Further, observers noted the quality of provider responses to questions by the consenters. One response was rated ‘Excellent’ and one was rated ‘Good’. Only one of the two consenters who asked questions was found receptive to the provided explanations.

Observers rated the mood of the consenters and providers during the process. The response options for this section were: a) Angry/Frustrated, b) Sad, c) Uninterested/bored, d) Overwhelmed, e) Other. The researchers subjectively categorized the responses in category ‘Other’ as: ‘Interested’, ‘Happy’, ‘Depressed’, ‘Confused’, ‘Annoyed’, ‘Optimistic’, ‘Nervous’ and ‘Neutral’. The most common response for the mood of the consenters between both facilities was ‘Interested’ (40%), followed by ‘Neutral’ (20%) and ‘Happy’ (20%). Some of the less common responses were: ‘Uninterested’ (5%), ‘Depressed’ (5%), ‘Confused’ (5%) and ‘Annoyed’ (5%). When it comes to the providers’ overall mood, the common response was ‘Interested’ (45%) followed by ‘Happy’ (30%) and ‘Neutral’ (20%). One (5%) provider each was rated as ‘Annoyed’, ‘Optimistic’ or ‘Nervous’.

Lastly, observers noted the consent choices of the consenters. Most participants (90%) chose to consent and share all personal health information with one or more entities (for e.g. PCP, family, care facility, etc.). Two (10%) participants chose to provide consent to share some information (such as restrict medications from PCP and treatment records from family) and restrict certain information. No explanations or information were provided by the providers to the consenter regarding decision to restrict the health information and any consequences.

Issues with the consent processes were reported as free text. We identified seventeen free text notes related to five themes: a) education methods, b) consenter-provider interaction, c) consent interruptions, d) use of electronic consent technology, and e) use of paper-based consents. In 6 (35.3%) sessions, no visual guide was provided for consenters during explanation referencing screen. In 4 (23.5%) sessions, observers recorded lack of interaction such as eye contact between consenters and providers. There were also interruptions due to location-based noise and distractions by persons accompanying the patients (17.6%). In 3 (17.6%) sessions, problems with the function of the electronic consent technology were noted. Lastly, in 1 (5.9%) session, a provider left the meeting space to print consent documents requested by the patient.

**Discovered Workflow:**

![Generalized Consent Process Workflow](image)

**Fig. 1** Generalized Consent Process Workflow, from observations at behavioral health Sites 1 and 2. In color Yellow, workflow step sometimes observed at Site 1, in color Blue workflow step (printing the filled consent forms when patient/legal guardian prefers signing paper-based consents) only observed at Site 2.

The students observed ten consent processes at each site and created a generalized workflow to describe the steps involved in the observed consent process (Figure 1). Any observed differences in consent workflows between Sites were noted in the workflow. Below we provide a step-by-step description of the observed consent workflows:
Step 1: The consent process starts with provider meeting the patient/guardian and accompanying members, if any, in ‘Patient Waiting Areas’ and escorting them to the meeting space.

Step 2: In the meeting space, the provider chooses one consent form from a potential set of forms and verbally explains it with the aid of their computer systems, based on the type of appointment and requirements.

Step 3: The consenter is asked if s/he has questions about the specific consent form.

Step 3a: If the consenter has questions, the providers clarifies consenter’s concerns.

Step 4: The consenter makes consent choices. In the case of release of information forms (Forms 3a-3e from Table 2), the consenter can choose whether to provide consent to share ‘all, some or none’ of their information and with whom.

Step 5: The provider either requests the consenter to provide an electronic signature or requests to sign the paper copies of the consent form (only observed at Site 2).

The described process (steps 2-5) is repeated for each consent form, based on the set of consent forms to be discussed.

Step 6 and 6a: Consenter is offered and provided a printed copy of the signed consent form/s, as per their preference (sometimes at Site 1).

Site 1 consenting sessions were performed before the therapeutic treatment sessions, instead of after the session, to better accommodate student observations. This modification could have introduced changes in the workflow and consent session length.

Discussion

Literature describing the use of electronic consent technologies is increasing. Much of the research has focused on consent for clinical research and patient recruitment. These have identified increased flexibility and interest, greater comprehension and patient engagement, reduced provider workload and higher satisfaction among patients and providers as potential benefits of electronic consent. As well, the transition from paper consent forms to tablet or web-based consent tools is noted to be feasible, effective and preferred by users. However, we identified no studies that analyzed the consent process in the routine ambulatory environment using electronic consent tools.

Informed consent for patients with behavioral health conditions is complicated by comprehension and decision-making issues. Disorders and treatments that impair cognitive and intellectual functioning interfere with the comprehension and retention needed to make informed decisions. Similarly, certain conditions may produce declines in intellectual processes, compromised judgement and insight impairment. These factors can contribute to increased cognitive load and result in confusion for some individuals, especially those with behavioral health conditions.

The lack of preexisting studies on consent processes and the effect of electronic consent tools on behavioral health patient and surrogate engagement convinced us of the importance of this research. It is particularly important for behavioral health patients that the consent process is understandable and low burden, but our study findings suggest that current processes differ from these expectations. Below we summarize our observations during the consent process at two outpatient behavioral health clinics, Site 1 and Site 2:

The consent forms offer granular choice: Both study sites currently offer patients a certain level of granular choices to share/restrict health information (Table 1, forms 3a-3e), specifically medications, assessments, diagnosis, substance abuse and HIV related information, treatment and discharge records and psychotherapy notes. This is consistent with the recommendation from the Office for the National Coordinator for Health Information Technology and with patients’ desires for more control over data sharing and in compliance with regulations.

The consent process is burdensome and time consuming: The consent process duration ranged from 3 to 36 minutes between both sites. Consenters completed multiple forms electronically (desktop or laptop) or paper-based forms during the process, with average of 5 forms and 2.25 pages of written information. In the worst case, one consenter was asked to complete 11 forms; this process consumed 34 minutes during which 15 pages were discussed. The length of the consent process and number of forms covered, in part reflect the distinct patient populations (general mental health vs. SMI) they serve. However, within each site we observed variation in the process based primarily on the type of visit, i.e. new or established and the nature of the presenting problem. Such variation reflects the necessity of personalizing the consent process for each patient within the structure of a standardized protocol and process.

The reading level of consent documents is high: An average American reads at an eighth-grade level. This was affirmed by Doak and Doak who studied the average reading level of patients in a public hospital and found that despite reporting high school education, most patients read at a seventh-grade level. A 2007 review of 154 informed consent forms from mental health studies in Massachusetts that found the consents were written at an average of 12-14.5 grade level. This is consistent with our study where the mean reading level of consent documents ranged from
grade 7.0 to 15.0. To accommodate patients with low reading levels or low health literacy, Paasche-Orlow and Taylor suggest the target reading levels of grade 4 to 6\(^2\).

The terminology and layout of consent forms are variable: Study Sites 1 and 2 use consent forms which serve similar purpose but are titled differently. For example, the Release of Information form is known as ‘Consent for Disclosure of Confidential Information’ at Site 1 and ‘Consent to Release Protected Health Information (PHI)’ at Site 2. There also appears to be an effort towards consolidation of some related forms, as exhibited in Site 2 (Table 2, forms 3a-3e and 4). An additional issue results from different titles for a similar process. Site 1 refers to the consent process as ‘Intake’ while Site 2, calls this segment the ‘Welcome Packet’. Such variations may cause confusion for the patient, especially as they visit other medical facilities and execute similar documents.

The physical space was not organized optimally: Effective patient-provider communication and shared decision making can improve the efficacy of the informed consent process\(^29\). However, we observed that the seating arrangement and the location of the consent technologies (i.e., desktop computers and signature pads) could be improved. Observers noted lack of eye contact and the absence of visual aids. Combined, these issues in setup can interfere with the consent interaction, affect consentor comprehension and may lead to a lack of shared decision-making between patient and provider.

Though the sample size was small and the observations limited to two clinics, we offer these corresponding consent process areas for improvement:

**Standardize the consent processes, forms and terminology, where possible:** With the involvement of the behavioral health facilities, attempts could be made to understand if some of the observed processes, forms and terminology could be standardized to reduce variability. We do acknowledge that the differences in patients’ demographics (general behavioral health vs. SMI), providers’ roles and relationships with patients, and types of services could impose constraints in the standardization process.

**Combine consent forms:** As feasible, multiple consent forms should be combined (as it was observed at Site 2, when ‘Release of Information’ and ‘Consent for Email/Voicemail Communication’ forms were consolidated), to streamline the consent process.

**Reduce reading levels and length of consent forms:** Embedded consent forms that are written for sixth-grade level and content divided into easily digestible thematic sections might help to increase patient focus, understanding and reduce cognitive load.

**Address challenges in provider-consenter interaction related to physical space and technology arrangements:** As consents are embedded in the EHR, providers take an active role while consenters are generally passive during the consent process. Electronic consent tools supporting multiplatform interfaces could allow providers and consenters to interact with the tool on any stationary (such as desktop) or mobile (such as tablets or smartphones) device. Potential use of handheld mobile devices can provide mobility to the providers in the meeting space, foster improved provider-consenter interaction and allow consenters to take an active role in the process.

**Incorporate multimedia educational material:** Multimedia education use has shown promise among physical and behavioral health patients with increased understanding, comprehension and interest towards treatment and care\(^29\). Embedding on-demand educational material relevant to consents, health data sharing and related consequences into the electronic consent forms would enhance direct patient engagement and involvement in the process.

**Personalize consenter experience:** Consent tools could be personalized based on the type of patient (general behavioral health diagnosis vs. SMI), patient appointment type (such as new, referral.transfer, reassessment), preferred language (English or Spanish) and past user experience (such as providing instructions when using the tool for first time). Personalized experiences and flexibility to access the tool virtually (inside or outside care facilities) could encourage patient engagement and help reduce the complexity and time to complete the consent process.

In terms of methodological improvements, the free text notes from observers helped to identify a need to amend the Observation Outline (Table 1) to reflect some consistent observations, including the lack of visual guides for patients, lack of eye contact between patient and provider and interruptions during the process. Revisions to the Observation Outline would include questions related to quality of the interaction, such as eye contact, conversation, etc., number and nature of interruptions such as environmental noise, disturbances by accompanying member or other staff, etc., technology arrangement and physical space, such as sitting arrangements, direction of computer screens, etc. Furthermore, observers noted challenges related to the use and absence of electronic technologies, such as issues with signature pads or absence of printers in provider meeting space. Observers often subjectively noted consenter or provider moods as ‘Interested’ or ‘Happy’. Certain positive moods can be accommodated in the revised outline. Such revisions could enhance future data acquisition using the proposed Observation Outline instrument.
An important limitation of our study is the number of observations. We observed ten consent processes for each site and included only one legal guardian. Also, each site has multiple providers responsible for the consent execution, but we observed only a limited number of providers. Another limitation is the lack of diversity within the observed population. Only English-speaking consenters and providers were included in the study. Differences in culture, lack of familiarity with the U.S. health care system or the use of translators could greatly affect the dynamics of the consent process. Similar studies should be conducted on a larger and more diverse sample of patients with behavioral health conditions, surrogates, and providers. Companion studies in the physical health environment will also be useful as technology permits greater data granularity.

Conclusions and Future Work
Informed consent establishes ethical safeguards for care and research and protects patient rights. Such protections are particularly important for individuals being treated for behavioral health conditions. Based on our observations of consent processes at two community behavioral health clinics, we identified several challenges and highlighted areas for improvement. These include standardizing the consent protocol and process; applying quality improvement techniques to minimize variation; redesigning consent artifacts to a reading level of grade six; re-imagining the physical space to include technology hardware; and upgrading the electronic consent software. More research is needed to better understand how electronic consent technologies and educational support will impact the informed consent practices and preferences of patients with behavioral health conditions, their surrogates and providers.

The outcomes from this observational study will be supplemented with semi-structured provider interviews and a survey of patients with behavioral health conditions and legal guardians to guide the development of electronic consent tools, such as My Data Choices.

Acknowledgements
This research was funded by the National Institute of Mental Health (NIMH) through My Data Choices, evaluation of effective consent strategies for patients with behavioral health conditions (1 R01 MH108992-01A) grant.

We thank our collaborating site members, Dr. Veronica Sabater and Byron Hoston, for their support in data collection.

References


Modeling Contextual Knowledge for Clinical Decision Support

Margarita Sordo, MSc, PhD;1,2,3 Priyaranjan Tokachichu, MD, MS;1 Christopher J. Vitale, PharmD;1 Saverio M. Maviglia, MD, MSc;1,2,3 Roberto A. Rocha, MD, PhD1,2,3

1Clinical Informatics, Partners eCare, Partners Healthcare System, 2Brigham and Women’s Hospital, 3Harvard Medical School, Boston, MA

Abstract

In theory, the logic of decision rules should be atomic. In practice, this is not always possible; initially simple logic statements tend to be overloaded with additional conditions restricting the scope of such rules. By doing so, the original logic soon becomes encumbered with contextual knowledge. Contextual knowledge is re-usable on its own and could be modeled separately from the logic of a rule without losing the intended functionality. We model constraints to explicitly define the context where knowledge of decision rules is actionable. We borrowed concepts from Semantic Web, Complex Adaptive Systems, and Contextual Reasoning. The proposed approach provides the means for identifying and modeling contextual knowledge in a simple, sound manner. The methodology presented herein facilitates rule authoring, fosters consistency in rules implementation and maintenance; facilitates developing authoritative knowledge repositories to promote quality, safety and efficacy of healthcare; and paves the road for future work in knowledge discovery.

Introduction

One well established best practice for designing rules for decision support systems (DSS) is that “decision rules should be atomic”.1 Although in theory this is widely agreed upon, in practice this is not always possible. What normally starts as a simple logic statement in the antecedent of a decision rule can develop into what looks as a procedural piece of code when additional conditions are added to the logic in order to restrict the overall rule to a specific context. These additional constraints may very well contain relevant information and are perfectly valid, yet from a knowledge representation perspective, they could be considered contextual knowledge that should be independent from the logic of the rule. For example, a decision rule that alerts for an abnormal laboratory test result should specify the laboratory test result in question, a comparison operator and a threshold value. However, the logic of such rule also tends to include conditions for age and gender, among other patient characteristics. This results in having a single rule with a long conditional expression, or multiple variations for the same rule to account for different age groups and genders. This approach creates rules that are less reusable, with large numbers of pre-coordinated logic expressions that are difficult to maintain.

Consistent with previous work2 we believe that decision rules could be modeled as adaptive agents capable of exhibiting specific behavior in response to their environment,3 where such environment is defined by contextual constraints. This approach allows us to identify and separate contextual knowledge from clinical knowledge and still preserve the intended functionality of decision rules. Furthermore, by clearly separating context from the logic of the rule, we not only facilitate rule implementation and maintenance, but also potentially promote knowledge discovery. The proposed approach to modeling contextual knowledge for decision rules is based on Ontologies, with some basic concepts borrowed from Complex Adaptive Systems and Contextual Reasoning. The work reported herein is part of ongoing efforts to further an authoritative knowledge repository to promote quality, safety and efficacy of healthcare.

Ontologies

An ontology is a conceptualized and agreed upon collection of entities, relations and instances in a domain of interest where all elements are unambiguously described by means of a declarative language and a shared vocabulary.4 From a modeling perspective, ontologies are of particular interest for this project as they align with ongoing efforts for extraction, modeling and curation of expert knowledge currently embedded in a variety of systems across Partners Healthcare.4 A large amount of this expert knowledge is in the form of decision rules, and collecting this knowledge from within a variety of disparate systems into a single platform optimized for curation is a crucial strategic initiative at Partners.10
Complex Adaptive Systems

Holland describes Complex Adaptive Systems (cas) as a large number of simple agents, each exhibiting their own behavior in response to external stimuli. There are seven basics (four properties and three mechanisms) central to all cas. Our previous work incorporated the four basics (aggregation, tagging, building blocks, and internal models) that we believed best contributed to modeling rule interactions. From those four basics, we further explore two mechanisms: a) Tagging mechanism facilitates the formation of aggregates and delimitation of boundaries, the latter being a key feature for describing behavior within a context. b) Building Blocks mechanism allows reusability and building complex things from repetitions and combinations of simple ones. Our modeling efforts align with both mechanisms in that by tagging a decision rule we successfully remove contextual information from the antecedent of the rule, and explicitly set the boundaries of the environment where such a rule applies. As a result, simplifying the logic of the rule makes it more atomic and more reusable.

Contextual Reasoning

Reasoning about context mainly arises from the problem of locality, namely, knowledge that only applies in a restricted world. Such a restricted application might be represented by a set of values that explicitly delimits the boundaries of a world where knowledge is valid. For example, the statement “It’s raining” is true only in the context of a given location and time. This is the contextual information we need to explicitly define a more concise representation of the world. There is no need to define all possible aspects of the world that may determine a particular context, only those relevant to a given circumstance. Using the above example, we do not need to state whether we are wearing a raincoat or not, or the color of the raincoat for that matter, unless, for some reason such information is relevant. As a result, a partial representation — that only includes location and time — suffices for the purposes of setting a context that allows us to determine the validity of the statement “It’s raining.”

For our modeling purposes we aim at characterizing such a world by borrowing concepts from knowledge representation and reasoning for Ontologies and Semantic Web technologies. We specifically base our approach on the metaphor of context as a box. This metaphor defines contextual metadata in terms of three basic elements: a set of parameters $P_i$; a collection of expressions representing the domain at hand; and three abstract forms of reasoning: expand/contract, push/pop, and shifting, each corresponding to an operation on one of the basic elements of the representation, i.e. parameter, value, and expression. Each parameter is seen as a dimension of a box (context), which in combination with a value indicate the position of the specific context within a multi-dimensional space. Hence delimiting what is “inside and outside the box.” The remaining of this section describes in more detail these contextual reasoning operations.

**Expand/Contract:** operations are based on the intuition that an explicit representation associated with a given context only contains a partial subset of facts — information inside the box. As a result, such subset could be expanded as more pertinent information is available, or contracted if information no longer relevant is removed.

**Push/Pop:** $Push$ operation removes information from inside the box and adds it as a parameter value outside the box; conversely, $pop$ removes a parameter value and adds it as information inside the box. For example, if our original statement about the weather were “It’s raining in Boston” then we would only have time as a context parameter (the location information is inside the box). However, we may choose to remove “Boston” from our statement and push it outside the box as location information. These operations have a direct effect on both the number of parameters and their values, while delimiting the box and the information inside it.

**Shifting:** operator allows us to move from one representation to another by changing the value of one or more contextual parameters, as long as the relationship between the parameter’s values and the statements inside the box is known. A simple example would be to state “Today is raining” in the context where the date/time parameter is set to today’s date, and then having the statement “Yesterday it was raining” and shifting the value of the date/time parameter to tomorrow’s date. We can see that both statements refer to the same fact, though the context was shifted.

These operations provide the mechanisms for handling the fundamental aspects of the proposed contextual representation: a) partiality, namely the portion of the world being represented; b) approximation or level of detail by which the portion of the world is being depicted and; c) the point of view or perspective from which the world is being observed. Authors showed that “at a suitable level of abstraction, a logic of contextual reasoning is precisely a logic of the relationships between partial, approximate and perspectival theories of the world.” In other words, from all possible representations of the world, we may choose a (partial) (approximation) of what we consider relevant to our purposes, and needs (perspective).
We believe that the context as a box metaphor is consistent with the Tagging mechanism for cas in that both approaches specify a context and delimit the scope of decision rules. Our only departure from the context as a box metaphor is that we do not allow any overlapping boxes (contexts); all boxes must be mutually exclusive, since any overlap may lead to ambiguity, and its inherent difficulties (e.g. limitations of existing rule engines to deal with ambiguity). In the following sections we will further expand on this notion of decision rules as “agents interacting inside a box.”

Furthermore, by applying these concepts to if...then decision rules we are able to explicitly circumscribe the context where actions specified in the consequent of a rule are executed when the conditions in the antecedent are satisfied.

Modeling Contextual Knowledge as Constraints

Capturing context is critical for understanding and handling knowledge. This is particularly true in a clinical setting where knowledge embedded in decision rules often times is tailored to specific settings and scenarios. However, it is also most desirable to preserve the generality of rules, ensuring a high degree of reusability and maintainability.

![Figure 1. Simplified Schema for Decision Rules.](image)

The schema for decision rules in Figure 1 consists of a) Generic Properties: Provenance, and Constraints for the overall content of the rule, e.g. a decision rule may apply to patients of all ages, and the constraint for AgeGroup would be set to “ANY”, or to both genders, and so the Gender constraint will be set to “ANY”; and b) Type Specific Properties, which model the rule expression (described in Sordo et al.) and represent the rule itself and its behavior. In other words, the “hasConstraint” generic property defines the full context where a rule applies (e.g. “big box”), whereas the type-specific “hasConstraint” defines more restrictive contexts (e.g. “small boxes”) and behavior of the rule. It is worth noting that all restrictive contexts (“small boxes”) must be contained in the full context (“big box”). For example, we could define the generic “hasConstraint” AgeGroup to “Adults”, and then define more specific ones as “Young Adult”, “Mature Adult” to restrict the scope. We currently focus on the type-specific property hasConstraint, and we will further describe it in the following section.

Type Specific Property “hasConstraint”

Rule execution can be constrained to narrower scopes by restricting it to more specific contexts. As long as we keep this in mind, we can define as many “sub contexts”, or boxes as needed for a single rule. For example, an alert for an abnormal laboratory test result may be relevant to all patients regardless of gender and age, and so the constraints in the generic properties of the rule should be set to Gender=“ANY”, AgeGroup=“ANY”. The “ANY” value means that such dimension is unrestricted. However, threshold value(s) for the rule may depend on the age of the patient. Therefore, by specifying such threshold values constrained by age groups in the data definition of the rule expression (type specific property), we can still model an alerting rule as simply as if LabResult <operator> threshold then alert; where the values assigned to threshold are constrained by the context (AgeGroup) where such values apply. Hence, the logic is the same, but the content (operator and threshold values) is defined by the context; in other words, the content – what is inside the box - is dependent on the parameters and values outside the box. This is consistent with the context as a box metaphor. Such metaphor and the operations presented in a previous section lay the foundation for our work and allow us to extend/restrict the scope of such rules.

Clinically-Relevant Constraints

In previous research authors identified clinically-relevant constraints that should be considered when modeling clinical scenarios. Such constraints are represented by three main context dimensions: a) Patient, further subdivided into age, gender, clinical condition, clinical protocol/trial, and health insurance plan; b) Provider-related including: group, role, and clinical specialty; and c) Care Setting consisting of care setting (inpatient/outpatient), geographic region, facility, department, unit, unit type, room, and bed. From these dimensions, we defined a partial approximation by choosing the following four: Patient Gender and Age Group, Care Setting Site, and Venue; and
abstracted away those dimensions that currently do not advance our purposes. However, as our clinical content evolves, we may choose to further expand such partial representation and include additional dimensions. We modeled the AgeGroup dimension as non-overlapping three subcategories: “Prenatal,” “Pediatric,” and “Adult,” with each subcategory potentially containing sub-sub categories. For example, “Adult” may contain “Young Adult,” “Middle Age Adult,” “Mature Adult,” etc. Patient Gender is modeled as “Male” and “Female.” Care Setting Site currently includes a limited number of Partners Healthcare hospitals: “Brigham and Women’s Hospital” (BWH), “Massachusetts General Hospital” (MGH) and, “Newton-Wellesley Hospital” (NW). Venue includes “Inpatient,” “Outpatient,” “Intensive Care Unit” (ICU), and “Emergency Department” (ED). All dimensions support a value of “ANY” to denote that a dimension is explicitly included, but it is left unrestricted. In other words, the scope of a dimension with a value of “ANY” includes all possible values available in such dimension. For example, an AgeGroup dimension with a value of “ANY” would include all three age groups, and be satisfied by any of these values.

These four dimensions define a box into which we place our decision rules. Values of “ANY” to specify the scope of a dimension allow us to always include all four dimensions, even if not restricted. This has the advantage of providing a fully explicit depiction of the dimensions of our box. As shown in Figure 2a, an unrestricted decision rule is placed in a box with all four dimensions set to a value of “ANY”. This means that this hypothetical rule inside the box will apply to any patient regardless of age or gender, at any site or venue. For example, if such rule were to measure potassium levels, Figure 2a would be specifying that such rule would apply at any Site, any Venue, and to any patient regardless of Gender or Age.

![Figure 2](image)

**Figure. 2.** (a) A generic decision rule inside an unrestricted box with dimensions unrestricted with an assigned value of ANY. (b) Rule with context-specific threshold value for adult inpatient females at BWH. (c) Rule with threshold value for adult inpatient males at BWH.

![Figure 3](image)

**Figure. 3.** Mutually exclusive “boxes” stacked to delimit multiple scopes of a single decision rule with context-specific threshold values.

However, if we chose to define a specific threshold value, e.g. 40 mg for adult inpatient women at the BWH, then instead of adding such constraints to the logic of the rule and changing the threshold value to 40 mg, we push the constraint values into the parameters (dimensions) outside the box and expand (add) the context-specific threshold value inside the box (Figure. 2b). This is equivalent to having both the rule and the threshold value dependent on the context. Additional context-dependent threshold values may exist for a given rule. This would be equivalent to “stacking” boxes. For example, we may define an additional threshold value of 30 mg for adult males for the same site and venue for the above rule. We shift the value for the Gender dimension and change the threshold value accordingly (Figure. 2c). This will result in two boxes for the same rule, each having its own context-delimited threshold.
For the sake of graphically representing this notion of “stacking boxes”, let us momentarily abstract away (remove) the Venue dimension, and draw a 3D representation (Figure 3). These two “boxes” are stacked side-by-side to cover the Gender dimension of the context box. Such (broader) Gender “box” is divided into two, more restrictive, non-overlapping boxes, each box describing a partial representation (portion) of the world, with its own threshold value.

We believe that by having fully circumscribed, mutually exclusive boxes representing partial views of the world we can build a robust contextual representation (consistent with cas internal models) of clinical decision rules that can be aggregated and reasoned upon. Similar to the building blocks mechanism of cas, the rationale behind the context as a box approach is to model simple, well-circumscribed behaviors for rules and, then, aggregate them to model more complex behaviors. In the remaining of this paper, we present some exploratory examples and discuss findings and limitations.

Results

We focused our initial analysis on the decision rules used by the Results Manager (RM) computerized application at Partners Healthcare. RM is an application in the outpatient setting consisting of 84 rules that enable clinicians to review, acknowledge, and act upon abnormal/critical results of chemistry, hematology, toxicology, radiology, and cytology tests in a timely manner.

The generic representation of RM alerting rules is as follows: If LabResult <Operator> <threshold> then Alert; where LabResult is the value of a given laboratory test result, e.g. “Potassium” (chemistry), “Acetaminophen” (toxicology), “INR” (hematology); Operator is a comparison operator to determine whether the laboratory test result is normal/abnormal when compared to a threshold value. Threshold value is a reference value for the laboratory test based on values found in a particular population cohort. Rules were enhanced to include variations on reference threshold values based on gender, age, and health status of the patient, resulting in a repository of 166 rules. For example, a patient in the ICU or with a chronic condition most likely has some abnormal laboratory test results. This is why, besides age and gender, venue is important to determine the threshold value of a test. We propose the following approach: starting with the general representation of the rule, with all dimensions set to “ANY”, we push the constraint on the venue dimension outside the box, and expand (add) the context-specific threshold value inside the box. Therefore, in this scenario, the venue would be set to “ICU”, and the threshold value for patients in critical condition would be added inside the box, with the desired effect of having a decision rule with a threshold value targeted to a specific context, that will only trigger an alert under these more delimited circumstances.

Some chemistry rules may need to be restricted to age ranges within an AgeGroup. For example, for patients with some suspected thyroid gland dysfunction, a thyrotropin (TSH) chemistry test might be ordered to check the endocrine function of the gland. TSH reference values are age-related, with specific reference values for Pediatric and Adult populations, with further subgroups in the Pediatric population with age ranges of [1 day, 4 days), [4 days, 8 days), and [8 days, 18 years). The initial generic representation of this rule is depicted in Figure 4a. In Figure 4b, we push the Adult AgeGroup outside the box and expand the operator and threshold value. In Figures 4c-4e, we shift the AgeGroup to Pediatric (PEDI) and include the additional age range, and replace (contract and expand) the facts inside the box for threshold and operator.
Figure 4. An example of pushing, expanding and shifting parameters for a single rule in multiple contexts.

Discussion

We have demonstrated the feasibility of the modeling strategy proposed above by implementing relevant test scenarios through our analysis of the clinical decision rules in the aforementioned application. Currently, our model supports non-overlapping, mutually exclusive contexts consistent with most our current decision rules. However, we anticipate the need for more flexible rules that could interact in environments where such clear boundaries do not apply. We have incorporated a “degree of membership” to our model to deal with potential overlapping “boxes”, though at the time of writing, we have not evaluated it.

We have shown that with a relatively simple approach we can model abstract contextual knowledge from a variety of rules while preserving the desired functionality. We found that by applying contextual reasoning operators we could expand/contract the context where decision rules apply, while keeping the rule representation as atomic as possible. Further, given that both contextual and clinical knowledge may lie on the same continuum with no fixed point separating them, having the flexibility to vary the degree of approximation of a given representation allows us to regulate the interplay between what goes inside (clinical knowledge) and outside the box (contextual knowledge). By doing so, decision rules can easily be adapted to new contexts, without encumbering the rule logic with extra conditions. We believe this approach is consistent with the internal models and building blocks concepts of complex adaptive systems, and it will allow us to model more complex behavior for rules while preserving a simple and sound representation. Likewise, from an ontology perspective, the same mechanisms can be applied to represent both contextual and clinical knowledge as “conceptual building blocks” that can be reasoned upon while preserving the correctness and expressiveness of the underlying ontology.

Conclusions

Sound and comprehensive approaches are key to accurately modeling knowledge content of any type. In the case of clinical decision rules, it is highly desirable to capture and model not only knowledge pertaining the logic and actions of such rules, but also the context where such knowledge becomes relevant and actionable. Currently, contextual knowledge is not identified as such at modeling time, and we strongly believe that it should be. Such contextual knowledge should be removed from the antecedent of a rule, so the logic remains as atomic as possible. The proposed approach provides the means for identifying and modeling contextual knowledge in a simple, yet sound manner. Furthermore, the methodology presented herein facilitates rule authoring, fosters consistency in rule implementation and maintenance; facilitates developing authoritative knowledge repositories to promote quality, safety and efficacy of healthcare; and paves the road for future work in knowledge discovery.
Acknowledgments

We would like to thank Beatriz H. Rocha for her valuable expertise and insights that greatly assisted the current research. We also thank Elisa Dell’Oglio, Amanda Fairbanks, Teal Aroy, David Dubois, Karen Bavuso, Sharon Bouyer-Ferullo, and Dan Bogaty for their contributions.

References

Reducing the Toxicity Risk in Antibiotic Prescriptions by Combining Ontologies with a Multiple Criteria Decision Model

Souhir Ben Souissi, MSc1,2,3, Mourad Abed, Prof1, Lahcen Elhiki, Dr2, Philippe Fortemps, Prof3, Marc Pirlot, Prof3

1LAMIH, University of Valenciennes and Hainaut-Cambrésis, Valenciennes, France; 2Research Institute for the Science and Management of Risks, and 3Department of Mathematics and Operational Research, University of Mons, Mons, Belgium

Abstract

We consider the risk of adverse drug events caused by antibiotic prescriptions. Antibiotics are the second most common cause of drug related adverse events and one of the most common classes of drugs associated with medical malpractice claims. To cope with this serious issue, physicians rely on guidelines, especially in the context of hospital prescriptions. Unfortunately such guidelines do not offer sufficient support to solve the problem of adverse events. To cope with these issues our work proposes a clinical decision support system based on expert medical knowledge, which combines semantic technologies with multiple criteria decision models. Our model links and assesses the adequacy of each treatment through the toxicity risk of side effects, in order to provide and explain to physicians a sorted list of possible antibiotics. We illustrate our approach through carefully selected case studies in collaboration with the EpiCURA Hospital Center in Belgium.

Introduction

Adverse drug events are one of the most important causes of mortality in the healthcare context. They claim each year between 700,000 and 1.5 million casualties in the United States1. From these, antibiotics are the second most common cause of drug related adverse events2,3,4 and one of the most common classes of drugs associated with medical malpractice claims5. Every antibiotic treatment is associated with a risk rate, i.e., an assessment of the degree of risk, which takes into account both the severity and the frequency of undesirable side effects. Determining the best treatments is highly dependent on the considered acceptance risk. The acceptance risk of a particular treatment varies according to the vulnerability of the patient. Therefore, physicians are in need of support that can give them both an overview of the situation, but also a detailed description of the risk and the vulnerability of each patient.

For many hospitals, guidelines constitute the solution to guide antibiotic prescriptions, by linking infection diagnosis to the relevant antibiotic therapies. As an example, our collaborators in the EpiCURA Hospital Center have been using a hundred-page guideline since 20116. Unfortunately, there is evidence showing that little or nothing has really changed regarding the malpractice issue in spite of the existence of guidelines. The main causes of this situation are that a) the physician cannot have a global view of the guidelines or complete explanations for the suggested treatment; b) in their current textual form these guidelines are completely static, making them hard to use and adapt to either the specific needs of the patient or the changing environment context (e.g. usage in emergency rooms).

Using decision support systems (DSS) in a clinical context is mainly about providing recommendations to physicians. Some of these Clinical DSS are using semantic technologies. From this perspective, Bright and coauthors7 have developed a formal ontology to structure the empiric antibiotic therapy guidelines of the New-York Presbyterian Hospital (NYP). The guidelines in this study have been explicitly entered in Protégé. This work was able to generate three kinds of prescribing alerts when the guidelines were not respected. Despite its advantages, this approach has serious limitations. Its main drawback is that both the basic data and the relationships between them should be explicitly entered in the system (i.e. the system cannot generalize to new data), making maintainability difficult, if not impossible. Other funded projects around semantic technologies for medical procedures such as REMINE8 and PSIP9 use data mining techniques to reduce drug adverse effects by taking into account the patient’s medical record.

Despite these efforts there is currently no widely accepted standardized framework that will help physicians in their day to day prescriptions needs, although some researchers have tried to move towards this direction10,11. The broadest approach has been taken by Doulaverakis and coauthors12,13 to cover drug to drug interactions and drug to diseases interaction but, apparently, with no consideration of patient vulnerability and the risk of the drug side effects. Other
Clinical DSS use Multiple Criteria Decision Aiding (MCDA). For example, the AHP (Analytic Hierarchy Process) method has been used to analyze and compare the treatment options for eradication of H. pylori infection in children. The main limit of this work is the lack of structured objective knowledge. The recommendation is highly sensitive to the stakeholders evaluations and weights.

Clinical DSS have another important task which is to provide physicians with explanations for recommendations. In a previous work, we combined semantic technologies with MCDA to sort alternatives (antibiotics) by their suitability to the subject (patient) but we did not consider the toxicity risk as is done by the physicians (assessing the risk level by combining severity and frequency).

Our objective in this work is to build a model for a clinical DSS that links knowledge structures for assessing the antibiotic prescription risk. The method we propose integrates MCDA with ontologies in order to provide to physicians sorted list of antibiotics assessed according to their toxicity risk for a given patient with an infection disease.

We use ontologically structured knowledge about the pharmacological characteristics of antibiotics and an ontology describing the critical criteria of patients. These ontologies are then linked through a set of rules structured in a Majority Rule Sorting model (MR Sort) with Veto. This model, which is a simplified version of ELECTRE TRI, assesses the alternatives by sorting them into ordered categories. This process results in antibiotic prescription recommendations categorized according to the risk of their side effect toxicity. We model the relations between concepts by simple interpretable rules, involving a small number parameters, to guarantee generality and maintainability of the knowledge model.

The main contributions of this paper are:

- A new model for integrating MCDA with ontologies for clinical Decision Support Systems
- An adapted version of the ELECTRE TRI – Majority Rule Sorting model (MR Sort) with Veto that is tailored to prescription recommendation with explanation.
- An experimental validation of the above model for categorization of antibiotics through risk assessment of side-effect toxicity

The rest of the paper is organized as follows: The next section (“The system”) describes our adaptation of the ELECTRE TRI – MR Sort rule, as well as our model for integrating the method with the systematic representation of patient profiles and antibiotic knowledge in ontological structures. Section “Illustration” details our validation of antibiotics categorization through the risk of side-effect toxicity. Finally in the “Conclusion” section, we present further research perspectives.

The system

The Semantic Model Our solution models a patient P that hosts pathogens which cause a bacterial infection. This patient goes to the hospital to seek medical care and requires a treatment by antibiotics. The first step of the decision support model for antibiotics prescription is to gather knowledge.

For this purpose we use an ontology O_P that models the patient and his/her relevant characteristics. These involve gender, age, comorbidities, allergies, and all the necessary patient information in order to assess the efficiency and the hazard of an antibiotic. All these characteristics influence the choice of an antibiotic in a way which is specified in reasoning rules. Indeed, a given antibiotic could suit a pregnant woman but not an old diabetic man, or conversely.

The variable \( SP_j \rightarrow \{0, 1, 2\} \) represents the sensitivity indication of patient to the side effect \( S_j \). A value 0 means “no sensitivity”, a value 1 means “minor sensitivity”, while 2 represents “major sensitivity”.

Our second ontology \( O_A \) provides us with the set of side effects \( S_j \) of a given antibiotic \( A_i \).

The variable \( RSA_{ij} \) is the risk of side effect \( S_j \) caused by antibiotic \( A_i \). It represents the relation between an antibiotic \( A_i \) and a side effect \( S_j \) with indication of harmfulness. \( RSA_{ij} \) combines the intrinsic (i.e., independent of the anti-
The values of the severity $GS_j$ are presented in Table 1. Value 0 indicates that side effect $S_j$ is not severe. Value 1 indicates that the side effect $S_j$ is severe. Similarly, value 3 corresponds to a harmful side effect. The frequency values $FSA_{ij}$ are presented in Table 2. Value 0 represents a less than $1/10,000$ chance that antibiotic $A_i$ has side effect $S_j$. Similarly, values 1 and 2 represent the chance being between $1/100$ and more than $1/100$, respectively.

Currently we deal with a total of 45 side effects organized in 8 categories for 60 antibiotics and 22 patient criteria.

**The MR Sort model with Veto for assigning objects to ordered categories** We use the MR Sort (Majority Rule Sorting) model with Veto (18,19,21) to classify the antibiotics for a given patient in ordered categories. MR Sort with Veto is a variant of ELECTRE TRI, which is a decision model belonging to the family of outranking methods (22, 23). The goal of MR Sort (and ELECTRE TRI) (24) is to sort alternatives in ordered categories based on their performance on several criteria. In our context, the alternatives are antibiotics $A_i$ and the criteria are the side effects $S_j$. The performance of an antibiotic w.r.t. a side effect for a given patient is determined by three elements: the severity $GS_j$, the frequency $FSA_{ij}$, and the sensitivity of the patient to the side effect $SP_j$.

Each category $C_k, k = 1, \ldots, K$, is associated a lower profile $t(C_k)$, which is a vector of levels on each criterion representing the minimal requirements to belong to category $C_k$. Actually, not all these requirements are to be fulfilled, but only a sufficiently large majority of them, in order for an alternative to be assigned to this category. The majority condition can be implemented by assigning weights to the criteria and selecting a majority threshold; the condition is fulfilled whenever the sum of the weights of the criteria on which the alternative is at least as good as the profile passes the majority threshold. A simpler version of this rule consists of counting the criteria on which the alternative is at least as good as the profile; the condition is fulfilled whenever this number passes a minimal fixed number of criteria (This is tantamount to assigning equal weights to all criteria). An additional condition for being assigned into category $C_k$ is that none of the alternative performances is “unacceptably bad”. Unacceptably bad performances are determined by veto thresholds on each criterion. Whenever the performance of an alternative on some criterion is worse than the corresponding profile performance by some specified quantity, a veto is activated which precludes assignment into the category.
To sum up, the principle implemented in MR-Sort with Veto in order to assign alternatives into categories is the following: An alternative is assigned to category $C_k$ or better (i.e. $C_{k+1}$ up to $C_K$) if the performances of the alternative are at least as good as those of the profile $t(C_k)$ on a majority of criteria and none of these performances falls below the veto threshold on each criterion.

In this work, we chose to use MR Sort with Veto for two reasons. On the one hand, it gives us structured rules to link the antibiotic knowledge structure to that of the patient. On the other hand, it provides us with an assessment of the suitability of the antibiotics to the patient. In addition, our method assigns the antibiotics (alternatives) in three categories ($K = 3$): $R$ (“recommended”), $P$ (“possible”) and $TBA$ (“to be avoided”), according to their toxicity risk.

The MR Sort with Veto assignment principles are applied to the context of antibiotic prescription, yielding the following rules:

- An antibiotic is assigned to category $R$ for a given patient if it has only a small number of side effects that the patient is sensitive to and if there is no unacceptable side effect (risk) for the same patient (no veto against $R$).
- A similar rule applies for an antibiotic being assigned to category $P$. The number of side effects tolerated is higher than those for category $R$ and the list of unacceptable side effects (risk) can possibly be smaller.
- If none of these conditions are fulfilled, the antibiotic is assigned to category $TBA$.

**Coupling ordered classification with the Semantic Model** MR Sort with Veto is combined with the Semantic models to assess one by one each antibiotic $A_i$ (as illustrated in Fig. 1).

These assignment principles are implemented using the following mathematical representation:

- $A_i, i \in 1, \ldots, n$ denotes the antibiotics that are potentially considered, i.e., antibiotics that cover the germs infecting the patient.
- $S_j, j \in 1, \ldots, m$ denotes the possible side effects of an antibiotic.
- $CSA_{ij}, i \in 1, \ldots, n, j \in 1, \ldots, m$ is a variable taking its values from $\{0, 1, 2\}$. It represents the level of risk of side effect $S_j$ of antibiotic $A_i$ for the patient (taking into account the sensitivity of the patient to side effect $S_j$.

$$CSA_{ij} = \begin{cases} 
0 & \text{if } SP_j = 0 \\
1 & \text{if } SP_j > 0 \text{ and } 1 \leq RSA_{ij} \leq 2 \\
2 & \text{if } SP_j > 0 \text{ and } RSA_{ij} = 3
\end{cases}$$

- $CSA_i$ is a number associated to antibiotic $A_i$. It counts the number of risky side effects $S_j$ such that $CSA_{ij} \neq 0$.
- Two tolerance levels $\lambda_R, \lambda_P$ with $\lambda_R < \lambda_P$ determine the maximal number of side effects (risk) that are compatible with an assignment in categories $R$ and $P$, respectively.
- An unacceptable risk ($CSA_{ij} > 1$) can prevent assignment of $A_i$ to category $R$ or $P$. The list of unacceptable risks for an assignment to category $R$ (resp. $P$) is a subset $Veto[R]$ (resp. $Veto[P]$) of the set of all side effects risks. The vetoes are presented in Table 3.

For a given Patient, the assignment of a suitable antibiotic $A_i$ to the class $R$, $P$ or $TBA$ is summarized in Table 4. $CSA_{ij}$ counts the number of side effects the patient is sensitive to. The first tolerance level, $\lambda_R$, is the maximum number of side effects the antibiotic could have to be in the $R$ (recommended) category. The second tolerance level, $\lambda_P$, is the maximum number of side effects the antibiotic could have to be in the $P$ (possible) category. In our application, $\lambda_R$ was empirically set to 6 and $\lambda_P$ to 12. Some side effects can induce a high risk on a patient, which
Figure 1: Using MR Sort method with Veto in the Semantic Model to link and to assess antibiotics for a patient by toxicity risk explains the usefulness of vetoes (Table 3). The first veto, Veto[P], is put when the antibiotic has an unbearable side effect for the patient. This antibiotic could not be prescribed, even though it only has this side effect. For example, this veto would be raised if the antibiotic contains penicillin (RSA_{ij} = 3) and if the patient has a major allergy to penicillin (SP_j = 2). With a Veto[P], the considered antibiotic is put in the TBA category. Similarly, a second veto, Veto[R], is put when the antibiotic has a very rare severe impact on the patient health (RSA_{ij} = 1), which, however, is not unbearable. For instance, an antibiotic which contains penicillin would get this Veto[R] if the patient has a minor allergy to penicillin (SP_j = 1). With the Veto[R], the considered antibiotic A_i is put either in the P (possible) or in the TBA (to be avoided) category, depending on the value of CSA_i.

Table 3: Vetoes

<table>
<thead>
<tr>
<th>Antibiotic (GS_j, FSA_{ij})</th>
<th>(0,0)</th>
<th>(0,1)</th>
<th>(1,0)</th>
<th>(0,2)</th>
<th>(1,1)</th>
<th>(3,0)</th>
<th>(3,1)</th>
<th>(3,2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient SP_j</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
</tr>
<tr>
<td>1</td>
<td>R</td>
<td>P</td>
<td>P</td>
<td>P</td>
<td>P</td>
<td>TBA</td>
<td>TBA</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>P</td>
<td>P</td>
<td>P</td>
<td>TBA</td>
<td>TBA</td>
<td>P</td>
<td>TBA</td>
<td>TBA</td>
</tr>
</tbody>
</table>

Illustration

In order to illustrate the model, we built a scenario through several meetings with practitioners of the EpiCURA Hospital Center (infectious diseases specialist, microbiologist) through which we were able to tune the sensitivities and the thresholds of our model. For the purpose of illustration, we present the following case:

Bill is a 68 years old man, he is suffering of alcoholism problems and he is diabetic. He comes to the emergency room with an increased temperature of 40.1°C and he has an inflammation in his leg (see Fig.2). It is later revealed, that he got injured 2 days before when he was gardening. The diagnostic yielded is “Erysipelas”.

Table 4: Rules for assigning antibiotics to classes R, P and TPA

<table>
<thead>
<tr>
<th>CSA_i &lt; λ_R and no veto [R]</th>
<th>Recommended</th>
</tr>
</thead>
<tbody>
<tr>
<td>CSA_i &lt; λ_P, no veto [P] and not Recommended</td>
<td>Possible</td>
</tr>
<tr>
<td>CSA_i &gt; λ_P or veto [P]</td>
<td>To be avoided</td>
</tr>
</tbody>
</table>
Table 5: Sensitivities of Bill

<table>
<thead>
<tr>
<th>$S_{Pa}$</th>
<th>$S_{Pb}$</th>
<th>$S_{Pc}$</th>
<th>$S_{Pd}$</th>
<th>$S_{Pe}$</th>
<th>$S_{Pf}$</th>
<th>$S_{Pg}$</th>
<th>$S_{Ph}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>


Figure 2: Erysipelas around the ankle

Figure 3: Guidelines recommendation for Erysipelas

The guidelines (Fig.3) inform us about the pathogens which cause the infection. In our example, the pathogens in question are *Staphylococcus Aureus* (MSSA) and *Group A streptococci* (GAS), as indicated in the second line of Fig.3. To suggest an appropriate antibiotic, the guidelines distinguish three situations with respect to penicillin: (a) a patient who is not allergic to penicillin, (b) a patient with a minor allergy and (c) a patient with a major allergy. In the latter two cases, it suggests two different antibiotics.

Table 5 gives us the set of side effects (we did not present all the side effects that Bill is sensitive to, because for this use case the set of all antibiotics (Table 6) that can help Bill don’t have them; in terms of frequency, $FSA_{ij} = 0$), that Bill is sensitive to. $SP_j$ denotes the sensitivity of Bill to the side effect $S_j$, $j = a, ..., h$.

The list shown in Table 6 gives us: The set of antibiotics which are effective against or cover the germs causing Bill’s infection; the set of side effects $S_j$; the severity evaluation of the side effects $GS_j$; the frequency $FAS_{ij}$ and the risk $RSA_{ij}$, where $i = 1, ..., 8$ is the set of antibiotics and $j = a, ..., h$ is the set of side effects.

We sort this list by suitability to Bill. The output of our system for this case is the following:

**P**: Oxacillin, $CSA = 2$, Veto[$R$]

P: Amoxicillin-Clavulanic, $CSA = 1$, Veto[$R$]

P: Piperacillin-Tazobactam, $CSA = 2$, Veto[$R$]

TBA: Clindamycin, $CSA = 2$, Veto[$P$]

TBA: Cefazolin, $CSA = 2$, Veto[$P$]

TBA: Ceftriaxone, $CSA = 2$, Veto[$P$]

TBA: Vancomycin, $CSA = 2$, Veto[$P$]

TBA: Moxifloxacin, $CSA = 4$, Veto[$P$]

Note that, currently, in order to get these results (without using our system), the physician has to manually cross-check and combine several different sections of the guidelines.

More precisely, for a patient that is as much vulnerable as Bill, at least one of the vetoes [$R$] and [$P$] are activated and therefore, no antibiotic is assigned to category “R”. For this kind of situation, indeed, we rarely find antibiotics sorted in the “Recommended” category. The best alternatives are usually sorted in the “Possible” category. The advantage of this kind of output is that it provides the prescribing physician with a global view of the sorted treatments. In addition it gives him/her a detailed explanation of the toxicity risk assessment for each particular patient.
Table 6: Frequency, severity and risk of side effects per antibiotic

<table>
<thead>
<tr>
<th>Antibiotics</th>
<th>$S_a$</th>
<th>$S_b$</th>
<th>$S_c$</th>
<th>$S_d$</th>
<th>$S_e$</th>
<th>$S_f$</th>
<th>$S_g$</th>
<th>$S_h$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oxacillin</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Cefazolin</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Ceftriazone</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Amoxicillin, Clavulanic</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Piperacillin, Tazobocamt</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Vancamycin</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Moxifloxacin</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Clindamycin</td>
<td>2</td>
<td>3</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

To describe in further details our model, we present here a second scenario. For this scenario we consider the case of "Abscess Perirectal" from the guidelines (as seen in Fig.4).

![Abscess Perirectal](image)

**Figure 4:** Guidelines recommendation for Abscess Perirectal

The guidelines inform us about the pathogens which cause the infection in the second line of Fig.4. For this example, the pathogens are Enterobacteria, Anaerobes, Enterococcus. To suggest an appropriate antibiotic, the guidelines distinguish three possible contexts with respect to penicillin: (a) a patient who is not allergic to penicillin, (b) a patient with a minor allergy and (c) a patient with a major allergy. In the last two cases, the guidelines suggest the same antibiotic.

The list presented in Table 7 gives the set of antibiotics which are effective against the germs causing the infection (Amoxicillin, Clavulanic, Piperacillin, Tazobocamt, Moxifloxacin). The set of side effects $S_j$, where $j = 1, \ldots, 22$ is presented underneath. These side effects can be given by the antibiotics: Amoxicillin, Clavulanic, Piperacillin, Tazobocamt and Moxifloxacin.

The severity evaluation of the side effects $GS_j$; the frequency $FAS_{ij}$ and the risk $RSA_{ij}$ (where $i = 1, \ldots, 3$ is the set of antibiotics and $j = 1, \ldots, 22$ is the set of side effects) are all presented in Table 7. The case description follows:
### Table 7: Frequency, severity and risk of 22 side effects for 3 antibiotics

<table>
<thead>
<tr>
<th>Antibiotics</th>
<th>$S_1$</th>
<th>$S_2$</th>
<th>$S_3$</th>
<th>$S_4$</th>
<th>$S_5$</th>
<th>$S_6$</th>
<th>$S_7$</th>
<th>$S_8$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amoxicillin, Clavulanic</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Piperacillin, Tazobocamt</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Moxifloxacin</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

### Table 8: Sensitivities of Alex

<table>
<thead>
<tr>
<th>SP_1</th>
<th>SP_2</th>
<th>SP_3</th>
<th>SP_4</th>
<th>SP_5</th>
<th>SP_6</th>
<th>SP_7</th>
<th>SP_8</th>
<th>SP_9</th>
<th>SP_10</th>
<th>SP_11</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SP_12</th>
<th>SP_13</th>
<th>SP_14</th>
<th>SP_15</th>
<th>SP_16</th>
<th>SP_17</th>
<th>SP_18</th>
<th>SP_19</th>
<th>SP_20</th>
<th>SP_21</th>
<th>SP_22</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Alex is a 29 years old patient, who is suffering from “Abscess Perirectal”. He is otherwise in good health without any history of medical incidents. His laboratory tests reveal that he does not have allergies and his creatinine level is 80ml/min.
For this case of infection, a surgical drainage is essential with antibiotic prescribing. Table 8 indicates the sensitivities $SP_j$ of Alex to the side effects $S_{ij}, j = 1, \ldots, 22$.

We sort this list by suitability for Alex. The output of our system for this case is the following:

- **P**: Amoxicillin_Clavulanic, $CSA = 1$, Veto[R]
- **P**: Moxifloxacin, $CSA = 1$, Veto[R]
- **P**: Piperacillin_Tazobocamt, $CSA = 2$, Veto[R]

If Alex has a “Major allergy to penicillin”, the sensitivities $SP_{19}$ and $SP_{21}$ change: $SP_{19} = 2$ and $SP_{21} = 2$. The output of the system becomes:

- **P**: Moxifloxacin, $CSA = 1$, Veto[R]
- **TBA**: Piperacillin_Tazobocamt, $CSA = 4$, Veto[P]
- **TBA**: Amoxicillin_Clavulanic, $CSA = 3$, Veto[P]

The recommendations of our system match the ones in the guidelines (as seen in Fig.4). Even in the case of “Major allergy to penicillin”, Moxifloxacin remains the only option.

**Conclusion**

We have developed a novel approach to help physicians with antibiotic prescriptions. Our solution combines semantic technologies with our own adaptation of the MR Sort method with Veto.

Our method sorts antibiotics in three categories: **R** (“recommended”), **P** (“possible”) and **TBA** (“to be avoided”) based on a small number of general rules. It is able to take into account a patient’s specific clinical criteria as well as generalize to new cases when – for example – a new antibiotic is added to the knowledge base. Using input from practitioners in the EpiCURA Hospital Center we have tuned the sensitivities and thresholds of our model and were able to validate our approach through examples that score prescription recommendations according to the risk of side effect toxicity.

In terms of future work we would like to apply our model to other types of drugs to determine if our method is applicable beyond antibiotics. We plan to expand our model to other dimensions of the prescription problem including: costs, drug-drug interaction and drug-disease interaction among others.

**Acknowledgement**

We are grateful to Dr Jean Pierre Sabot – medical director of the EpiCURA Hospital Center – for his warm welcome and availability considering our research. We also want to thank Dr. Sammy Place (infectiologist), Dr. Lorenzo Filippin (microbiologist) as well as all the EpiCURA Hospital Center staff for numerous helpful meetings that allowed us to improve and validate our approach.

**References**


A Data-Driven Method for Generating Robust Symptom Onset Indicators in Huntington’s Disease Registry Data

Zhaonan Sun1, PhD, Ying Li1, PhD, Soumya Ghosh1, PhD, Yu Cheng1, PhD, Amrita Mohan2, PhD, Cristina Sampaio2, MD, PhD, Jianying Hu1, PhD
IBM T.J. Watson Research Center, Yorktown Heights, NY1; CHDI Management/CHDI Foundation, Princeton, NJ2

1 Introduction

A disease registry is an organized system that uses observational study methods to collect uniform data to evaluate specified outcomes for a population defined by a particular disease, condition, or exposure, and that serves one or more predetermined scientific, clinical, or policy purposes[1]. Different from Electronic Health Record (EHR), whereas the research of understanding a particular disease belongs to the secondary use of data, the disease registries serve as a primary source to study the target disease. As a primary source, the registry data usually involves data generated from known and comprehensive clinical assessments for the target disease, and therefore a disease registry can be a powerful tool to observe the course of disease, to understand variations in treatment and outcomes, to examine factors that influence prognosis and quality of life, and to assess the clinical cost and quality of care.

The natural course of a disease can be characterized by the onset and progression of symptoms. Measures from clinical assessments are collected in a disease registry for tracking various symptoms of a target disease. However, such measures can be biased since they not only can be affected by the progression of the target disease, but also can be influenced by non-disease-related factors, e.g. selection bias, clinical instrument sensitivity and participant compliance. Moreover, for a symptom of interest, its onset and progression indicator are decided by threshold values of recorded clinical measures. Such threshold values are currently determined based on evaluation of measures recorded in comparable historical studies or by domain expert, and are applied to all participants in a registry. Consequently, they do not take individual variations into account. In order to address the aforementioned issues, data-driven methods are needed to generate robust (less biased) and personalized symptom onset indicators, and as a result, in return for better understanding the natural course of a target disease.

A disease registry may include only people with disease of interest, or may also include one or more comparison groups for which data are collected using the same methods during the same period. Hereinafter, we refer to patients with disease of interest as case participants, while patients in the comparison group who are not at risk as control participants. Control participants may share some similar traits or are exposed to similar environmental factors as case participants. In this article, we propose a data-driven method to generate robust symptom onset indicators based on assessments collected in disease registry data with the requirement that this disease registry should have included control participants. According to our knowledge, the proposed method is the first of its kind that uses control participants in a disease registry to adjust the biases inherited in the raw clinical measurements among case participants. The biases are caused by non-disease-related factors such as natural aging process, education level and marital status. In the remainder of this paper, we will exemplify the application of this novel method to integrated data from observational Huntington's Disease studies.

Huntington’s Disease (HD) is an autosomal dominant fully-penetrant neurodegenerative disorder, which is caused by an abnormal expanded trinucleotide (CAG) repeat in the Huntingtin (HTT) gene[2]. Owing to its monogenic nature, predictive genetic testing is able to determine whether the disease will manifest in an individual. Among genetically confirmed HD patients, a clinical diagnosis of HD is typically made when an individual exhibits overt, otherwise unexplained extrapyramidal movement disorder. While motor impairment is currently the primary indicator of clinical onset, cognitive[3] and certain behavioral disorders[4] are also known to surface years before motor onset. As such, clinical measurements along these dimensions are also important for understanding the progression of the disease. Functional assessments in particular are important in measuring overall quality of life of individuals with HD and prove useful for a descriptive characterization of HD progression.

In recent years, several large-scale observational studies have been conducted in HD gene expansion carriers (HDGECs)
with the hope to understand the natural history and pathophysiology of the disease. A diverse range of clinical assessments have been designed in HD observational studies to record the triad of motor, cognitive/behavior, and functional symptoms of HD. While accessibility to a wide range of clinical assessments from these domains has helped gain insightful information about the natural history of HD, these clinical assessments are often influenced by factors other than HD disease status and progression. For instance, natural aging processes are especially known to affect participants cognitive and functional abilities. Therefore, absolute changes in most cognitive and functional assessment scores can be attributed to multiple factors including both HD disease progression and the natural aging process, rendering assessment scores less robust for tracking disease progression.

Onset of new symptoms is useful in characterizing the course of HD. Certain clinical measures collected have pre-defined thresholds to indicate reliability of clinical diagnosis. For example, a Diagnostic Confidence Level (DCL) of 4 is used to indicate motor symptom onset, thereby leading to a confirmed diagnosis of HD onset. However, other measures (e.g. the SDMT score, which is used to assess cognitive abilities) do not have clearly defined thresholds for disease onset. To address this inconsistency in a systematic manner, a data-driven method is needed for generating robust symptom onset indicators for better understanding of the progression of the disease.

In this article, we propose a data-driven procedure for adjusting the values of clinical assessments and generating robust symptom onset indicators. Its worth pointing out that although this article and methodology described here relies on HD observational data sets, the methodology can be generalized to other disease registry data if they have also recruited control participants.

For readers benefit, the contents of this paper are organized as follows. In Section 2, we describe the HD observational data used in this work. In Section 3, we describe a novel method for generating robust clinical assessments from the outcome measures recorded in the observational datasets. In Section 4, we use two cognitive assessments to demonstrate the properties of the generated assessment score. Lastly, Section 5 provides a summary and brief discussion of the new method.

2 Data Sources

In this study, we integrated data from four large prospective observational studies of HD, namely Enroll-HD[5], REGISTRY[6], TRACK-HD/TRACK-ON[7, 8], and PREDICT-HD[9], respectively.

Enroll-HD is a worldwide observational study of Huntington’s Disease families. The study aims at providing a platform to support the design and conduct of future clinical trials, improving the understanding of the phenotypic spectrum and the disease mechanisms of HD and improving health outcomes for the participant/family unit. The study monitors how HD appears and changes over time in different subjects. It recruits confirmed HD patients, HD at-risk patients, HD genotype negative participants as well as control participants from HD family. Study participants are required to visit study sites annually, and undergo a comprehensive battery of clinical assessments. In this work, we used the ENROLL-IDS-2015-10-R1 version of the Enroll-HD periodic data, which contains un-monitored data from 7614 subjects who made their baseline visits prior to October 2015. Among the participants, 5475 are Huntington’s disease gene expansion carriers (HDGECs) with CAG length greater than 35, 1613 participants are control subjects (i.e. CAG length ≤ 35), and the other 527 have unknown CAG length. Subjects have up to four annual visits, with an average number of visits being 1.44.

REGISTRY is a multi-center, multi-national observational study, managed by the European Huntington’s Disease Network (EHDN), with no experimental intervention. REGISTRY aims at obtaining natural history data on many HD mutation carriers and individuals who are part of an HD family, relating phenotypical characteristics of HD, expediting the identification and recruitment of participants for clinical trials, developing and validating sensitive and reliable outcome measure for detecting onset and change over the natural course of pre-manifest and manifest HD. The REGISTRY cohort used in this study consists of 12108 participants, among which 7988 are HDGECs (i.e. CAG > 35), 758 are control participants (i.e. CAG length ≤ 35), and the other 3894 participants do not have CAG length information. Participants have up to 15 annual visits, with the average number of visits equals to 2.9.

TRACK-HD is a multinational study of HD that examines clinical and biological findings of disease progression in individuals with pre-manifest HD and early-stage HD. Participants in the study underwent annual clinical assessments
for 36 months. At the baseline visit, 402 participants were enrolled. Among the participants, 127 participants were control subjects, 144 participants were pre-manifest subjects who had not reached HD clinical onset, and 130 were post-manifest subjects who had already reached HD clinical onset. 298 participants completed the 36-month follow-up, among which 97 were controls, 104 were pre-manifest subjects at their baseline visits, and 97 were post-manifest subjects at their baseline visits.

TRACK-ON is a follow-up study of TRACK-HD with the aim of testing for the compensatory brain networks after structural brain changes in TRACK-HD pre-manifest participants. Participants in the study underwent annual clinical assessment for 24 months. At the baseline visit, 245 participants were enrolled, among them 181 were participants of TRACK-HD who have not reached HD clinical onset at the end of TRACK-HD, and 64 participants were newly recruited in the study. 112 participants in TRACK-ON were control subjects, and others are HDGECs.

PREDICT-HD is another longitudinal observational study of subjects who chose to undergo predictive testing for the CAG expansion in the HD gene but did not meet criteria for a diagnosis of HD (Diagnostic Confidence Level = 4). Participants were recruited from 32 sites worldwide beginning in October 2002. The goal of PREDICT-HD is to define the neurobiology of Huntington’s disease (HD) and to develop tools to allow clinical trials of potential disease-modifying therapies before at-risk individuals have diagnosable symptoms of the disease. It collected a variety of biosamples including MRI, blood and urine samples, and comprehensive assessments of cognitive, motor, functional and psychiatric outcomes to characterize the pre-manifest syndrome in HD, to document the rate of change of these variables during the years leading up to and following a clinical diagnosis of HD, and to investigate the relationship among neurobiologic factors, clinical diagnosis and CAG repeat length. The PREDICT-HD data used in this study consists of 1481 participants. Among them 316 were control subjects. Participants have up to 14 annual study visits, with the average number of visits equals to 5.2.

3 Materials and Methods

3.1 Integration of Multiple HD Data Sets

The four studies introduced in Section 2 contain a diverse set of clinical assessments that span a spectrum of clinical symptoms expressed by HD patients. In this section, we briefly describe the process of integrating data from these four studies.

We began by matching subjects across studies using a unique Recoded HD participant ID. This unique identifier also allows us to recognize the subjects who participated in multiple studies. In the four HD observational studies, participants visited study sites approximately annually and were evaluated by a diverse range of clinical assessments. In the rest of this paper, we refer to the data generated from one visit of one participant as an observation. In each of the four studies, the date of a participant’s first study visit in the study, referred to as the baseline visit, was used as the reference date for the participant and was set to 0. The visit dates of all his follow-up visits in the same study were aligned with the reference date and measured in days. In addition, for a subject who participated in multiple studies, the time gaps between the multiple reference dates from different studies were also available. Therefore, subjects’ records from multiple studies could be stitched together when they were available.

The second step of combining the multiple data sets was matching and merging variables. Not all variables were named consistently across studies. We analyzed data dictionaries, study protocols and guidelines from the four studies and manually matched variables across studies. We also corrected coding inconsistencies across studies.

We categorized variables into two groups, namely, assessment score and demographical information. The assessment score group consists of measurements from clinical assessments performed at annual study visits to capture wide range of clinical symptoms among HDGECs, such as motor impairment, cognitive deficits, functional decline, and behavioral disorder. The demographic information group includes participants’ demographics (e.g. age, sex, education level, etc.), CAG length, medical history (e.g. drug abuse history, alcohol abuse history, etc.), and other information related to study designs (e.g. region, study site). The integrated data set contains 106 variables from the participants demographic information group and 2079 variables from the assessment scores group.

Finally, we performed cross-study distributional check to filter out obvious erroneous measurements in the integrated
After all these steps, we ended up with a data set containing 55782 observations from 16553 HDGECs and 2716 control participants. The average number of observations per participant of the integrated data is 2.9.

3.2 Flow of the proposed new method

The aim of this work is to develop a data-driven method to generate robust symptom onset indicators based on the clinical assessment scores collected in HD registry data. In the rest of this paper, we will refer to the original assessment scores as the ‘raw assessment scores’, and the newly generated scores as the ‘robust scores’. A robust score is generated from its raw assessment score. As proof of concept and application, we will showcase the proposed method using two cognitive assessments in the integrated data, which are Symbol Digit Modalities Test (SDMT) total correct score, and the Stroop Word Reading Test (SWRT) total score.

As discussed in Section 1, raw assessment scores can be influenced by both disease-related and non-disease-related factors. While influence of the disease-related factors on these assessment scores is desirable, influence of non-disease-related factors on these scores can result in misleading follow-up analysis and conclusions. Figure 1 illustrates an example from the raw SDMT total correct scores. The left panel (1a) depicts the distributions of raw SDMT scores vs. the levels from International Standard Classification of Education (ISCED). The right panel (1b) depicts the distributions of raw SDMT scores vs. HD stages of participants in the aggregated data. HD clinical stage has been discretized here into early pre-manifest, late pre-manifest, HD1, HD2, HD3 and HD4. It is used here as a surrogate for HD disease progression. From the figures, it is clear that SDMT scores not only depend on HD disease status and progression, but also depend on other factors such as education levels.

The integrated HD observational data includes control participants who do not carry HD gene expansions. Control participants by definition are not affected by HD disease progression. Therefore, raw assessment scores of control participants are expected to be influenced by non-disease-related factors. Throughout this study, we assume that the underlying effects of a non-disease-related factors on an assessment score are the same for both control participants and HDGECs. Any differences in the effects between the two groups can be attributed to the disease. The basic idea of the proposed procedure is to utilize the control cohort to evaluate the effects of non-disease-related factors on an assessment score of interest. Then for a case participant, the predicted control value of the assessment score can be produced from the control-based model. The predicted control value gives an estimate of what the expected assessment score look like for a hypothetical control participant with similar characteristics as the case participant. Subtracting the predicted control value from the observed value can remove the effects of non-disease-related factors. The remaining part of the assessment score is less subject to changes of non-disease-related factors, and is more robust in terms of reflecting the effects of HD related factors. The onset of a new symptom (e.g. cognitive impairment) is defined to be
the critical point at which a case participant exhibits significant difference from control participants. Comparing the observed assessment scores with the distributions of the predicted control values could lead to personalized symptom onset indicators.

Next we describe the work flow of generating the robust measure of a target assessment score. For each target assessment score, the framework consists of a sequence of steps: (1) check missing values. If there are missing values in the data, perform imputation to generate multiple sets of complete data sets. (2) For each imputed dataset, build a model for the target assessment with participants’ characteristics, using control subjects only. The model is referred to as the control model for the imputed dataset. (3) For each imputed dataset, get predicted control values and the prediction confidence interval for HDGEC based on the control model from step (2). (4) For each imputed dataset, generate robust assessment scores for target assessment scores. (5) Aggregate the robust assessment scores from multiple imputed datasets. If there is no missing values in the data, steps (2)-(4) will be performed on the observed data set, and step (5) will no longer be needed. We discuss each step in detail below.

**Step 1.** Check missing values and perform multiple imputation.

All four HD observational studies have missing values. Therefore, the aggregated data contain missing values. To cope with this problem, we performed Multiple Imputation (MI) with the Fully Conditional Specification method [10] and Predictive Mean Matching [11] to impute the missing values and generated multiple sets of complete data sets. Multiple Imputation [12] is a statistical technique for analyzing incomplete data sets. Instead of filling in a single value for each missing value, MI procedure replaces each missing value with a set of plausible values. Uncertainty about the value to impute can be represented by the multiple imputed values. These multiple imputed data sets are analyzed individually. Results from the multiple sets of complete data sets are then aggregated to generate the final results. In this paper, we applied the MI procedure using the MICE package in R [13] and generated ten sets of complete data sets.

**Step 2.** Build control models.

With each imputed data set, we build a model for the target assessment score using available patient characteristics as the predictors. The goal of this step is to build a model with high predictive power. For each target assessment score of interest on each imputed dataset, we compared multiple candidate predictive models, and choose the one with the highest predictive power (measured by R-squared) as the model of choice for the target assessment score on the imputed dataset. In this study, we used three types of models as candidate control models in the experiments: the generalized linear regression model, Support Vector Machines (SVM) with RBF kernel, and Multivariate Adaptive Regression Splines (MARS). The proposed method is not limited to the three types of models. Other types of predictive models can be included in this step.

**Step 3.** Get prediction confidence interval of a target assessment score on HDGECs from control models.

Once a candidate model is selected as the control model for a target assessment score on an imputed data set, we obtain predicted control values of the target assessment scores for HDGECs on the imputed data set. We also obtain the lower and upper bounds of the 95% confidence interval of the predicted control value (PCI) for each case observation. In this paper, we used bootstrap method to obtain the PCI for case observations on each imputed dataset.

**Step 4.** Generate robust assessment scores.

The PCI obtained from Step 3 gives an interval estimate of what an assessment score would be for a hypothetical control participant with similar characteristics as a case participant. We define the symptom onset as the event when a case participant can be distinguished from the control participants. The PCI obtained from the previous step can be used to mark the boundary to determine whether a case participant presents significant difference from the controls. Therefore, the time of symptom onset is defined to be the first time that an assessment score of a case participant...
Figure 2: Description of the method in Step 4

falls out of the PCI. Following the above definition, the robust assessment score is defined as the distance between the observed assessment score and the boundary of the PCI. Figure 2 summarizes the method. If the raw assessment score decreases with time (left panel of Figure 2), the difference between the observed value and the lower bound of the PCI is used as the robust assessment score. If the raw assessment score increases with time (right panel of Figure 2), the difference between the upper bound of PCI and the observed value is used as the robust assessment score. The sign of a robust assessment score from an observation can indicate whether the participant show significant difference from controls in the symptom assessed by the corresponding raw assessment score. A positive sign indicates that the participant cannot be distinguished from controls. A negative sign indicates that the participant can be distinguished from controls. The value 0 serves as a natural threshold for deciding the onset of a new symptom. Therefore, the robust assessment score can be used as a symptom onset indicator. Note that a robust assessment score can be generated from each raw assessment score. A robust assessment score assesses the same symptom as the symptom targeted by its corresponding raw assessment score. Multiple robust assessment scores together could be used to provide a comprehensive view of the progression pathway of the target disease.

Step 5. Aggregate results from multiple imputed data.

Up to Step 4, for each HDGEC observation, we generate ten robust assessment scores, each from one imputed data set. The final step is to aggregate the ten sets of robust assessment scores. In this study, we used the average across the ten sets as the final robust assessment scores. When there is no missing values in the observed data set, steps 2-4 will be performed on the observed data set and step 5 will be skipped.

4 Results

We applied the proposed procedure to the integrated HD observational data. In this section, we demonstrate the characteristics of the robust assessment scores using two cognitive assessments: the Stroop Word Reading Test (SWRT) score and the Symbol Digit Modalities Test (SDMT) total correct score.

Both the two assessments have missing values in the data. Step 1 of the proposed method generated multiple complete data sets using the Multiple Imputation method. We first check the quality of the imputed data. The imputed values should 1) have the same support as the observed data; 2) have similar distributions as the observed data. We adopted the Predictive Mean Matching method in the MI step, therefore the imputed values were guaranteed to have the same support as the observed data. In this section, we compare the distributions of the imputed values with the observed values by visual inspection. One example of the inspection is showed in Figure 3. The left panel of Figure 3 shows the distribution of observed SDMT scores in each age group, and the right panel shows the distribution of imputed SDMT scores in each age group. The plots demonstrate that the distributions of the imputed value are similar to that of the observed values. Similar inspections were performed for observed and imputed values versus other factors. Due to lack of space, we do not show the details in this paper.

After obtaining the ten sets of imputed data, we build control models for SDMT and SWRT separately on each
Figure 3: Boxplots of imputed and observed SDMT scores in each age group

Patient characteristics, such as study ID, age, gender, and education levels, were used as predictors in the control models. For each assessment score on each imputed dataset, three types of candidate models (Generalized Linear Regression (GLM), Support Vector Machine (SVM) and Multivariate adaptive regression splines (MARS)) were compared. The type of model with the highest R-squared value was selected to build the control model on the imputed data set. A summary of the R-squared values from the selected models are listed in Table 1. After building the control models, bootstrap method was used to obtain the 95% PCI for each case observation. The robust assessment scores for each complete dataset were calculated following the descriptions in Section 3. In the final step, we aggregated the robust scores from the ten sets of complete data sets by calculating the mean values across the ten sets of complete data sets.

<table>
<thead>
<tr>
<th></th>
<th>Mean of R-squared</th>
<th>std. of R-squared</th>
</tr>
</thead>
<tbody>
<tr>
<td>SDMT</td>
<td>0.3994</td>
<td>3.46 × 10^{-5}</td>
</tr>
<tr>
<td>SWRT</td>
<td>0.2519</td>
<td>1.68 × 10^{-5}</td>
</tr>
</tbody>
</table>

Table 1: Summary of selected model types and R-squares of the control models

Next we discuss the properties of the robust assessment scores. Values of raw assessment scores were not only influenced by HD disease status and progression, but also by other non-disease-related factors. The proposed method utilizes the control cohort to model and adjust the effect of non-disease-related factors. The robust assessment scores are expected to be less subject to the non-disease-related factors. Figure 4 and 5 show two examples comparing the distributions of raw and robust SDMT scores among the HDGECs versus two patient characteristics, which are age groups and ISCED education levels. The left panels show the distributions of the raw SDMT scores vs. age groups and ISCED education levels, respectively. The right panels show the distributions of the robust SDMT scores vs. age groups and ISCED education levels, respectively. The raw SDMT scores demonstrate strong correlation with age and education levels, while the robust SDMT scores demonstrate decreased correlation with the two factors.

Table 2 summarizes the influences of non-disease-related factors in raw and robust SDMT/ SWRT scores. For categorical factors, we calculate the average Cohen’s d effective sizes of the scores between pairs of levels of the factor. For continuous factors, we report the Spearman’s correlation coefficients. A smaller average effective size or smaller absolute value of Spearman’s correlation coefficient indicates decreased influence of the non-disease-related factor.
Most non-disease-related factors show decreased influence in the robust scores. The few exceptions can be attributed to the imbalance of factors’ effects in the control model. In general, the robust scores are less subject to changes in non-disease-related factors.

The values of raw SDMT scores range from 0 to 120. A few previous literature [14, 15] reported some threshold values to distinguish normal vs. impaired cognitive abilities. However, whether these threshold values can be applied to HDGECs has not been systematically tested. There is no clear threshold to decide when a HDGEC participant starts to show the symptom of cognitive impairment. A robust SDMT score comes with a natural threshold (i.e. 0) for deciding whether a HDGEC starts to show cognitive impairment. Since influences of the non-disease-related factors for the observation have been adjusted in the robust assessment score, the threshold is personalized and specific to HD. Similarly, a value of 0 serves as a natural personalized threshold for other robust assessment scores to determine the onset of other symptoms in HDGECs. It is worth mentioning here that ‘onset’, for a case participant, is being identified as a deviation from the model learned based on control participants. This data driven ‘onset’ is a hypothesis that needs further clinical validation.

<table>
<thead>
<tr>
<th></th>
<th>Study</th>
<th>Age</th>
<th>Marital Status</th>
<th>Education Level</th>
<th>Gender</th>
<th>Region</th>
<th>Tobacco abuse history</th>
<th>Drug abuse history</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMDT</td>
<td>0.798</td>
<td>-0.378</td>
<td>0.493</td>
<td>0.918</td>
<td>0.216</td>
<td>0.715</td>
<td>0.574</td>
<td>0.674</td>
</tr>
<tr>
<td>Robust SDMT</td>
<td>0.679</td>
<td>-0.269</td>
<td>0.277</td>
<td>0.287</td>
<td>0.309</td>
<td>0.622</td>
<td>0.596</td>
<td>0.514</td>
</tr>
<tr>
<td>SWRT</td>
<td>0.793</td>
<td>-0.320</td>
<td>0.438</td>
<td>0.902</td>
<td>0.169</td>
<td>0.653</td>
<td>0.660</td>
<td>0.636</td>
</tr>
<tr>
<td>Robust SWRT</td>
<td>0.587</td>
<td>-0.280</td>
<td>0.254</td>
<td>0.250</td>
<td>0.129</td>
<td>0.795</td>
<td>0.511</td>
<td>0.407</td>
</tr>
</tbody>
</table>

Table 2: Correlations and average effective sizes of raw and robust SDMT/SWRT scores with patient characteristics

5 Discussion

In this paper, we proposed a method for generating robust assessment scores and symptom onset indicators from patient registry data set which have recruited both case and control participants. The generated scores are more robust in the sense that they are less subject to changes due to non-disease-related factors. Therefore are more relevant when evaluating subjects’ disease status and tracking disease progression. The signs of the robust assessment scores indicate whether an observation can be distinguished from the control cohort. The value 0 serves as a natural indicator
to indicate the onset of the symptom assessed by the score.

We applied the proposed procedure to an integrated HD observational data set and discussed the properties of the robust assessment scores using two cognitive assessments. The proposed procedure is not limited to these two assessment scores. It could be applied to other assessment scores in HD observational data.

The proposed method may also be applied to other disease registry data that have recruited both case and control participants. However, the application of the proposed method to other disease registry data should be conducted with caution. Owing to the monogenic nature of HD, the identification of the control participants in HD observational data is relatively clearer compared to other types of diseases. Participants in the control cohort of HD data sets by definition are not expected be affected by any direct (or known) HD disease-related factors. The identification of a control cohort in other disease registry may require more efforts such as matching and stratification.

One limitation of the proposed method comes from the nature of observational studies. The proposed method can only adjust the effects of non-disease-related factors that are available among both case and control participants. If a factor is not collected in the patient registry data or is only available in one of the groups, its influences cannot be adjusted. In other words, the proposed procedure only mitigates the effects of non-disease-related factors, but does not guarantee elimination of their effects entirely.

Despite the aforementioned limitations, the proposed procedure is useful in improving the quality of assessments scores in patient registry data. The value of zero serves as an indicator for detecting symptom onsets. Better understanding of the course of the disease could be obtained by comparing the times and order of multiple symptoms. We will explore the clinical insights of the generated symptom onset indicators in future work.

Acknowledgement

Enroll-HD: Data used in this work was generously provided by the research participants in ENROLL-HD and CHDI Foundation, Inc that financially supports the study.

TRACK (TRACK-HD & TRACK-ON): Data used in this work was generously provided by the research participants from the TRACK-HD and TRACK-ON studies and made available by the TRACK Investigators.

REGISTRY-HD: Data used in this work was generously provided by the the participants of the European Huntington’s Disease Network (EHDN) REGISTRY study and made available by the EHDN REGISTRY Investigators.
PREDICT-HD: Data used in this work was generously provided by the participants in PREDICT-HD and made available by the PREDICT-HD Investigators and Coordinators of the Huntington Study Group, Jane Paulsen, Principal Investigator.

References


Designing Decision-Support Technologies for Patient-Generated Data in Type 1 Diabetes

Si Sun, PhD1, Kaitlin L. Costello, PhD2
1IBM T.J. Watson Research Center, Yorktown Heights, NY
2Rutgers University, New Brunswick, NJ

Abstract

People living with type 1 diabetes generate data as a byproduct of diabetes management. The development of decision support technologies can be enabled by harnessing these patient-generated data, but a major challenge is for these technologies to provide meaningful and highly personalized guidance to support individual patients’ decision-making processes. In this paper, results from a year-long qualitative study were reported. Twenty-six people with type 1 diabetes were interviewed regarding the types of self-generated data they use for decision-making, their decision-making processes using self-generated data, and the difficulties they experience when attempting to use this data for decision-making. These patients’ behaviors and difficulties point to new approaches to designing decision support technologies for personal use, including patient-centered and automated data entry, automated and individualized data analysis, and humanized output.

Introduction

Patients with type 1 diabetes are forced into frequent decision-making processes related to their health on a daily basis. Data generated by type 1 diabetes patients (e.g., blood glucose levels, food consumption, mood, stress levels, and co-morbidity medication) can be a valuable resource for patient decision-making. This paper examines how patients with type 1 diabetes use various types of self-generated health data in their everyday decision-making processes. Understanding this issue will help improve consumer decision support technologies and enhance our capacity to use patient generated data in clinical decision support technologies.

In this paper, self-generated health data refers to health information that is collected by patients about their condition (e.g., blood glucose readings collected with a glucometer), their actions (e.g., exercise intensity), or their experiences (e.g., pain). Patients with type 1 diabetes can achieve better health outcomes by using their self-generated data to make decisions about their health, such as dietary choices spurred by frequent monitoring of their blood glucose levels.

However, maintaining and using self-generated health data can be challenging. An average diabetes patient spends about an hour every day on self-care. Collecting and maintaining health data adds to patients’ responsibilities and intensifies patients’ stress. Also, in many cases this patient work is invisible to health care providers, making it difficult or impossible to address in clinical environments.

Research on consumer-oriented technologies point to the potential value that they have in assisting patients’ efforts to collect and use self-generated health data for decision making. For example, technologies that support frequent monitoring (e.g., continuous glucose monitors) may help patients to maintain better glycemic control. Real-time feedback through mobile phones for diabetes patients based on their blood glucose levels, diabetes medications, and lifestyle choices may lead to improved glycated hemoglobin test results. The medical informatics community is also exerting increased effort in incorporating patient-generated data in clinical information technologies (e.g., patient portal systems and electronic health records), but more research is needed to investigate the impact of this data and potential approaches to facilitate the process.

This paper offers a patient’s perspective toward self-generated health data, including its value and use by patients with type 1 diabetes for decision-making. Our purpose is to improve the usefulness and adoption rates of decision support technologies for these patients by informing technology design through naturalistic research. The research questions include: (a) What types of self-generated health data do patients with type 1 diabetes consider useful for diabetes decision support? (b) How do patients use this data for decision-making? (c) What difficulties do patients encounter when using self-generated data for decision-making?

Related Studies

Studies reporting the types of diabetes patient-generated data used in decision support technologies usually focus on health indicators and health behavior. For example, a randomized controlled trial on a community-based decision
support system, COMPLETE II, tracks patients’ blood pressure, cholesterol, glycated hemoglobin, foot check results, kidney function, weight, physical activity frequency, and smoking frequency\textsuperscript{10}. Other decision support systems reported using similar data to generate personalized decision support, including health outcome (e.g., blood glucose levels\textsuperscript{7,11–16}, Hemoglobin A1c\textsuperscript{17}, blood pressure\textsuperscript{17}, lipid\textsuperscript{17}), diet (e.g., carbohydrate intake\textsuperscript{7,11}), medication\textsuperscript{7,11}, and exercises\textsuperscript{11}. One data type that is not currently well supported by these systems is the illness experience, an important aspect of decision-making processes for patients with diabetes\textsuperscript{1}.

The aforementioned decision support systems were mainly developed for shared decision-making between healthcare providers (HCP) and patients. The types of data entered are predefined, and data entry into these systems requires patient work with the help of sensor technologies. Decision support, however, is mostly directed toward HCPs, not patients. Some data that patients find useful is considered irrelevant by HCPs\textsuperscript{18}. In other words, when using these decision support systems, diabetes patients contribute to data entry but they do not play a major role in making decisions for themselves using the generated data. For a chronic disease that involves day-to-day self-care, type 1 diabetes calls for patients’ effort in daily decision-making\textsuperscript{1}. To date, not enough research has examined patients’ requirements for these technologies\textsuperscript{19}, particularly since they may use the data generated and stored in these technologies for daily self-care efforts.

Recently, the roles that patients and their self-generated data play in decision-making have been explored in several studies. This research suggests that personal preferences, values and context can shape patients’ decision-making processes\textsuperscript{1,20,21}. Disregarding these individualized requirements to pursue standardized goals may not necessarily lead to improved long-term health outcomes\textsuperscript{19,20}. Indeed, research on various diabetes patient decision aids (e.g., cards and apps) attempts to take personal preferences into consideration\textsuperscript{22}. Nevertheless, much remains unknown about how to design technologies that put patients at the center of the decision-making process to support their decision-making activities.

**Method**

**Participant Recruitment**

Participants were recruited through snowball sampling via the researcher’s personal network. The first author recruited the first group of participants among people she knows in person. Then, the first author asked the existing participants to introduce this study to people in their personal networks who were potential candidates. This process carries on through the researcher’s and existing participants’ networks. Snowball sampling is a recruiting method that can effectively and efficiently reach stigmatized populations, making recruitment for this study possible\textsuperscript{23}.

Twenty-six eligible participants with type 1 diabetes were recruited. Among them, 23 live in New York City and 3 are located in other cities in the East, North and Midwest regions of the United States. To protect participants’ privacy, the names of the 3 cities are not reported in this paper. Participants’ characteristics are presented in Table 1.

**Data Collection**

The data presented in this article were collected through one-on-one semi-structured interviews. The first author carried out all of the data collection using an interview guide, which was tested and revised iteratively through a pilot study with 30 diabetes patients; data from the pilot study is not reported in this paper. The major questions

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>n (%)</th>
<th>Characteristic</th>
<th>n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td></td>
<td>Occupation</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>6 (23.1)</td>
<td>Artist</td>
<td>6 (23.1)</td>
</tr>
<tr>
<td>Female</td>
<td>20 (76.9)</td>
<td>Business administrator</td>
<td>5 (19.3)</td>
</tr>
<tr>
<td>Age (years)</td>
<td></td>
<td>Engineer</td>
<td>3 (11.6)</td>
</tr>
<tr>
<td>19-24</td>
<td>1 (3.8)</td>
<td>Student</td>
<td>3 (11.6)</td>
</tr>
<tr>
<td>25-44</td>
<td>8 (30.8)</td>
<td>Lawyer</td>
<td>2 (7.7)</td>
</tr>
<tr>
<td>45-69</td>
<td>12 (46.2)</td>
<td>Accountant</td>
<td>1 (3.8)</td>
</tr>
<tr>
<td>70+</td>
<td>5 (19.2)</td>
<td>Doctor</td>
<td>1 (3.8)</td>
</tr>
<tr>
<td>Years of experience</td>
<td></td>
<td>Home maker</td>
<td>1 (3.8)</td>
</tr>
<tr>
<td>0.5-2</td>
<td>4 (15.4)</td>
<td>Journalist</td>
<td>1 (3.8)</td>
</tr>
<tr>
<td>3-19</td>
<td>8 (30.8)</td>
<td>Nurse</td>
<td>1 (3.8)</td>
</tr>
<tr>
<td>20-39</td>
<td>7 (26.9)</td>
<td>Pharmacist</td>
<td>1 (3.8)</td>
</tr>
<tr>
<td>40-49</td>
<td>6 (23.1)</td>
<td>Teacher</td>
<td>1 (3.8)</td>
</tr>
<tr>
<td>50+</td>
<td>1 (3.8)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
asked in the interview include: (a) What health data about yourself and generated by yourself do you collect? (b) Is this data useful for your diabetes management? If yes, in what way(s)? (c) (If participant mentioned decision making) How do you use this data to make decisions about diabetes? (d) Did you encounter difficulties in this process? (e) What kind of technologies do you think can help you overcome these difficulties?

Data Analysis

All interviews were transcribed and inductive coding was performed for each of the 5 major interview questions. The analysis follows this iterative procedure: The first author (a) coded a sample, (b) discussed the coding with colleagues, (c) revised the codebook and the coding procedure, (d) coded the old sample(s) along with a new sample, (e) repeat the previous two steps until the codebook is no longer updated, which is 8 iterations, and (f) code the other 18 interviews. A second coder coded 5 interviews with high agreement (Krippendorff’s Alpha = 0.817; Cohen’s Kappa = 0.818). The number of participants who mentioned each theme is not reported, as is common in research that uses semi-structured interviews for data collection24. Specifying the number of participants mentioning each theme may not fully represent their exact behaviors.

Table 2 Types of self-generated diabetes data used by participants for decision support

<table>
<thead>
<tr>
<th>Types</th>
<th>Detail</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blood glucose</td>
<td>Blood glucose levels</td>
<td>120mg/dl</td>
</tr>
<tr>
<td></td>
<td>Blood glucose test time</td>
<td>1/2/2012 11:35am (exact); before breakfast; 2h after lunch (relative)</td>
</tr>
<tr>
<td></td>
<td>Blood glucose test location</td>
<td>L3Y (i.e., third finger from thumb on left hand, as described by a participant)</td>
</tr>
<tr>
<td>Other health indicators</td>
<td>Physiological</td>
<td>shortness of breath; fatigue; tingling in extremities</td>
</tr>
<tr>
<td></td>
<td>Psychological</td>
<td>work-related stress; mood changes</td>
</tr>
<tr>
<td>Insulin</td>
<td>Insulin type</td>
<td>NovoLog, Humalog, and Lantis</td>
</tr>
<tr>
<td></td>
<td>Insulin dosage</td>
<td>4 units and half a unit</td>
</tr>
<tr>
<td></td>
<td>Insulin injection location</td>
<td>left buttock; right thigh; arm; belly</td>
</tr>
<tr>
<td></td>
<td>Insulin injection time</td>
<td>9:00pm</td>
</tr>
<tr>
<td>Other treatments</td>
<td>Treatment type</td>
<td>metformin; medication for lupus; cancer treatments</td>
</tr>
<tr>
<td></td>
<td>Treatment dosage</td>
<td>1 pill</td>
</tr>
<tr>
<td></td>
<td>Treatment time</td>
<td>with dinner</td>
</tr>
<tr>
<td>Food</td>
<td>Carbohydrate count</td>
<td>“15 carbs” for a slice of apple</td>
</tr>
<tr>
<td></td>
<td>Portion size</td>
<td>a big apple may be equal to 3 servings of apples</td>
</tr>
<tr>
<td></td>
<td>Type of diet</td>
<td>vegan; paleo; high protein; high fat; low carb</td>
</tr>
<tr>
<td></td>
<td>Types of food</td>
<td>red meat; fish; fruits; grains; vegetables</td>
</tr>
<tr>
<td></td>
<td>Specific food</td>
<td>cereals; cookies; yogurt; cheese; lentil soup.</td>
</tr>
<tr>
<td></td>
<td>Dining time</td>
<td>7:30pm</td>
</tr>
<tr>
<td>Exercise</td>
<td>Exercise type</td>
<td>running; tennis</td>
</tr>
<tr>
<td></td>
<td>Exercise intensity</td>
<td>marathon; first run after recovery from injury</td>
</tr>
<tr>
<td></td>
<td>Exercise time</td>
<td>3pm</td>
</tr>
<tr>
<td></td>
<td>Exercise length</td>
<td>1 hour</td>
</tr>
<tr>
<td>Perceived environment</td>
<td>Seasons</td>
<td>winter; summer</td>
</tr>
<tr>
<td></td>
<td>Weather</td>
<td>hot; sunny</td>
</tr>
</tbody>
</table>
Results

In this section, results for each of the three research questions are presented. First, the types of patient generated health data collected and used by participants for decision support are presented. Next, four decision-making patterns reported by participants are described. Finally, the difficulties that participants experience when engaging in each of the four decision-making patterns are discussed.

Types of Self-Generated Health Data that Patients Collect and Use in Decision Making

Participants in this study report collecting, maintaining, and using a wide range of self-generated health data for making health-related decisions. A list of the various types of self-generated health data that participants use to support decision-making processes is presented in Table 2. Note that not all participants report using all types of data. In other words, participants may use a subset of the types listed in Table 2. Also, participants’ preferences for different data used in decision-making processes may develop and change over time. For example, P19 was diagnosed with lupus shortly before the interview and her new procedure for diabetes decision-making takes into consideration the ways in which lupus treatments would impact her diabetes-related routine.

The types of self-generated health data used by participants for decision-making cover many aspects of their health condition, related treatments, regimens, and personal contexts. The scope of this data sometimes extends beyond that advised by participants’ HCPs. Some participants recall their HCPs asked them to collect blood sugar levels and test time and date. Others add that food diaries detailing the types of food and corresponding portion size are recommended by their HCPs. However, participants were not required to keep records on other types of data that they use for decision making or keep uninterrupted records of the types of data preferred by their HCPs.

Use of Self-Generated Health Data in Decision Making

Participants report four patterns of decision-making processes that rely on the use of self-generated data related to diabetes. These four patterns include: (a) connecting causes and effects – “What caused these health outcomes?”, (b) establishing priorities – “What should I do next?”, (c) negotiating outcomes – “What may I give up?”, and (d) setting visibilities – “What do I want to see?”. In the rest of this section, each of these four patterns will be introduced with detailed descriptions and examples.

Pattern 1: Connecting Causes and Effects

Participants make decisions about future actions by identifying specific factors, based on the data they collect, that positively or negatively impact their health outcomes. In Table 2, blood glucose and other health indicators are considered health outcomes by participants and the other types of self-generated health data (e.g., insulin, food and exercise) are considered causes that may affect health outcomes. Three approaches were taken by participants to connect cause and effect, including observation, statistic testing, and experimentation.

Observation. In all cases of observation, participants observe changes in their self-generated health data by using time-stamps, which allow them to solve problems related to diabetes management. Some of these observations occur over short periods of time. For example, all participants test their blood sugar levels before and after meals. When their test results are unexpected, they estimate or check the carbohydrate content in their meals to find out which items are responsible for causing specific test results. Likewise, for participants who perform a variety of exercises and experience various treatments, they tend to associate unexpected health outcomes with changes in their exercises and treatments occurring immediately before the observed changes in their health outcomes. Participants who observe immediate health outcome changes following activities they perform tend to adjust their activities to amplify positive outcomes and temper negative ones. For example, consuming food with high carbohydrate levels (e.g., orange and potato chips) lead to spikes in some participants’ blood glucose levels immediately afterwards, which makes them reconsider similar food options in the future.

However, not all cause and effect connections are easily observable in the short term. For example, participants complain that injecting insulin at the same site on the body can affect the effectiveness of the insulin, potentially because of the building up of scar tissues. This phenomenon cannot be observed in the short term, and participants often suspect other factors that can have a more direct impact on blood glucose levels (e.g., food consumption and exercise) before moving on to the less obvious factors (e.g., insulin injection site and temperature). In this circumstance, keeping long term records of self-generated health data is a necessity. For example, P2 keeps paper records of his blood sugar levels over the long term and he color-codes these records according to how high the blood sugar levels are when measured. He keeps a large board and pins his records on it so he can easily observe
patterns of codes and colors in his records. When he observes a cluster of certain color, he reads the annotations in these records on the contexts of these records (e.g., food intake, insulin injection location, and finger used for a blood sample) to find out what may have caused these clusters of certain blood sugar levels. Participants who observe cause and effect connections over the long term tend to approach behavior change decisions more cautiously. Some participants reported consulting other sources (e.g., HCPs, support groups, online communities, and medical and health articles) for reassurance before taking actions.

Statistical analysis. Participants who maintain electronic records of their self-generated health data sometimes conduct statistical analyses with their observations in order to locate potential problems. For example, P18 built annotated records of his blood sugar levels with Microsoft Excel spreadsheets. He describes how he uses statistical analysis to identify issues of importance in these records, “I noticed the clustering. I noticed the change on weekends, I noticed the changes on the seasons, because my overall requirements are a little higher in the summertime, and down a little bit in the winter time. Took me years of charts to try to work that out. ...In an effect, it speaks to me. Also in terms of time series. I go back, and then look at it month before, I look at it years before, around this time of year, and I know I’m having a seasonal change. ...This is why a data driven approach was for me.” P18 was confident with his findings because they were based on models using longitudinal data. These findings alleviated P18’s concerns about unexpected changes in his blood glucose levels during the winter and prompted him to monitor his condition more closely and make necessary adjustments to his insulin intake during the same period.

Other participants report benefitting from statistical analysis provided by applications for diabetes management, such as those that work with their continuous glucose monitors. Participants note that these applications usually help identify patterns of changes in health outcomes (e.g., blood glucose levels), but do not associate these outcomes with a wide variety of potential causes (e.g., exercises, stress and temperature), although some connect carbohydrate intake with blood glucose levels. Also, participants complain that these applications require manual entry of nearly all of their self-generated health data, except for blood glucose levels, which are automatically measured and documented by their continuous glucose monitors and related applications. Due to the lack of functionality and the efforts required for manual data entry, participants who report using such applications find their data entry incomplete (e.g., P7 had interim blood glucose records on her continuous glucose monitor) or that their usage declines and stops after a few months. These participants explain that these applications were not very useful for them, but their HCPs can help them with decision-making after reviewing the blood glucose levels they collected over 10 days to 2 weeks before their clinical visits, depending on their HCP’s requirements.

Experimentation. Some participants experiment on themselves to extract the causes of certain health outcomes. Experimentation is usually employed by participants who do not identify relationships between causes and their effects through observation and statistical analysis. An experimentation method reported by participants is to establish a routine of daily activities and test the effect of new behaviors (e.g., eating a new type of food, performing a new physical exercise, changing insulin injection schedule, mixing different types of insulin, using expired insulin, and switching medical devices) on specific health outcomes. Participants also voice that diabetes management is a “noisy system”, and they usually repeat the same experiment multiple times before making a conclusion if they do not consider the changes in their health outcomes large enough to justify the necessary behavior changes.

Pattern 2: Establishing Health Priorities

When cause and effect connections are established, participants plan the steps they will take for future behavior changes. While some changes may be perceived by participants as straightforward (e.g., reducing carbohydrate intake to an amount recommended by HCPs), others may involve multiple processes when the cause and effect relationships are intricate.

Take P16’s treatment of chronic fatigue as an example. P16 was living with cancer and type 1 diabetes, and one of the treatments for cancer made her suffer from chronic fatigue. The symptoms of the fatigue were so severe that she struggled to get out of bed every morning. Every time when she attempted to test her blood sugar, she falls asleep without realizing it. This fatigue, coupled with her unstable blood sugar levels, became dangerous because it increased the risk of her inability to react during a hypoglycemia episode.1 To treat the fatigue, P16 had three options, and could choose any of the options or a combination of the options: (a) change her cancer treatment, (b) change her diabetes treatment (e.g., adjust insulin dosages and dosing methods), (c) change her diabetes regimen (e.g., increase exercises

1 Hypoglycemia: Situations of blood glucose levels dropping below a certain threshold. Hypoglycemia may be more difficult to recover in patients with type 1 diabetes that use insulin. If left untreated, hypoglycemia can worsen and lead to severe health consequences, such as seizures, unconsciousness.
and reduce carbohydrate intake). P16’s health care team offered her a wide range of choices for potential treatments and regimens, but they did not have a clear sense of how effective these changes would be for P16 as an individual patient. She decided to immediately change her cancer treatment; then, she joined a physical training class in the same hospital where her health care team was located. Finally, she changed her diabetes treatments by acquiring a continuous glucose monitor. These changes took course over a 2-year period, and P16’s fatigue was largely reduced.

P16’s decision-making process involves considerations for cause, convenience and cost; she prioritizes different considerations for different reasons, at different times during her decision-making and her resultant behavioral changes. First, she removes the cause of negative health outcomes (e.g., fatigue-inducing cancer treatment). Then, she increases positive health outcomes by participating in convenient behaviors (e.g., exercising). Later, she aims for less convenient and costlier causes of positive health outcomes (e.g., a continuous glucose monitor).

Other participants also report considering cause, convenience and cost, but with different priorities. For example, P13 acquired a continuous glucose monitor shortly after her diagnosis because being able to monitor her blood glucose levels in real time helps her make decisions on micro-adjustments and maintain constantly in-range blood glucose readings. In this case, P13 prioritizes cause and convenience over cost.

**Pattern 3: Negotiating Outcomes**

Participants report that they have specific goals in mind when they make connections between causes and effects and they negotiate between these goals to decide what actions to take. Some of these goals are straightforward health outcomes, such as achieving stable blood glucose levels and not experiencing diabetes complications. While many health outcomes pursued by participants require similar actions (e.g., timely insulin injection around meal time can help type 1 diabetes patients avoid symptoms of hyperglycemia and also reduce the chances of complications in the long run), some health outcomes have competing requirements. For example, P22 describes her fears of hypoglycemic episodes, “I’m not aggressive enough with it [i.e., carbohydrate counting]. I always underestimate. It goes back to my fear of lows.” In this case, P22 believes that maintaining an in-range average blood glucose level increases the chance of hypoglycemia. She negotiates between the immediate and potentially devastating effects of the acute hypoglycemic episodes and the relatively mild but degenerative effects of hyperglycemia. To reduce the chances of the acute episodes, P22 decides to loosen her standards for carbohydrate counting and maintain a relatively high average blood glucose level.

Participants also have other considerations that may compete with their pursuit for better health outcomes. Such considerations reported by participants include quality of life, time, and social life. Participants complain that not being able to enjoy their favorite food affects an important aspect of their quality of life. For example, P6 enjoys a small portion of ice cream every day, but her blood glucose is sensitive to the sugar in the ice cream. In order to manage diabetes, P6 has to cut back on her daily ice cream portion and take extra insulin to cover for the blood glucose spikes. Also, participants found the lack of time due to factors such as the pursuit of a career, caring for family, and traveling a major reason that may compromise their diabetes management endeavor. P8 attempts to maintain the best blood glucose levels with a busy graduate school schedule. Because of her busy schedule, she is not able to guarantee timely food intake, so she keeps snacks at her desk and micro-manages her insulin injection through a pump to obtain the best possible outcomes. Further, participants find that their social lives impact their decision-making processes as well. For example, P17 explains his struggles to negotiate between diabetes management and a normal social life: “...to let down, let's say a friend or not make it to a social arrangement because of a hectic diabetic disruption. It's difficult for someone else to understand, because, in a lot of ways, it's an unreasonable disease. ... I think I have trouble with those moments saying to friends, you know like, it's a long, hard day and I know were aren't right words, but I think it's a psychological weight to kind of carry for this irrational disease that can disrupt at any moment.” Other participants also mention social eating and religious fasting as difficult moments where they must balance healthy social relations and the temptation of food with the possibility of negative health outcomes.

When negotiating between aspects of their lives that compete with health outcomes, rather than giving up or reducing the frequency and intensity of certain actions, participants in this study also report seeking alternative actions that have less impact on their health outcomes. For example, P9 enjoys sweet foods and drinks and finds them to be an important aspect of her quality of life, but they usually create spikes in her blood glucose levels and require a greater insulin dosage. To achieve a balance between quality of life and stable, in-range blood glucose levels, P9 stockpiles Glucerna, a sweet beverage that she discovered would not create sugar spikes that were as severe, and she uses it to substitute her usual breakfast.

**Pattern 4: Setting Visibilities**

When participants use their self-generated health data to make decisions, they have different preferences regarding
the type of data they want to see in different circumstances; they also have different preferences for the quantity and timing of such data.

**Data selection.** All participants do not use all types of data listed in Table 2 for decision-making. Some participants follow their HCPs’ recommendations such as using carbohydrate counting and food consumption to determine insulin dosages, while other participants use other types of data for the process, depending on their daily activities and preferences. For example, P2 performs the same moderate exercises on a daily basis, so he does not consider exercise-related data relevant for decision-making. On the other hand, P12 is a marathon runner with varying exercise schedule, and exercise-related data is a major input for her to make treatment and regimen decisions.

**Quantity and timing.** Participants have different preferences regarding the quantity of data they use for decision-making. Some participants prefer large quantities of self-generated health data because more data means greater possibilities of discovering recurring patterns and hidden connections. On the other hand, participants who do not keep records of their health data (e.g., blood glucose levels) but use it for real time decision support (e.g., test blood glucose to determine the next insulin dosage) tend to prefer smaller quantities of data collected shortly before or during the decision-making episode. The reasons participants give for these preferences are that too much data is overwhelming, and data collected too long before the decision-making episode has little bearing on their ever-changing health statuses (e.g., blood glucose readings).

**Decision-Making Difficulties**

Participants in this study identify various difficulties that they experience with their current decision-making processes and supporting technologies. These findings are reported in Table 3.

The difficulties reported by participants point to three disconnections between their decision-making processes and the support they receive, including: (a) data entry does not reflect data used for decision making, (b) data analysis is difficult to carry out and results are hard to comprehend, and (c) output of analysis is difficult to understand and put to practice. The four decision-making patterns are influenced by different disconnections, with (a) connecting causes and effects and negotiating outcomes mainly caused by data entry and analysis problems, and (b) establishing priorities and setting visibilities mainly related to issues with data output. Decision support technologies that address these three disconnections will offer better utilities and experiences for patients with type 1 diabetes. The next section details some of the ways that these disconnections can be addressed in decision-support technology design.

<table>
<thead>
<tr>
<th>Decision-making patterns</th>
<th>Difficulties</th>
<th>Detail</th>
</tr>
</thead>
<tbody>
<tr>
<td>Connecting causes and effects</td>
<td>Identify causal factors</td>
<td>Unable to locate causes of certain health outcomes.</td>
</tr>
<tr>
<td></td>
<td>Identify intensity of causal factors</td>
<td>Difficult to determine how much influence each causal factor have on certain health outcomes.</td>
</tr>
<tr>
<td></td>
<td>Identify solutions</td>
<td>Unable to find viable ways to address the causal factors for negative health outcomes.</td>
</tr>
<tr>
<td>Establishing priorities</td>
<td>Stay accountable</td>
<td>Difficult to stick to the plan.</td>
</tr>
<tr>
<td>Negotiating outcomes</td>
<td>Identify current trends</td>
<td>Unable to monitor the trending of current health outcome indicators.</td>
</tr>
<tr>
<td></td>
<td>Estimate individualized outcomes</td>
<td>Difficult to predict health outcomes based on personal health history and actions taken.</td>
</tr>
<tr>
<td>Setting visibilities</td>
<td>Too much information</td>
<td>Irrelevant information causes fatigue to notifications and negative emotions.</td>
</tr>
<tr>
<td></td>
<td>Inconvenient access</td>
<td>Medical terms that are difficult to understand, tools that have awkward displays, not sure how to interpret the data, not sure what questions to ask.</td>
</tr>
</tbody>
</table>

Table 3 Participants’ difficulties with using self-generated health data for decision-making
Discussion and Implication

This paper describes the multiple types of self-generated data that type 1 diabetes patients use in their everyday decision-making processes, the ways in which they use this data to make decisions, and the difficulties they experience in these processes. These results point to multiple venues for the design and redesign of decision support technologies for patients with type 1 diabetes that will not only provide direct support for patients, but will also harness patient generated data for clinical use. In the clinical environment, individual patients’ longitudinal data that is used by patients in their decision-making processes can enrich and enhance the data used by clinicians for medical decision making in clinical care. Collections of this data, gathered as part of patients’ daily routines, may also enable the identification of cause and effect connections not commonly considered in clinical settings, as well as potential treatments and regimen recommendations used by patients.

Patient-Centered Data Entry

People living with type 1 diabetes use a wide range of self-generated health data for decision-making. The different types of data they generate, collect, store, and use extends beyond HCPs’ requirements. Some of this data is patients’ perception of themselves and their environment, such as mood and weather. The nuances of health data captured by patients and used by them in everyday settings for decision-making are usually not incorporated in the design of decision support systems. To facilitate using patient generated data to support patients’ decision making, allowing patients to define the types of data to enter into their decision support technologies is necessary. For example, P2 considers blood glucose testing location and insulin injection location important for him to interpret his blood sugar readings, especially when he compares his readings across testing and injection locations. In contrast, P11 believes that stress is a major cause of her diabetes and is highly conscientious about work-related stress levels. Considering these highly specialized needs, it is not sufficient to offer these two participants a standard diabetes decision support application that allows them to document their carbohydrate counts and insulin dosages. A better, more flexible solution is to allow patients to define data types that are beyond the standard setup, enabling them to name new data entries (e.g., variables), their measurements (e.g., categorical or continuous), and their entry method (e.g., choose from preset options or enter free text) as well as incorporate this data to support their decision-making processes.

Furthermore, the larger amount of data collected by patients in the home points to the insufficiency of patient data routinely collected in the clinical setting. If patient-generated data is incorporated into clinical records, it may be useful for improving patient-provider communication and clinical care by offering HCPs data-driven approaches from patients’ perspective, rationales of patients’ activities, and possibilities of discovering new working regimen initiated and tested by patients.

Automated Data Entry and Analysis

Participants in this study expressed their needs for decision support systems that can automatically identify potential causes for changes in their health outcomes. Some participants also experienced difficulties with identifying solutions for certain health-related problems and desire decision support technologies that will aid them in finding solutions for these problems. Some participants mentioned that existing technologies (e.g., continuous glucose monitors) can address some of their needs, but others complained that the automated functionalities of these technologies are limited to blood glucose related tracking and recommendations. There is currently not enough support to automate the collection of other types of self-generated data (e.g., diet and exercise). When possible, incorporating existing self-generated data from other devices using open APIs, such as step-tracking software on many mobile phones or weight data collected by smart scales, may be one way to aggregate and store self-generated data for patient use.

In an effort to automate the data entry and decision-making processes, the medical informatics community has made recent developments in consumer-oriented technologies, such as a diabetes wound management app based on image recognition technologies and a meal-camera system used in telemedicine for diabetes patients. More research is needed to facilitate communication between applications that automate the entry and analysis of different data to offer more coherent and meaningful consumer experiences.

Individualize Data Analysis

Findings in this paper demonstrate that type 1 diabetes patients’ decision-making processes are highly personal. Some decisions are made within the participants’ resource constraints and their personal preferences. These constraints and preferences differ between patients and may change over time. This issue points to the need to not only use patients’ personal health data for individualized care, but also to allow patients to shape the underlying logic of how this data can be viewed, analyzed, and used. For example, a piece of decision support technology that
models a type 1 diabetes patient’s blood glucose levels on weekly, monthly, and yearly basis may not clearly show patterns for all three time intervals, particularly if the patient has a non-standard work schedule. In this case, allowing the patient to set her preferred time interval to match her specific schedule may help the technology determine if the patient’s work schedule plays a significant role in affecting her blood glucose levels.

**Humanize Decision Support Output**

Some participants voice their frustration and confusion when engaging in their existing decision support aids for their lack of intuitive interfaces (e.g., continuous glucose monitors and their corresponding applications and carbohydrate counting books). Participants report that they prefer decision support technologies designed for laypeople, who do not have specialized knowledge in medicine or expertise with using search engines. Participants would prefer to communicate with decision support technologies as if they were human (e.g., both the patient and the technology can use and understand plain, natural language), and would like these technologies to help hold them accountable for their health-related decisions (e.g., recording blood glucose test results and maintaining diet) like HCPs would.

These requirements point to three potentially useful aspects of decision support technologies for patients with type 1 diabetes: (a) the ability to translate between natural language and search queries; (b) the ability to translate between medical terms and lay terms; and (c) the ability to socialize with the patient at checkpoints (e.g., making sure that the patient keeps her daily carbohydrate count at under 200 grams). This study does not yield enough information on how to determine what incidents qualify as checkpoints or how frequent these checkpoints should be. The literature on using telemedicine reminders to support self-management and promote adherence in people with chronic conditions does not agree on how frequently reminders should be sent, and there is not enough research on the detection and determination of checkpoints to date. For example, Tao and colleagues argued that the frequency of electronic reminders does not seem to moderate medication adherence. Eleches and colleagues found that the frequency of reminders has an impact on patients’ adherence to antiretroviral therapy (i.e., an HIV treatment) and that weekly reminders seem to have a more significant positive impact than daily reminders. Research also suggests that allowing teenage asthma patients to set the frequency of reminder text messages may enhance patients’ self-management. Therefore, a future direction for this work is to determine when and how to build checkpoints into these technologies. It is possible that, like the recommendation to individualize data analysis, checkpoints should also be a setting that patients can customize to their own preferences.

**Conclusion**

This qualitative study explores type 1 diabetes patients’ decision-making behaviors and designs that can support these behaviors with a focus on patient generated data. The findings have implications for decision support for patients in their everyday lives, and may also extend to the clinical space. As this is a qualitative study with a purposive sample, the results are not generalizable but may be transferrable to other patient populations where decision-making is frequent (e.g. coronary heart disease, lung cancer) or when patients may generate and track data beyond their HCPs’ requirements (e.g., pregnancy). More research is needed to investigate the effects of the design recommendations and explore certain aspects of decision support technology design (e.g., the timing of checkpoints to promote adherence).

**References**

7. Quinn CC, Shardell MD, Terrin ML, Barr EA, Ballew SH, Gruber-Baldini AL. Cluster-Randomized Trial of a Mobile Phone Personalized Behavioral Intervention for Blood Glucose Control. Diabetes Care. 2011 Sep


Causal Phenotyping for Susceptibility to Cardiotoxicity from Antineoplastic Breast Cancer Medications

Deyu Sun, PhD1, Gyorgy J. Simon, PhD2,4, Steven Skube, MD3, Anne H. Blaes MD4, Genevieve B. Melton, MD, PhD2,3, Rui Zhang, PhD2,5
1School of Statistics, 2Institute for Health Informatics, 3Department of Surgery, 4Department of Medicine, 5College of Pharmacy, University of Minnesota, MN

Abstract
Cardiotoxicity is a relatively common and particularly important adverse event caused by chemotherapy for breast cancer patients. Typical associative phenotypes, such as risk factors associated with diabetes, can often be detected solely based on the data elements existing in electronic health records; however, causal phenotypes, such as risk factors causing cardiotoxicity, require establishing causation between chemotherapy and determining new heart disease, and cannot be directly observed from EHR. We propose three phenotyping algorithms to assess breast cancer patients' susceptibility to cardiotoxicity caused by five first-line antineoplastic drugs: (1) causal phenotype model to predict the patients' risk of cardiotoxicity as the difference between the heart disease risks with exposure and non-exposure to the drugs; (2) regular predictive model; (3) combined predictive model of the above two models. Concordances for three methods were 0.60, 0.62, and 0.68. When considering all exposed patients, concordances were 0.66, 0.58 and 0.65 at 280 days after treatment. The study demonstrates the potential utility of causal phenotyping.

Introduction
Phenotyping is a quintessential tool for constructing patient cohorts that consist of patients with a set of desired physiologic or other physical characteristics (referred to as phenotype) [1-3]. Finding such cohorts is helpful for different applications such as clinical trials or intervening for direct patient care for improved population health. Given an intervention that potentially benefits patients suffering from a particular condition (phenotype), phenotyping algorithms can be used to search large electronic health record (EHR) databases to identify patients who very likely have a particular condition (cases) and also patients who very likely to not have that particular condition (controls).

It has been abundantly demonstrated that machine learning methods can be applied as techniques to help solve the phenotyping problem [4]. From a machine learning perspective, phenotyping is a detection problem, where the machine learned model associates EHR data elements with the condition in question. This approach can be very successful owing to the fact that many diseases have associated data elements that are specific to the disease. For example, the eMERGE phenotyping algorithm for diabetes [5] utilizes factors directly associated with the disease such as diagnosis codes of diabetes, blood sugar levels, and diabetes medications for detecting the disease in the EHR. For this use case, diabetes diagnosis codes are only (or mostly in the case of drugs) given to patients with diabetes and high blood sugar levels are one of the main defining characteristic of diabetes. We refer to phenotypes that can be detected directly with associated EHR data elements as associative phenotypes.

In this work, we propose to explore the concept of causal phenotypes. In contrast to associative phenotypes, causal phenotypes do not have directly associated EHR data elements. For example, cardiotoxicity, which is defined as heart disease due to chemotherapy, is not directly associated with any diagnosis code, lab test, procedure or treatment. From the EHR, we can detect that the patient has received the chemotherapy that may trigger and cause heart disease. We can also identify that the patient has suffered an incident (new) of heart disease after the chemotherapy, but we cannot determine from the EHR whether these two elements have a causal relationship since the patient may have suffered from heart disease even if she had not received chemotherapy.

The eMERGE library of phenotypes [6] does provide an example of causal phenotype: drug induced liver injury. Briefly, the algorithm identifies patients who had normal liver function before taking a drug that may cause liver injury and have abnormal liver function after taking the drug. This algorithm implicitly attributes all deterioration of liver function to the drug. While this algorithm may be mostly correct in the case of liver injury (patients with chronic kidney disease are excluded), such a strategy may also incorrectly attribute liver injury to a drug in cases where patients have polypharmacy or undocumented exposure to other hepatotoxic elements. Also, this approach will certainly be incorrect for some cases of cardiotoxicity. This is because numerous conditions (e.g., ischemia, neutropenic fever,
gradual decompensation) can cause heart disease in a patient outside of chemotherapy. To ascertain cardiotoxicity from chemotherapy, we need to quantify the causal effect of chemotherapy on heart disease.

A second key difference between causal cardiotoxicity and the typical disease phenotyping algorithm is that we are interested in understanding a given patient’s susceptibility to the disease. While many of the existing phenotyping algorithms are particularly apt for detecting existing disease in patients after diagnosis, they tend not to be useful for quantifying susceptibility to disease before or as it develops. Quantifying susceptibility is essentially risk prediction, in which the existing machine learning phenotyping algorithms could be easily re-trained to perform; however, predicting cardiotoxicity requires prediction of the causal effect of an intervention, which is a much more challenging task. This is the task we set out to explore and ultimately solve in this paper.

In our work, we are focusing on cardiotoxicity in breast cancer. Breast cancer is the most common cancer in women worldwide [7]. Chemotherapy is an essential part of the treatment regimen for many patients diagnosed with breast cancer [8]. Chemotherapy can be extremely beneficial to patients with breast cancer by improving survival and decreasing recurrence; however, its utility can also be limited because of adverse events and side effects [8]. Cardiotoxicity is a particularly important and relatively common adverse event caused by chemotherapeutic agents used for breast cancer. Cardiotoxicity is a general term that signifies toxicity affecting the heart [8]. In other words, cardiotoxicity is heart disease caused by chemotherapy. In the Food and Drug Administration Structured Product Labels (FDA SPL), two of the main chemotherapy drugs used in breast cancer treatment have black box warnings of cardiotoxicity. These drugs are doxorubicin (Adriamycin) and trastuzumab (Herceptin or Ado-Trastuzumab Emtansine) [9-11]. Doxorubicin causes Type I chemotherapy-induced cardiotoxicity (cell death, cumulative, permanent) while trastuzumab causes Type II chemotherapy-induced cardiotoxicity (cell dysfunction, not cumulative, reversible) [8]. Risk factors and predictors of which patients will develop cardiotoxicity are still lacking in both doxorubicin and trastuzumab [12-13]. Also, epirubicin (Ellence) and pertuzumab also have black box warnings of cardiotoxicity [14-15]. All of these agents associated with chemotherapy-induced cardiotoxicity were included in this study.

The manifestation of cardiotoxicity in cancer patients is varied and shares similar symptoms as heart disease. Presentation can be acute, subacute, or chronic with the timeline ranging from hours to years after chemotherapy administration [8, 12]. Acute changes associated with cardiotoxicity include tachycardia, arrhythmias, electrocardiogram (EKG) changes, and changes in markers for cardiac injury (troponin). Subacute and chronic changes are often related to decreased ventricular function as noted on echocardiogram, symptoms of congestive heart failure (CHF), and issues with arrhythmias. Less commonly, pericarditis and myocarditis can occur [8, 12, 13]. Patients at risk for chemotherapy-induced cardiotoxicity are often co-managed by cardiologists and followed closely through imaging studies and outpatient assessments. There are also several approaches taken to reduce the potential for cardiotoxicity and to reduce the effects of cardiotoxicity; however, this is more individualized as there is no consensus guideline for the treatment or prevention of chemotherapy-induced cardiotoxicity [8].

We propose three phenotyping algorithms to assess patients' susceptibility to cardiotoxicity in breast cancer caused by five first-line breast cancer drugs: ado-trastuzumab emtansine, trastuzumab, doxorubicin, epirubicin, and pertuzumab. First, we present the casual phenotyping algorithm. This algorithm aims to estimate the causal effect of chemotherapy on heart failure. The algorithm requires labels of heart failure but does not require gold-standard labels of cardiotoxicity. Heart failure is an associative phenotype and thus determining whether a patient has heart disease from EHR is relatively simple. On the other hand, obtaining gold-standard cardiotoxicity labels requires that we determine whether the treatment is causally related to the heart disease and is thus difficult. The second method is the “regular” Cox model. If we have gold-standard labels, we can construct a regular predictive model (Cox proportional hazard model) to directly estimate the patients' risk of cardiotoxicity based on the available physiological data. Third, we can combine the above two models to incorporate the estimated causal effect into the Cox model.

We evaluated the method on retrospective cohort data collected from University of Minnesota Medical Center (UMMC) and a set of gold standards generated by a clinician. We found that the causal phenotype model works reasonably well without being trained with gold-standard labels and the third model which combined the causal phenotype model and regular predictive model works the best. The ability of the causal phenotype to achieve reasonable performance without gold-standard labels is particularly important. Generally, machine learned models cannot be simply ported from one site to another; they typically require re-training [16], which in turn, requires gold-standard labels. When obtaining gold-standard labels is difficult, our proposed model's ability to operate without gold-standard labels is particularly valuable. When gold standard labels are available, the proposed causal effect improves the performance of the machine learned phenotyping algorithm trained on physiological data.
Methods

Data Collection

In this retrospective cohort study, 4084 patients diagnosed with breast cancer and treated with chemotherapy at UMMC from 2000 to 2017 were included. The risk factors of heart failure, including age, body mass index (BMI), high-density lipoprotein (HDL), low-density lipoprotein (LDL), hemoglobin A1c (HbA1c), troponin, B-type natriuretic peptide (BNP), coronary artery disease (CAD), systolic and diastolic blood pressure (systolic BP and diastolic BP), and heart rate (HR), were collected as a baseline prior to the start of chemotherapy treatment. We also included infection markers such as white blood cell count (WBC) and temperature. For patients who took the breast cancer drugs with cardiotoxicity warnings, the baseline is when the first dose of the drug was taken; for patients who received other interventions, the baseline is the first time the breast cancer was diagnosed, because the date of the intervention is not always known. The endpoint is the first time the patients were diagnosed with heart disease related to cardiotoxicity; for patients who did not have symptoms of cardiotoxicity, the endpoint was arbitrarily chosen (12/31/2016). The characteristics of the cohort are shown in Table 1.

Table 1. Characteristics of two populations at baseline.

<table>
<thead>
<tr>
<th>Patients Exposed to the Drugs of Cardiotoxicity Warnings (N=653)</th>
<th>Patients Without Exposure to the Drugs (N=3431)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Data</td>
</tr>
<tr>
<td># of patients with heart diseases related to cardiotoxicity</td>
<td>80 (12.25%)</td>
</tr>
<tr>
<td>Age (years)</td>
<td>54.11 (12.59)</td>
</tr>
<tr>
<td>BMI (kg/m²)</td>
<td>28.47 (6.38)</td>
</tr>
<tr>
<td>HDL (mg/dl)</td>
<td>49.05 (12.07)</td>
</tr>
<tr>
<td>LDL (mg/dl)</td>
<td>105.70 (36.50)</td>
</tr>
<tr>
<td>HbA1c (%)</td>
<td>6.23 (1.20)</td>
</tr>
<tr>
<td>Troponin (µg/l)</td>
<td>0.03 (0.06)</td>
</tr>
<tr>
<td>BNP (pg/ml)</td>
<td>NA</td>
</tr>
<tr>
<td>WBC (10⁹/l)</td>
<td>6.60 (2.46)</td>
</tr>
<tr>
<td>CAD</td>
<td>4 (0.61%)</td>
</tr>
<tr>
<td>Systolic BP (mmHg)</td>
<td>124.7 (16.60)</td>
</tr>
<tr>
<td>Diastolic BP (mmHg)</td>
<td>73.87 (11.09)</td>
</tr>
<tr>
<td>Temperature (F)</td>
<td>97.94 (0.75)</td>
</tr>
<tr>
<td>Heart Rate (beats/minute)</td>
<td>80.74 (12.94)</td>
</tr>
</tbody>
</table>

* Data are means (SD) or numbers (%).

We used two-sample t-test to compare the risk factors in two groups. No significant differences in risk factors between two groups of patients were detected at the significance level of 0.05. No patient who took the drugs with cardiotoxicity warnings had prior BNP test results. As demonstrated in Table 1, the rate of missingness for a number of these variables is high. A high rate of missingness is a common and difficult problem for many secondary uses of EHR data. For instance, Hu et al. compared various imputation methods for EHR data; in their analyses, imputing the mean values of the cohort while adding an indicator variable signaling imputation (“dummy variable”) not only has acceptable performance but also can correct bias due to higher rates of missingness [17]. Here, we treat the missing values in the risk factors by imputing the mean values of the cohort, and an indicator variable signaling imputation was also added.

Extraction of Heart Diseases Related to Cardiotoxicity

Types of heart disease that may be caused by each breast cancer drug of interest are recorded in the black box warning section of the package inserts or FDA SPLs. Using a previously described approach, we first extracted all diseases from the black box warnings through natural language processing (NLP) algorithms [18]. Then, the clinicians reviewed the extracted diseases and assigned the ICD-9-CM as well as ICD-10-CM codes to each one (Table 2).
Acquisition of Cardiotoxicity Gold Standards

Rather than review all charts including those with no indication of heart disease, the charts of the 80 patients who took the drugs with cardiotoxicity warnings and showed symptoms of heart failure were manually reviewed and adjudicated by a physician (Table 1). Clinical notes, laboratory data, and imaging were reviewed by the clinician. Among 80 patients (with heart failure), 42 patients were found to have chemotherapy-related cardiotoxicity. Common examples of heart failure that is not chemotherapy-related included pulmonary complications with symptoms incorrectly classified as cardiac, cardiac arrhythmias thought to be secondary to acute illness or decompensation rather than chemotherapy, and transient post-operative cardiac issues thought to be unrelated to chemotherapy. The outcome of the physician’s adjudication served as a gold standard to evaluate our phenotyping models.

Figure 1 shows the proportion of patients that truly have chemotherapy-related cardiotoxicity among those with heart failure. We can see that the heart disease that developed 280 days after the initial chemotherapy tends to be more related to the breast cancer treatment (i.e., heart disease before 280 days is potentially more often caused by factors other than the chemotherapy).

Figure 1. The proportions of cardiotoxicity over time. Each circle estimates the proportion of the patients developing symptoms that have cardiotoxicity at a certain time since baseline. The red curve is the estimated trend by fitting the data points. The dashed horizontal line indicates the average proportion (0.525 or 42/80).
Proposed Phenotyping Methods

We designed three phenotyping methods, one of which does not use the gold-standard labels, and then evaluated each algorithm’s performance in terms of concordance. Concordance is the probability that in a randomly chosen pair of patients, where one developed cardiotoxicity earlier than the other (or the other did not develop it), the one with earlier cardiotoxicity has higher risk. We used bootstrapped cross-validation with 100 iterations to estimate the concordance and its sampling distribution.


The Causal Phenotyping method relies on the Neyman potential outcomes framework [19] to estimate the excess risk of heart failure attributable to the chemotherapy agents in this study. In the potential outcome framework, we need to estimate the patients’ risk of developing heart failure when they receive chemotherapy ($\hat{Y}_1$) and the risk of heart failure when they do not receive chemotherapy ($\hat{Y}_0$). For patient $i$, the causal effect of chemotherapy is the difference between these two risks $\hat{Y}_{1i} - \hat{Y}_{0i}$.

As shown in Figure 2, we compute $\hat{Y}_{0i}$ using the patients who did not actually receive chemotherapy. We fit a Cox proportional hazard model (denoted by $M_0$ in Figure 2) predicting heart failure. These heart failure events are clearly not cardiotoxicity because the patients lack exposure to the potentially cardio-toxic agents. We then use this model to make predictions for patients who actually received chemotherapy, obtaining their expected number of (cumulative hazard of) heart failure by the end of the study.

We compute $\hat{Y}_{1i}$ by fitting a Cox proportional hazard model (denoted by $M_1$ in Figure 2) to the patients who received the potentially cardio-toxic chemotherapy agents and make predictions for these patients in a cross-validated fashion using the jackknife procedure (also known as leave-one-out cross-validation). This prediction is the cumulative hazard (expected number) of heart failure events until the end of the study under chemotherapy.

The difference between the two risk estimates is the excess risk of heart failure conferred on the patient by the chemotherapy agent. We call this quantity the causal effect. With higher causal effect, heart failure is more likely to be a result of chemotherapy than an independent reason (e.g. ischemia).

Notice that this procedure does not require gold-standard labels of cardiotoxicity. We only used the gold-standard labels for evaluating the performance of the method.

![Figure 2](image.png)

**Figure 2.** Flow chart of causal phenotype model. $M_0$ is the Cox proportional hazard model fitted for the patients who did not receive chemotherapy drugs with cardio-toxic side effect; $M_1$ is the Cox proportional hazard model fitted for the patients who received chemotherapy drugs with cardio-toxic side effect.

2. Regular predictive model.

The regular predictive model is the Cox proportional hazard model built on the population with exposure to the potentially cardio-toxic chemotherapy agents using physiological variables as predictors and the gold-standard labels as training labels.
3. Combined predictive model.

The combined predictive model integrates the above two models together by adding the causal effect acquired by the causal phenotype model as an additional predictor to the second model. We expected the combined model to take the advantages of both models and perform best.

Results

We first estimated the performances of the model M₀ and M₁ by cross validation. M₀ and M₁ are Cox proportional models which predict the expected number of heart disease events for any individual from the two groups, i.e. without exposure to the drugs and with exposure, respectively, after almost infinite time (10,000 days since the baseline). The concordance of the model M₀ and the model M₁ are 0.76 and 0.63, respectively. Cardiotoxicity can be more of a chronic disease than acute heart disease that can be well predicted by physiological data at baseline. The model built on the population exposed to the chemotherapy mixed the risks of both cardiotoxicity and acute heart diseases. Thus, the performance of the model M₁ is not as good as that of M₀.

Due to differences in baseline characteristics and missingness patterns of information between the exposed and unexposed populations, the distributions of the predicted risks by M₀ and M₁ have similar shapes, but different scales and a minor shift between the modes of the two distributions is also visible. (Figure 3 depicts the two distributions.) This necessitates the recalibration of M₀, which we carried out by a linear transformation of the predictions of M₀ to match the predictions of M₁ by having the same mean and scale.

![Different Predicted Risks, Before Re-calibration](image)

**Figure 3.** The difference in the predicted risks of heart diseases by using M₀ and M₁.

As discussed in the Methods section, the difference in the predicted risks between M₁ and M₀ on the same individual is the net, causal effect of exposure to the drugs on the risk of cardiotoxicity.

Figure 4 shows the concordances of three models estimated by bootstrapped cross-validation. We estimated both the overall performance and the long-term (>=280 days) performances. We can see that the combined predictive model has the best overall performance as well as long-term performance (0.68 and 0.65, respectively), while causal phenotype model has comparable long-term performance to the combined predictive model (0.66 vs 0.65).
In this paper, we proposed the concept of causal phenotyping, where the phenotype is defined as a causal effect of a condition on another. Specifically, we constructed phenotyping algorithms for cardiotoxicity caused by certain breast cancer chemotherapy agents. Cardiotoxicity cannot be directly observed from the her. We can observe directly that the patient has received chemotherapy utilizing the agents in this study and we can also observe that the patient has developed heart failure, but we cannot observe whether the heart disease is a direct consequence of the chemotherapy. We used causal inference to ascertain the degree to which chemotherapy contributed to the heart failure.

In particular, we built three different models. First, we built a causal model, which simply quantifies the causal effect of chemotherapy on heart failure risk for each patient. We can then rank the patients on their personal estimate of causal effect: the patients with the highest causal effect are most likely to develop cardiotoxicity (as opposed to heart failure not attributable to chemotherapy). Overall this model has (median) concordance of .60 when we consider all patients who received chemotherapy and .66 when we only consider long-term (“chronic”) cardiotoxicity. We defined long-term as ‘at least 280 days after treatment initiation’.

**Discussion**

![Figure 4](image)

**Figure 4.** Concordances estimates by bootstrapped cross validation using: (a) all patients and (b) patients with at least 280 days follow-up (patients with heart failure within 280 days are excluded).
Not only does this model offer reasonable performance, but it also does not require gold standard cardiotoxicity labels for training; it only requires heart failure labels, which can be obtained inexpensively with no assistance from clinicians. In contrast, obtaining gold-standard cardiotoxicity labels requires clinicians’ adjudication for each patient, making it labor-intensive and expensive. This is particularly beneficial in light of the observation that machine learned models often require retraining when they are ported from one healthcare provider site to another.

For comparison, we also constructed a “regular” Cox model, using gold-standard cardiotoxicity labels (and time to cardiotoxicity) as outcome. This model was constructed using physiological variables as predictors and achieved superior performance on the entire population as compared to the causal phenotyping algorithm: a median concordance of .63 as opposed to the causal phenotyping algorithm’s .60. Its performance, however, dropped to .58 when we only considered long-term cardiotoxicity.

The key difference between the causal phenotyping and the “regular” Cox model is that causal phenotyping is more sensitive in predicting long-term cardiotoxicity, while the “regular” model built on baseline physiological data is more sensitive towards short-term cardiotoxicity. Cardiotoxicity (“acute cardiotoxicity”) can happen immediately upon the initiation of the chemotherapy, but typically occurs on the longer scale (“chronic cardiotoxicity”) presumably through accumulation of damage. Our Figure 1 supports this common belief. It is interesting that the “regular” model could predict the short-term cardiotoxicity from physiological data, which could suggest that subclinical manifestation of the disease may have been present at the initiation of the treatment.

We also constructed a “combined” model, which uses both the causal effect and patients’ physiological data. As expected, this model utilized the physiological data to predict short-term cardiotoxicity and successfully utilized the causal effect towards predicting long-term cardiotoxicity. Consequently, the “combined” model achieved the highest median concordance of .68 on the entire set of chemotherapy patients; and achieved similar performance to the causal phenotyping algorithm on the long-term cardiotoxicity (median concordance of .66).

We wish to remind the readers that the proposed models are predictive models as opposed to the typical detection models. Predictive models utilize risk factors that are present before the disease manifests itself clinically, while detection models utilize data elements that are consequences of the diagnosis and subsequent treatment of an already existing disease. Therefore, the performance of predictive such models is necessarily much lower than that of detection models. Concordance of .6 to .7 is statistically significantly non-random but is a testament to the limits of predictability. Examples of heart failure in this study not due to chemotherapy include ischemia, rapid deterioration secondary to some other condition, misdiagnosis (patient did not actually have heart failure) and neutropenic fever. Given that the physiological data is collected at the initiation of the chemotherapy, we can possibly detect the incorrect diagnosis and ischemia, but we certainly cannot detect neutropenic fever or rapid deterioration. We did include WBC as an indication of possible immunosuppressive treatment, but baseline levels of WBC and temperature have little relation to neutropenic fever months or years later.

The models introduced in this paper mainly used the risk factors extracted from structured data. Based on guidelines and knowledge of domain experts, in the future we will include more potentially confounding risk factors extracted from both structured and unstructured data, such as smoking status and alcohol use, ejection fraction, dose of chemotherapy drugs taken, history of chest-directed radiotherapy, history of anthracycline use, history of diabetes, and history of related heart diseases. Lastly, since collecting gold standard values requires intensive chart review and is time-consuming, we plan to recruit more physician domain experts to enhance both the quantity and quality of gold standard acquisition, and will include inter-rater reliability as a means to demonstrate internal validation of the gold standard.

**Conclusion**

In this paper, we proposed the concept of causal phenotyping in the context of assessing breast cancer patients’ susceptibility to cardiotoxicity as a result of undergoing chemotherapy utilizing certain potentially cardio-toxic agents. The proposed method achieved reasonable predictive performance even though it does not require gold-standard cardiotoxicity labels for training. Its performance for predicting longer-term (at least 280 days after the initiation of chemotherapy) cardiotoxicity was higher than that of a Cox model trained on the gold-standard labels. The Cox model, in turn, had superior performance in predicting shorter-term cardiotoxicity (less than 280 days after the initiation of chemotherapy). Unsurprisingly, combining these two models achieved the overall best performance, since the model successfully predicted both short- and long-term cardiotoxicity. Training the combined model, however, requires gold-standard labels.
Acknowledgement

This research was supported and publication costs were covered by the Agency for Healthcare Research & Quality grant (#1R01HS022085) (GBM), National Center for Complementary & Integrative Health Award (#R01AT009457) (Zhang), and the University of Minnesota Clinical and Translational Science Award (#8UL1TR000114) (Blazer). The authors thank Fairview Health Services for support of this research.

References


A Multi-Task Framework for Monitoring Health Conditions via Attention-based Recurrent Neural Networks

Qiuling Suo, Fenglong Ma, Giovanni Canino, PhD, Jing Gao, PhD, Aidong Zhang, PhD, Pierangelo Velti, PhD, Agostino Gnasso, MD

1Department of Computer Science and Engineering, University at Buffalo, NY, USA
2Department of Surgical and Medical Sciences, Magna Graecia University, Catanzaro, Italy
3Metabolic Diseases Unit, Department of Clinical and Experimental Medicine, Mater Domini Hospital, Magna Graecia University, Catanzaro, Italy

Abstract

Monitoring the future health status of patients from the historical Electronic Health Record (EHR) is a core research topic in predictive healthcare. The most important challenges are to model the temporality of sequential EHR data and to interpret the prediction results. In order to reduce the future risk of diseases, we propose a multi-task framework that can monitor the multiple status of diagnoses. Patients’ historical records are directly fed into a Recurrent Neural Network (RNN) which memorizes all the past visit information, and then a task-specific layer is trained to predict multiple diagnoses. Moreover, three attention mechanisms for RNNs are introduced to measure the relationships between past visits and current status. Experimental results show that the proposed attention-based RNNs can significantly improve the prediction accuracy compared to widely used approaches. With the attention mechanisms, the proposed framework is able to identify the visit information which is important to the final prediction.

1 Introduction

Disease monitoring is often limited by physician experience, test time, economic barriers and so on. The Electronic Health Record (EHR), which consists of longitudinal health information of patients, is a valuable source for exploratory analysis to monitor diseases and assist clinical decision making. However, due to the complexity of EHR data, the efficient mining of EHRs is not trivial. Firstly, EHR data is heterogeneous which contains various types of features. For example, type of visit is a categorical feature while body mass index is continuous. In addition, some features are static through the lifetime while some change dynamically. Models should be able to capture the essence of heterogeneous features. Secondly, the data is inherently sparse and noisy, due to patient’s irregular visits, absence of tests, and incomplete recording, etc. Thirdly, result interpretation in healthcare applications is essential, and the lacking of interpretability often hinders the adaption of models in clinical settings. Thus, how to correctly model heterogeneous and sparse EHR data and reasonably interpret the prediction results is a challenging problem for disease prediction.

Recent work has made rapid progress in utilizing EHRs for predictive modeling tasks in healthcare, including predicting unplanned readmission, early prediction of chronic disease, adverse event detection and monitoring disease progression. In these settings, the EHRs are typically represented as temporal sequences of medical visits, and each visit contains a set of objects (such as diagnosis and procedure codes). The main idea is to learn a good representation of a patient’s historical health information, in order to improve the performance of the prediction for future risks. To capture the progression of the patient’s health status, much effort has been made on regression models and Markov models. However, these models cannot take into account the long-term dependencies of diagnoses, which may miss several severe symptoms in the past and reduce the performance of disease monitoring.

In order to model the dependencies of diagnoses, deep leaning techniques, such as recurrent neural networks, can be employed. Recent work shows that deep learning can significantly improve the prediction performance. To handle the temporality of multivariate sequences, dynamically modeling the sequential data is necessary. Recurrent neural networks (RNNs), in particular Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU), have achieved state-of-the-art performance in handling long-term dependencies and nonlinear dynamics. Taking advantage of the capability of RNN in memorizing historical records, multiple recent models based on RNNs are employed for deriving accurate and robust representations of patient visits. The work by Lipton et al. applies LSTM to a multilabel classification task for diagnosing multiple diseases in the future, and the contemporary work by Choi et
al.\textsuperscript{11} applies GRU to predict codes for subsequent visits. Both of them show the efficacy of basic RNN models in modeling longitudinal healthcare data. Among the state-of-the-art models, RETAIN\textsuperscript{12} adopts a temporal attention generation mechanism to learn both visit level and code level weights; GRAM\textsuperscript{13} is a graph-based attention model, which uses medical ontologies to handle data insufficiency and combines with an RNN to learn robust representations; and Diploe\textsuperscript{14} uses bidirectional recurrent neural networks (BRNNs) to further improve the prediction accuracy.

The aforementioned models are RNN-based frameworks which use medical codes as inputs to predict whether the diagnosis or treatment will appear in the future visit, i.e., binary prediction. However, for some diseases, the doctor may care about the transition and severity level of the clinical event, i.e., multiple prediction. For example, if a person is likely to have osteopenia, the doctor may suggest more exercises and supplements, while if osteoporosis occurs, medications will be necessary. To measure the severity of diagnoses, the diagnosis values may be discretized into multiple status: normal range and abnormal range of different severity (i.e. low/high abnormal range), following doctor's advice or medical references. As a disease may be characterized by multiple important observations or diagnoses, we need to monitor these variables simultaneously.

In this paper, our goal is to predict the status of multiple diagnoses (or observations), with each diagnosis having multiple severity levels. We form our problem as multi-task learning, which first learns a shared representation from all the features, and then performs task-specific predictions. We propose an attention-based RNN model to monitor patient’s longitudinal health information. First, we use an RNN to memorize all the information from historical visits, and then attention mechanisms to measure visit importance. Based on the latent representation, we train multiple classifiers and each focuses on the prediction of a specific task. We perform our model on two applications: predicting chronic states for bone health, and monitoring BloodTest values for cardiovascular disease. Our main contributions can be summarized as follows:

- We propose a multi-task framework to monitor the future status of different clinical diagnoses. We process the monitored diagnoses to multiple severity status following medical references, which can help doctors to make more precise decisions on controlling risks.
- We employ three attention mechanisms to evaluate the importance of previous visits to prediction tasks. This gives the explanation of visit importance, which can provide suggestions for doctors to pay more attention on the information from a specific timestamp.
- Our experiments show promising results of using RNNs to handle historical health information from longitudinal records. We empirically show that the proposed attention-based RNNs outperform widely used methods in multi-diagnoses prediction on real world EHR datasets.

This work will result in an effective tool for the physicians to monitor disease progression for early treatment in a more efficient way. This framework can be used real-time on a regularly scheduled basis which highlights patients whose disease state is more likely to worsen.

2 Method

In this section, we first introduce the format of our healthcare datasets and some basic notations. Then we describe the details of the proposed framework, including the preliminary of RNN structure, proposed attention mechanisms, and the multi-task model. Finally, we describe the interpretation for analyzing the importance of different visits.

2.1 Basic Notations

The EHR data contains heterogeneous variables such as diagnosis results, lab data, and physical functions. Among them, diagnosis results are what we care about most, and we want to monitor their progression. Other variables are risk factors that may potentially influence patient’s health status. Assume that there are $N$ patients and $M$ diagnoses to be monitored, and the total number of visit records for the $n$-th patient is $T^{(n)}$. The health record of a patient can be represented by a sequence of visits $V_1, V_2, \ldots, V_{T^{(n)}}$. Each visit $V_t$ is denoted by a vector of feature variables $x_t$. To monitor patient’s health status progression, the diagnosis results are discretized into several classes, indicating
the severity level of the disease, following doctor’s opinions or medical references. For example, in a patient’s visit for bone health test, bone mineral density (BMD) in different areas such as femoral neck and intertrochanteric is measured, and the X-ray scan results can be diagnosed as normal, osteopenia and osteoporosis. We want to predict the severity range of BMD value in each bone area in this patient’s next visit. For simplicity, we describe the proposed method for a single patient and drop the superscript \((n)\) in the following notations when it is unambiguous. Figure 1 illustrates the health records of one patient in our data. The patient has multiple visits, and each visit contains multiple variables. Each monitored variable falls into a severity range. Suppose that we are currently at time \(t\) and want to know the diagnoses at time \((t+1)\), this patient’s historical records from \(V_1\) to \(V_t\) can be utilized for the training of the model.

2.2 Model

The basic component of our framework is gated recurrent unit, which is a state-of-the-art deep learning architecture for modeling long range sequences. To further improve its performance, we apply attention mechanisms to measure the importance of historical sequences. To predict the status of multiple diagnoses, we add a multi-task classification layer on top of the learned representations.

Recurrent Neural Network

Recurrent neural network (RNN) captures the characteristics of the input sequence by recursively updating its internal hidden states. Figure 2(a) shows the unfolded RNN structure for a general classification task. For the first visit \(V_1\) (i.e., \(t = 1\)), RNN learns a hidden state \(h_1\) to represent the input feature vector \(x_1\); as time moves to \(t = 2\), feature vector \(x_2\) together with \(h_1\) are fed into the RNN to update parameters in the network, and the learned hidden state \(h_2\) contains information from both \(x_2\) and \(x_1\). Through updating the network parameters recursively, the hidden state \(h_t\) learns all the previous information from \(x_1\) to \(x_{t-1}\). Then a softmax classifier is applied on \(h_t\) to perform classification. As the parameters of the network are shared by each visit, RNN can handle patients with different visit lengths.

We implement our RNN with Gated Recurrent Units (GRU)\textsuperscript{15}, which has been shown to have comparable performance as Long-Short Term Memory (LSTM), while employing a simpler architecture. The structure of GRU is shown in Figure 2(b). A GRU has two gates, a reset gate \(r\) and an update gate \(z\). Intuitively, the reset gate determines the combination of the new input and the previous memory, which allows the hidden layer to drop irrelevant information that is not useful to the prediction, and the update gate controls how much information from the previous hidden layer
Attention Mechanism

As mentioned in the above section, RNN can remember the past information for future prediction. However, it is limited to only a few latest steps, with more impact from later ones, and may not be able to discover major influences from earlier timestamps. Therefore, we apply attention mechanisms to memorize the effect from long-time dependencies, which have gained success in many tasks. In neural machine translation, the attention mechanism can be intuitively described as follows: given a sentence of length $S$ in the original language, RNN is adopted to generate the word representations $h_1, \ldots, h_{|S|}$. To find the $t$-th word in the target language, we assign each word in the original language an attention score $\alpha_{t,i}$, and then calculate a context vector $c_t = \sum_{i=1}^{S} \alpha_{t,i} h_i$ to perform prediction. Through attention mechanism, RNN can focus on specific words when generating each target word. Similarly, in diagnoses prediction, we use a temporal attention mechanism to predict medical results in the $(t+1)$-th visit, according to visit time $t$. The mathematical formulation of GRU can be described as follows:

$$
\begin{align*}
    z_t &= \sigma(W_z x_t + U_z h_{t-1} + b_z), \\
    r_t &= \sigma(W_r x_t + U_r h_{t-1} + b_r), \\
    \tilde{h}_t &= \tanh(W_x x_t + r_t \odot U h_{t-1} + b_h), \\
    h_t &= z_t \odot h_{t-1} + (1 - z_t) \odot \tilde{h}_t
\end{align*}
$$

where $\odot$ denotes the entry-wise product, $\sigma$ is the activation function, $r_t$ and $z_t$ represent the reset gate and update gate at time $t$ respectively, $h_t$ is the intermediate memory unit, and $h_t$ is the hidden unit. Matrices $W_r, W_z, W, U_r, U_z, U$ and vectors $b_r, b_z, b$ are model parameters to be learned. At time $t$, we take the hidden state $h_t$ to predict the labels of time $(t+1)$. The figure illustrates both location-based attention and general attention mechanisms.
records from \( x_1 \) to \( x_t \). The hidden state \( h_t \) from the \( t \)-th visit can be estimated as a representation for the \((t + 1)\)-th visit. However, it may not contain enough long-term visit information. Therefore, we need to derive a context vector \( c_t \) which captures relevant information to help prediction. We propose three methods to compute attention score \( \alpha_t \) in order to obtain the context vector \( c_t \): location-based attention, general attention and concatenation-based attention.

The attention mechanisms are illustrated in Figure 3. The general procedure goes as follows: we first obtain a set of hidden states through the GRU layer, and then calculate the attention score \( \alpha_i \) for each of them, in order to obtain the context vector \( c_t \): an attentional hidden state \( h_t \) is then calculated by combining \( c_t \) and \( h_t \). Thus \( h_t \) contains both current and historical information. Location-based attention, as in Figure 3(a), calculates the attention score solely from each individual hidden state \( h_i \) (\( 1 \leq i \leq t - 1 \)) using formula: \( \alpha_{ti} = W_{ai}h_i + b_{a} \), where \( W_{a} \) and \( b_{a} \) are parameters to be learned. Since location-based attention mechanism only considers each hidden state individually, it does not capture the relationships between the current hidden state and all the previous hidden states. The other two attention mechanisms, as shown in Figure 3(b), calculate attention weight \( \alpha_{ti} \) by considering the relationship between \( h_t \) and \( h_i \). General attention uses a weight matrix \( W_{a} \) to connect \( h_t \) and \( h_i \) through formula: \( \alpha_{ti} = h_{ti}^T W_{a}h_i \). For concatenation-based attention, we first concatenate the current hidden state \( h_t \) and the previous state \( h_i \), and then calculate a latent vector by multiplying a weight matrix \( W_{a} \). Thus the attention weight vector is generated as: \( \alpha_{ti} = u_{a}^T \text{tanh}(W_{a}[h_t; h_i]) \).

After obtaining \( \alpha_t \), we can obtain the context vector \( c_t \) through formula \( c_t = \sum_{i=1}^{t-1} \alpha_{ti} h_i \), which contains the weighted hidden representations of the past visits from \( x_1 \) to \( x_{t-1} \). To combine the information from context vector \( c_t \) and the current hidden state \( h_t \), we employ a simple concatenation layer to generate an attentional hidden state \( h_t \) using \( \tilde{h}_t = \text{tanh}(W_{c}[c_t; h_t]) \), where \( W_{c} \) is the weight matrix to be learned. \( h_t \) contains all the information from \( x_1 \) to \( x_t \), such that the prediction task can be performed on top of \( h_t \).

**Multi-task Diagnosis Prediction**

Our task is to predict the status of multiple measurement results at the time \((t + 1)\) given the historical records from \( x_1 \) to \( x_t \). Figure 4 shows a high-level overview of the proposed model. Given the information from time 1 to \( t \), the \( i \)-th visit’s health record \( x_i \) is fed into an RNN network, which outputs a hidden state \( h_i \) as the representation of the \( i \)-th visit. Along with the set of hidden states \( \{h_i\}_{i=1}^{t-1} \), we compute their relative importance \( \alpha_i \), and then obtain a context state \( c_t \). From the context state \( c_t \) and the current hidden state \( h_t \), we can obtain an attentional hidden state \( \tilde{h}_t \), which is used to predict diagnoses in the \((t + 1)\)-th visit. For the prediction, we use \( M \) softmax classifiers, which correspond to the \( M \) different diagnoses, to predict the severity level for each diagnosis. The representation \( h_t \) contains the visit information of all the input features, and the task-specific classifier focuses on the prediction of each diagnosis.

![Figure 4: Overview of the proposed model.](image-url)
To perform the multi-task classification, we feed the hidden state $\tilde{h}_t$ into each task through a classification layer. Thus the information of the $k$-th diagnosis in the $(i+1)$-th visit can be produced by $\hat{y}^k_t = \text{Softmax}(W^k_s\tilde{h}_t + b^k_s)$, where $W^k_s$ and $b^k_s$ are parameters to be learned. We use cross-entropy between the ground truth $y_t$ and the predicted $\hat{y}_t$ to calculate the classification loss. The total loss is the sum of cross-entropy among all the diagnosis categories in predicted visits of patients. The loss function $L$ can be described as:

$$
L = -\frac{1}{N} \sum_{n=1}^{N} \frac{1}{T^{(n)}} \sum_{t=1}^{T^{(n)}} \sum_{k=1}^{M} \left\{ (y^k_t)^{\top} \log(\hat{y}^k_t) + (1 - y^k_t)^{\top} \log(1 - \hat{y}^k_t) \right\},
$$

where $N$ is the number of patients, $M$ is the number of monitored diagnoses, and $T^{(n)}$ is the number of visits of the $n$-th patient. In the training procedure, we estimate parameters in the proposed models by minimizing the loss function (2).

**Interpretation**

In healthcare applications, giving interpretation of the learned representations is important. Here we evaluate the contribution of the past visits to the prediction of future status in the process of learning latent representations. Since we adopt attention mechanisms, the importance of each visit can be found by analyzing its attention score. For example, for the $t$-th prediction, if the attention score $\alpha_{ti}$ is large, then the probability of the $(i+1)$-th visit information related to the current prediction is high. In most cases, the last visit is usually important for chronic diseases, as patient’s health status usually does not change much during two visits. However, since disease progression is complex and affected by many factors, the disease can get better or worse. Thus the health information of specific earlier visits may be more important for some patients. Therefore, the attention mechanism can help doctors to pay attention to specific important visits in the past.

3 Experiments

We conduct experiments on two real-word datasets, and evaluate the performance of the proposed attention-based RNN models compared to other prediction methods. Moreover, we use case studies to understand the behavior of the proposed models.

**Datasets**

*Study of Osteoporotic Fractures Dataset.* The study of osteoporotic fracture (SOF)\textsuperscript{20} is the largest and most comprehensive study focused on bone diseases. It includes 20 years longitudinal data about osteoporosis of 9,704 Caucasian women aged 65 years and older. Potential risk factors and confounders belong to several groups such as demographics, family history, and lifestyle. We process people’s bone health diagnoses of different areas using the bone mineral density (BMD) values by comparison with young healthy references\textsuperscript{17}, resulting in three BMD levels: normal, osteopenia and osteoporosis.

*BloodTest Dataset.* This dataset\textsuperscript{21} contains multivariate blood tests of 3,000 patients affected by cardiovascular disease from the University Hospital of Catanzaro, Italy. For each patient, there are several blood tests during their in-hospital stay, such as hemoglobin, triglycerides, glucose, and calcium. As suggested by doctors, we pick 12 blood analytes variables which are important to cardiovascular. Each variable has a normal range provided by doctors. Knowing variable transitions in advance can alarm doctors to take actions before the abnormal occurs, in order to reduce the risk of diseases.

As a common issue of EHR, these datasets are irregular sampled and sparse, so that data preprocessing is needed. For each person, we remove those visits without any monitored variables recorded, and remove patients with less than three visits. We use simple imputation to fill missing variables. For the SOF data, we fill the missing variables with the values in the previous visit. For the BloodTest data, we impute missing sequences (where a single variable is missing entirely) with a clinical normal value. This is based on an assumption that clinicians believed it to be normal so that they did not measure it. Other missing variables are filled with the median value of other patients. After data preprocessing and extraction, we obtain the datasets with statistics shown in Table 1.
Table 1: Statistics of datasets.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>SOF</th>
<th>BloodTests</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of patients</td>
<td>5,318</td>
<td>2,055</td>
</tr>
<tr>
<td>Number of visits</td>
<td>22,313</td>
<td>18,758</td>
</tr>
<tr>
<td>Average number of visits per patient</td>
<td>4.19</td>
<td>9.13</td>
</tr>
<tr>
<td>Number of normal claims</td>
<td>25,145</td>
<td>221,642</td>
</tr>
<tr>
<td>Number of low abnormal claims</td>
<td>55,399</td>
<td>17,407</td>
</tr>
<tr>
<td>Number of high abnormal claims</td>
<td>31,021</td>
<td>79,837</td>
</tr>
<tr>
<td>Total number of features</td>
<td>42</td>
<td>17</td>
</tr>
<tr>
<td>Number of monitored diagnoses</td>
<td>5</td>
<td>17</td>
</tr>
</tbody>
</table>

Experiment Setup

For each patient, we want to predict the diagnosis results of each visit based on his/her previous records. To validate the performance of the proposed models in this diagnosis prediction task, we conduct experiments on two categories of methods: baselines and RNN-based models.

We set up two kinds of baselines. The first baseline is to use the median value of each monitored variable from \( V_t \) to \( V_t \) to predict \( V_{t+1} \) for continuous variables. This is based on a heuristic assumption that the most frequent state is more likely to occur. For each patient, we use his/her most popular health status as the current status, regardless of time variations. The second baseline is a multi-task logistic regression (LR). To predict information at \( V_{t+1} \), we feed the health records at \( V_t \) to a logistic regression model with multiple softmax classifiers. This can be viewed as a simplified model of Figure 4 without using RNNs and attention mechanism to learn latent states. This model only considers the effect from previous one time step, rather than long time history.

For the proposed methods, we have several variants, including a plain RNN or attention-based RNNs. For RNN, the architecture is similar to the proposed model, but without the attention mechanism. RNN\(_l\), RNN\(_g\) and RNN\(_c\) are three attention-based RNN models, whose architecture is shown in Figure 4. RNN\(_l\), RNN\(_g\) and RNN\(_c\) stands for location-based, general and concatenation-based attention respectively. The attention mechanism of RNN\(_l\) can be seen in Figure 3(a), and attention mechanism of RNN\(_g\) and RNN\(_c\) can be seen in Figure 3(b).

The proposed approaches are implemented with Theano 0.7.0\(^{22}\), Adadelta\(^{16}\) with a mini-batch of 50 patients is used to optimize Eq. (2). To evaluate prediction performance, we define the accuracy as the ratio between correctly predicted severity status and the total number of variables to be predicted.

Results of Diagnosis Prediction

Table 2 shows the accuracy of the proposed approaches in comparison with baselines on the two datasets. For each patient in the testing set, we predict the health conditions for the subsequent visits using his/her historical health records. For the SOF dataset, we predict the probability of BMD states of normal, osteopenia and osteoporosis for different measurements such as hip and femoral neck. For the BloodTest dataset, we predict the probability of each blood analyte falling into normal, low abnormal and high abnormal. The results are averaged over 5 random trials of 5-fold cross validation. \( \text{Avg.\# Correct} \) represents the average number of correctly predicted claims of 5 random trials. \( \text{Accuracy} \) represents the ratio between correctly predicted claims and total number of claims to be predicted.

We can observe that RNN based methods outperform other baselines. The method of predicting with median values considers each variable separately, without taking into account the time trend and feature relationships. This intuitive method is very sensitive to noise, and cannot capture the correlation between variables. It performs the worst in the BloodTest dataset, possibly due to the reason that variables in that dataset are not independent but have strong correlations (e.g. MYO-CKM-TRHS-GPT-GL-GOT-LDH). Logistic regression (LR) takes the whole inputs into a classifier with multiple softmax functions, in order to classify each monitored variable. The structure of LR can be viewed as the framework of Figure 4 without latent representations to memorize historical information. The inputs of
LR include all the features, such that features can impact on each target task. However, there is no way for logistic regression to memorize historical information, as it only takes information from the nearest one visit. Attention-based RNNs outperform the above baselines. This owes to the capability of RNN in memorizing long-term dependencies of patient’s longitudinal health records, and the attention-based mechanisms can further enhance this capability. For the two datasets, RNN\(_l\), RNN\(_g\) and RNN\(_c\) can clearly outperform plain RNN. Since the prediction of RNN mostly depends on recent visits, it may not memorize all the past information. Through attention-mechanism, RNN\(_l\), RNN\(_g\) and RNN\(_c\) can fully take all the previous visit information into consideration, assign different attention scores for past visits, and achieve better performance compared to RNN.

Table 2: Prediction performance on two datasets.

<table>
<thead>
<tr>
<th>Method</th>
<th>SOF</th>
<th>BloodTest</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Avg.# Correct</td>
<td>Accuracy</td>
</tr>
<tr>
<td>Median</td>
<td>10,509</td>
<td>0.8209±0.0057</td>
</tr>
<tr>
<td>LR</td>
<td>10,125</td>
<td>0.7909±0.0069</td>
</tr>
<tr>
<td>RNN</td>
<td>10,769</td>
<td>0.8412±0.0042</td>
</tr>
<tr>
<td>RNN(_l)</td>
<td>10,822</td>
<td>0.8454±0.0031</td>
</tr>
<tr>
<td>RNN(_g)</td>
<td>10,805</td>
<td>0.8440±0.0027</td>
</tr>
<tr>
<td>RNN(_c)</td>
<td>10,816</td>
<td>0.8449±0.0023</td>
</tr>
</tbody>
</table>

Visit Interpretation

The attention mechanism can be used to understand the importance of historical visits to the current visit. As an example, here we analyze the concatenation-based attention mechanism on the SOF dataset. Figure 5 shows a case study for predicting the diagnoses in the sixth visit through the previous five visits. The concatenation-based attention weights are calculated for the visits from \(V_2\) to \(V_5\), according to the hidden states \(h_1, h_2, h_3\) and \(h_4\). Thus, we have four attention scores corresponding to the visits from \(V_2\) to \(V_5\). In Figure 5, we select five patients for visualization. The X-axis represents patients, and Y-axis is the attention score calculated for each visit. We can observe that for different patients, the attention scores learned by this attention mechanism are different.

For chronic diseases, the last visit is often the most important since patients’ health conditions change slowly. As in the figure, for the first, fourth and fifth patients, the importance of visit increases with time going on. However, this is not always the case due to the complexity of disease progression and impact from risk factors. Table 3 shows the variation of bone mineral density (BMD) diagnoses and attention scores of different visits of the second patient. In each visit, there are five different BMD diagnoses, and the values in the table indicate the severity of bone density.
Table 3: BMD diagnoses and attention scores of one patient with six visits on SOF dataset. 0 is normal, 1 is osteopenia, and 2 (osteoporosis) does not occur for this patient.

<table>
<thead>
<tr>
<th>Diagnoses \ Visits</th>
<th>V_2</th>
<th>V_3</th>
<th>V_4</th>
<th>V_5</th>
<th>V_6</th>
<th>Attention weights</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total hip</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.290</td>
</tr>
<tr>
<td>Femoral neck</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0.361</td>
</tr>
<tr>
<td>Intertrochanteric</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.187</td>
</tr>
<tr>
<td>Trochanteric</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.162</td>
</tr>
<tr>
<td>Wards</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>–</td>
</tr>
</tbody>
</table>

loss. Although V_4 and V_5 are closer to V_6 in terms of time, V_2 and V_3 have the same condition as V_6. Thus health records of V_2 and V_3 are more important to V_6. We can see that the attention mechanism correctly assigns larger weights to V_2 and V_3. As for the BloodTest dataset, using attention mechanism to memorize all the past information is also important. An abnormal blood analyte can temporarily turn into normality via medicine, but it may fall back after some time. Therefore, interpreting visit importance through the attention mechanism can help to better monitor disease progression.

In diagnosis prediction, making decisions using very recent record is usually not enough, and it is important to lookup long term health information. To understand the relationship between the length of patient medical history and the prediction performance, we select 1,000 patients from the BloodTest dataset with more than seven visits. Table 4 shows the accuracy of RNN in predicting the diagnoses from V_2 to V_7. We can see that with the number of visit increasing, the performance can often improve. We believe that it is due to the fact that RNN is able to learn better estimates of patient information as it memorizes longer health records.

Table 4: Prediction accuracy for V_2 to V_7 on BloodTest dataset.

<table>
<thead>
<tr>
<th>Visit</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>V_2</td>
<td>0.8579</td>
</tr>
<tr>
<td>V_3</td>
<td>0.8624</td>
</tr>
<tr>
<td>V_4</td>
<td>0.8706</td>
</tr>
<tr>
<td>V_5</td>
<td>0.8792</td>
</tr>
<tr>
<td>V_6</td>
<td>0.8780</td>
</tr>
<tr>
<td>V_7</td>
<td>0.8735</td>
</tr>
</tbody>
</table>

4 Conclusions

In this paper, we introduce attention-based RNN architectures to predict patients’ disease progression. In particular, we monitor multiple diagnoses status simultaneously, based on patients’ historical health records. By employing recurrent neural network, our model can remember hidden knowledge learned from previous visits. Three attention mechanisms allow us to interpret the prediction results reasonably. Experimental results on two real world EHR datasets show the effectiveness of the proposed attention-based RNN models for simultaneously predicting multiple diagnoses. Analysis shows that the attention mechanisms can assign meaningful weights to previous visits when predicting the future visit information. The proposed approach can be widely used for the prediction of a variety of different diseases.

Acknowledgement

The authors would like to thank Dr. Giovanni Cuda (MD, PhD) for his help and advice. This work was supported in part by the US National Science Foundation under grants NSF IIS-1218393 and IIS-1514204. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.
References

20. https://www.sof.ucsf.edu
Physician Information Needs in Managing Delirium

Teresa Taft, BS1,2, Scott D Nelsen, PharmD MS3, and Stacey Slager, MS2,
Charlene Weir RN, PhD1,2

1IDEAS 2.0, George E Whalen VA Medical Center, Salt Lake City, Utah;
2Department of Biomedical Informatics, University Of Utah, Salt Lake City, Utah
3Department of Biomedical Informatics, Vanderbilt University Medical Center,
Nashville, TN

Abstract
Delirium has the highest occurrence rate of any complication in hospitalized adults over the age of 65. The study objective was to determine physician information needs for use in the development of electronic clinical decision support for physicians managing the care of patients with delirium. Critical incident interviews were conducted with 8 experienced internal medicine physicians and 1 cardiologist. Thematic analysis revealed the following 6 themes: 1) Clinician’s experience an impoverished information field for mental status, 2) Uncertainty is pervasive, 3) Extensive information foraging effort is required for cohesive story building, 4) Goal Conflict leads to missed diagnosis and early closure, 5) Diffusion of Responsibility for treating delirium is common, and 6) Use of structured delirium resources is minimal. Elicited information needs were identified and physician recommendations for improving access to information needed in managing the care of patients with delirium are reported. Information elicited in this study is useful for designing delirium clinical decision support that supports physician cognition.

Introduction
Delirium is acute confusion with attention deficits, which may fluctuate. It is the most common complication of hospitalized adults over age 65, occurring in 14-56% of hospitalized patients, depending on the associated conditions. Delirium is costly, as it is associated with increased complications, longer hospital stays, increased rates of institutionalization, and increased mortality. The annual economic burden in the United States has been estimated as high as $164 billion. Identification of delirium is an imperative because it is often confused with other conditions having a similar appearance. In addition, delays in treatment have been shown to decrease the chance of a full recovery. Studies have shown that delirium can be prevented and when it is caught early and treatment is begun, duration can be reduced. However, delirium is harder to reverse once it has occurred and because delirium is undiagnosed in up to 2/3 of cases, identifying risk is essential, decision support is needed.

Multidisciplinary efforts to reduce incidence of delirium have effectively reduced severity, duration, and incidence of delirium rates. However, previous decision support for attending physicians has had mixed results. For example, a letter and phone call from a geriatrician and pharmacist informing primary care providers of a patient’s ‘Drug Burden Index’ (a measure of exposure to anticholinergic and sedative medications), and advising them to consider whether anticholinergic and sedative medications could be reduced did not reduce the number of these medications that patient’s were receiving in the intervention group any more than in the control group. Physicians reported 1) being uncomfortable altering specialist initiated prescriptions; 2) previous failure at reducing medication, 3) patient unwillingness to alter medication, 4) unawareness that patient was receiving anticholinergic/sedative medications, and disagreement with recommendation. However, geriatric consultation in the hospital reduced delirium by 1/3 and severe delirium by 1/2. In another informatics implementation, physicians received a computer generated highly specific ‘interruptive’ alert that the patient’s delirium associated medication put them at risk for cognitive impairment and a recommendation for preferred alternative medications. This method of alerting physicians did not change prescribing behavior. Alternative methods of designing delirium CDS should be developed. An important first step in a radical design change is an investigation of the information needs and decision-making processes or cognitive tasks of doctors managing the care of these complex patients. Cognitive task analysis (CTA) or a variant of it such as critical incident interviews has been shown to be effective for elicitation of decision processes data.

Objectives
Characterize the decision process, strategies and information needs of physicians managing diagnostic dilemmas in patients with delirium.
Methods
The design of the study was a Qualitative Descriptive Study using Critical Incident Interview methodology. Eight internal medicine physicians (4 of whom reported that they were experienced geriatricians) and 1 cardiologist (N=9) were recruited from the inpatient medical wards at Salt Lake City VAMC and University of Utah Hospital. They were selected because of their extensive experience and all self identified as having overseen the care of an elderly patient with delirium. Clinicians were recruited either in person or through email. A time was set aside for the interview at the physician’s convenience.

Interview Procedures: Physicians were asked to report a complex case in diagnosing a patient over age 65 with delirium. Critical Incident Interview technique, an effective method for elicitation of cognitive processes, judgments, and decision20 was used to elicit qualitative and descriptive aspects of diagnosis and orders through the following steps: 1) The participant is asked to recall and give a brief report of a recently hospitalized patient with delirium where there was a diagnostic dilemma; 2) The researcher asked the participant to create a timeline of critical decision points in order to promote deeper recall and a common reference; 3) Probing was then used to elicit deeper processing of information seeking and the decision processes, including expectations, goals, surprises and “what if” scenarios. After the critical incident interviews, physicians were asked to make recommendations for improving access to the information that they needed in managing patients with acute mental status changes. The interviews, which took approximately 30 minutes, were audio recorded and then transcribed to a word document with any identifying information removed and the recordings were erased.

Data Analysis: Bradley, et al.21 described a method of data analysis and coding to uncover themes in health services research, “for understanding phenomena within their context, uncovering links among concepts and behaviors, and generating and refining theory.” Themes are recurrent and emergent unifying concepts that characterize the specific experiences of individual participants. Thematic analysis has been widely used in understanding decision processes and facilitates the move from a description of what is happening to an understanding of the process by which it is happening.22-24 This study uses Bradley’s methods applied to a cognitive task analysis of a critical incident interview investigation to identify information relevant in designing clinical decision support in the context of acute onset mental status change.

Qualitative analyses were conducted, including thematic analysis and development of content coding taxonomy. Analysis began with three investigators independently reading and identifying relevant text and applying pre-codes to three interview transcripts using procedures recommended by Patton.24 Then the investigators met to discuss and refine the pre-codes. After several iterations the remaining transcripts were independently coded, with additional pre-codes added and further discussion as the need arose. Finally, codes were aggregated into higher-level constructs. Sub-categories emerged which were grouped into themes.

Results
The results are presented in three parts. The first is the alerting method by which physicians became aware of the patient’s altered mental status, the second is the result of the thematic analysis, the third is a report of information needs and recommendations given by the physicians for improving access to information within the medical record.

Altered Mental Status Alerting & Features: Physicians reported on patients with delirium caused by adverse drug events, infection, dehydration, dementia, blood gas imbalance, and distress about hospitalization. For every case reported in our study an initial alerting event prompted investigation by the physician. Physicians reported that this alerting always came from human sources who knew the patient well enough to judge that there had been an acute change from the patient’s normal baseline mentation (as there is no lab value available). The reported sources of the verbal alert were family members, visitors, clinical staff, nursing home staff, and the physician’s own personal knowledge of the patient. In the absence of a verbal alert or personal knowledge, clinicians often reported initially assuming the patient had dementia (because the patient described was older).

Identifying features of delirium, in the cases reported by our study physicians, were: changes in personality, looking frightened or confused, inability to focus, uncooperativeness in a usually compliant patient, inability to recognize family or friend, post hoc diagnosis based on delayed recovery from altered to normal mentation, inability to cooperate, reduced ability to converse, fidgeting, talking crazy, disorientation, agitation, pulling on lines, decreased responsiveness, lack of interest in addiction, and not making connections.

Thematic analysis revealed six major categories. 1) Clinician’s experience an impoverished information field for mental status, 2) Uncertainty is pervasive, 3) Extensive information foraging effort is required for cohesive story
Theme 1. Impoverished Information Field: Interviews revealed that medical records lack information needed to establish mental status baseline. Physicians did not consider an acute mental status change to be definitive for delirium, however, diagnosis required a rule out, “Is this a case of delirium or is it confusion better explained by psychosis, depression, withdrawal, brain injury or a new baseline mentation (dementia)?” Physicians often lacked information needed in making this determination; people who knew the patient and could provide baseline information were often not present when the attending physician was available. Even when information was available, physicians reported that it was often inaccurate, lacked credibility or was hidden in voluminous notes. They often felt that family/friend reports were not accurate. Physicians trusted their personal knowledge and judgment of the patient’s baseline mental status. Physician’s reported that mental status data is difficult to locate in the volume of clinical notes that often accompany elderly patients and terms describing delirium varied a great deal among providers. The following subthemes of ‘Impoverished information field’ were identified: 1) Baseline data is sparse or missing, 2) Medication lists in the record are not accurate, 3) Available information often lacks credibility or is of low value, 4) Mental Status Data is embedded in the note, 5) Searching is complicated by vague terms. See Table 1 for examples.

Table 1. Theme 1: Impoverished Information Field

<table>
<thead>
<tr>
<th>Subtheme</th>
<th>Quote</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline data is sparse or missing</td>
<td>“There really wasn’t much in the history that helped inform us about her previous state.”</td>
</tr>
<tr>
<td></td>
<td>“The last week she had been treated at a different hospital where we didn’t really have access to their records.”</td>
</tr>
<tr>
<td>Medication lists in the record are not accurate</td>
<td>“Re medication information: “We got them from a piece of paper sent with her from her care facility.”</td>
</tr>
<tr>
<td></td>
<td>Re accuracy of medication list in EMR: “It was not accurate... a lot of times it’s the wife ... that’s the more reliable resource.”</td>
</tr>
<tr>
<td>Available information often lacks credibility or is of low value</td>
<td>“I think most families don’t adequately understand the degree of impairment..... so I guess I don’t always trust the family fully.”</td>
</tr>
<tr>
<td></td>
<td>“Most of the time when ‘doctors’ write their history or their physical exam, the mental status portion is either inaccurate.....”</td>
</tr>
<tr>
<td>Mental Status Data is embedded in the note. Searching is complicated by vague terms</td>
<td>Re using a general search to find mental status information: “some of these things, the terms are too broad or there’s too many words involved.”</td>
</tr>
</tbody>
</table>

Theme 2. Uncertainty Is Pervasive: The lack of credible information in complex settings related to mental status and delirium gives rise to a pervasive sense of uncertainty. Not only did physicians have difficulty determining baseline mental status and the timing of the patients mental status change, they often had significant difficulty finding a cause. Attribution of the mental status change was ambiguous and multifactorial. Regardless of family reports, delirium was often viewed as dementia if clinical staff did not directly observe fluctuations in mental status in these older patients. Clinicians made assumptions to build sensible stories when credible information was unavailable, often assuming dementia. Physicians reported that delirium can only be diagnosed if it responds to treatment (the treatable medical cause is found). Post-hoc confirmation following the patients return to normal mentation was sometimes needed before a clinician was confident attributing the altered mental status to delirium. The following subthemes of ‘Uncertainty is Pervasive’ were identified: 1) Difficulty determining baseline mental status & timing of change, 2) Attribution is ambiguous and multifactorial, 3) Assumptions of dementia in elderly patients with altered mental status. See Table 2 for examples.
Table 2. Theme 2: Uncertainty is Pervasive

<table>
<thead>
<tr>
<th>Sub theme</th>
<th>Quote</th>
</tr>
</thead>
<tbody>
<tr>
<td>Difficulty determining baseline mental status &amp; timing of change</td>
<td>“This was especially difficult ... not knowing a baseline essentially and not seeing a report of that in the medical record.” Re availability of family: “When she initially came in the hospital he was in Vegas and not there at admission.”</td>
</tr>
<tr>
<td>Attribution is ambiguous and multifactorial</td>
<td>“We did reverse her known risk factors that we knew about; but were there others out there undetected? Maybe.” Re attribution: “I have no idea and that is why we don’t have any good dementia treatment measures.”</td>
</tr>
<tr>
<td>Assumptions of dementia in elderly patients with altered mental status</td>
<td>“...you just expect somebody who has got advanced dementia to maybe not be very responsive.” Everyone else just thought she had dementia and this was sort of since we hadn’t met her before... we didn’t really realize this was an acute change.” “.... looked like a typical dementia case that was undiagnosed”</td>
</tr>
</tbody>
</table>

Theme 3. Extensive Effort is Required to Forage for Information: Physicians reported that significant time and effort is required to get the needed mental status information as mental health information comes from a variety of disparate sources: (1) from outside care providers and other people who personally knew the patient, (2) from within the EMR, and (3) through home investigation and monitoring of the patient. In addition, significant effort is required to coordinate the information, to put it in sequence and to match it with other information. Creating a coherent story and mental model could take hours of sorting through notes, contacting care-givers, and talking to nurses and family. Physicians reported that compensation for these efforts was minimal. The following subthemes of ‘Extensive Effort is Required to Forage for Information’ were identified: 1) Requesting Records from external healthcare providers, 2) Tracking down a person who knows the patient 3) Searching for information hidden in the EMR, 4) Home Analysis, 5) Continuous monitoring. See Table 3 for examples.

Table 3. Theme 3: Extensive Effort in Information Foraging

<table>
<thead>
<tr>
<th>Subthemes</th>
<th>Quotes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Requesting Records from external healthcare providers</td>
<td>“We had to obtain faxed copies from the other hospital.”</td>
</tr>
<tr>
<td>Tracking down a person who knows the patient</td>
<td>“We went back and talked to the team and everyone that took care of her... Tracking them down and talking to them could take half an hour to an hour.” “... having time to sort of track down the nurse and ask them those specific questions is tough.” “... we did talk with, uh, her care facility and what they had been experiencing over the past week and then what she was like prior to that.”</td>
</tr>
</tbody>
</table>
Table 3 Cont.

<table>
<thead>
<tr>
<th>Subthemes</th>
<th>Quotes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Searching for information hidden in the EMR</td>
<td>“You have to go through each note individually to find certain things”</td>
</tr>
<tr>
<td></td>
<td>“The extensive chart review would take over an hour. Often times I go back to the first ever VA visit in the system.”</td>
</tr>
<tr>
<td>Home Analysis</td>
<td>“We had to get a home visit—report—...we didn’t know what was going.”</td>
</tr>
<tr>
<td></td>
<td>“if this is an un-resolving delirium v. is this a dementia...It takes months and a lot of work.”</td>
</tr>
<tr>
<td>Continuous monitoring</td>
<td>“you’d just watch them and see if it waxes and wanes and gets better or worse and try to do that over time”</td>
</tr>
</tbody>
</table>

**Theme 4. Goal Conflict Leads to Early Closure:** Analysis revealed that competing goals result in distribution of attention and often the result is to obscure mental health information and reduce the emphasis on delirium. Physicians reported disparate goals such as treating co-existing problems, seeking an attribution for the delirium, keeping the patient safe, discharging patients to appropriate care once they had stabilized, and addressing family concerns. Co-occurring health conditions, although not always the reason for hospitalization, often took precedence over delirium and this sometimes resulted in loss of mentation information during transfer between one hospital department and another. Physicians in our study did not feel it was appropriate to add a delirium diagnosis to a patient’s problem list, explaining that delirium is temporary and fluctuating by definition and that patient’s are discharged only after they have stabilized, therefore they no longer have delirium. Patient’s who did not recover from delirium prior to discharge were given an official diagnosis of altered mental status or dementia. The following subthemes of ‘Goal Conflict Leads to Early Closure’ were identified: 1) Treating co-existing problems, 2) Cost/benefit, 3) Managing Patient behavior, 4) Early closure defaults to dementia. See Table 4 for examples.

Table 4 Theme 4: Goal Conflict Leads to Early Closure

<table>
<thead>
<tr>
<th>Subthemes</th>
<th>Quotes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treating co-existing problem</td>
<td>“She... had fluid overload and so they were just focusing on that, weren’t looking at her mental status”</td>
</tr>
<tr>
<td>Cost/benefit</td>
<td>“He still wasn’t back to baseline. He still - he knew who I was by the time he was discharged, but he was still kind of confused ... he was confused enough that he wasn’t an active participant in his rehab and so orthopedic finally said, we can’t keep him here. We got to send him to rehab facility.”</td>
</tr>
<tr>
<td>Managing Patient behavior</td>
<td>“...so she had pulled out her line and then of course, the medical team... ended up replacing the line and putting her in restraints. She managed to get out of her bed in her restraints but fell hard enough that she broke her hip.”</td>
</tr>
<tr>
<td>Early closure defaults to dementia</td>
<td>“The team discharged her with dementia and figured she was demented, and didn’t really appreciate this was an acute delirium in front of them.”</td>
</tr>
</tbody>
</table>

**Theme 5. Diffusion of Responsibility:** There was a significant hesitance to claim jurisdiction over the problem of delirium or an acute mental status change. Physicians reported that delirium is more effectively treated in a setting
outside the hospital or should be under the purview of psychiatrists. Despite being a physiological problem, physicians were not comfortable making a diagnosis of delirium without a psychiatric consult. The following subthemes of ‘Diffusion of Responsibility’ were identified: 1) No easily discernable treatment that can be solved during my period of responsibility, 2) Mental health is often viewed as the arena of psychiatrists, even though delirium is a medical problem, 3) Significant Hesitance to Claim Jurisdiction over the Problem of Solving Delirium. See Table 4.5 for examples.

### Table 5. Theme 5: Diffusion of Responsibility

<table>
<thead>
<tr>
<th>Subtheme</th>
<th>Quotes</th>
</tr>
</thead>
<tbody>
<tr>
<td>No easily discernable treatment that can be solved during my period of responsibility</td>
<td>I think we got her on a stable regimen and got a good diagnosis; excluded a lot of other possibilities and then just decided she would probably be better off in a stable, calm environment of a skilled nursing facility.</td>
</tr>
<tr>
<td>Mental health is often viewed as the arena of psychiatrists, even though delirium is a medical problem.</td>
<td>“We consulted psychiatry to, you know, address the possibility of a hypoactive delirium.”</td>
</tr>
<tr>
<td></td>
<td>Re distinguishing delirium from dementia: “... get a psychiatry or geriatrics you know, evaluation to come in or psychologist come in and do some further testing or you but I think in reality for us, you know, as being the primary team, I think, usually we don’t do that.”</td>
</tr>
<tr>
<td>Significant Hesitance to Claim Jurisdiction over the Problem of Solving Delirium</td>
<td>“He still wasn’t back to baseline... he was confused enough that he wasn’t an active participant in his rehab. And so orthopedic finally said, we can’t keep him here...”</td>
</tr>
</tbody>
</table>

### Theme 6. Minimal Use of Formal Clinical Decision Support Tools:

Despite knowledge of their existence, physicians in our study did not use anticholinergic load calculators, lists of medications to avoid in the elderly, or specific delirium screening tools, and physicians reported that they usually do not see medical record alerts. The use of general mental status assessment tools was also minimal as physicians reported that they could get a sense of the patient’s mental status by speaking with them, therefore, they did not require a formal evaluation. The following subthemes of ‘Minimal Use of Formal Clinical Decision Support Tools’ were identified: 1) Recommended delirium assessment tools are not used, 2) Medication guidelines & tools are rarely used, 3) Alerts don’t reach the right people or are ignored. See Table 4.6 for examples.

### Table 6. Theme 6: Minimal use of Formal CDS Tools

<table>
<thead>
<tr>
<th>Subtheme</th>
<th>Quotes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recommended delirium assessment tools are not used</td>
<td>“I think it’s ideal to ask them, you know, always be asking them about their orientation and especially asking nighttime nurses about their behaviors, but often I feel like I don’t have time to do that or you think it might be insulting to ask a patient who seems, oriented, some silly questions every single day.”</td>
</tr>
<tr>
<td>Medication guidelines &amp; tools are rarely used</td>
<td>“we didn’t pull out the Beers list and look specifically, just relied kind of on knowing which medications where commonly found on the Beers list.”</td>
</tr>
<tr>
<td></td>
<td>“Anticholinergic load calculators are available, but I didn’t use one.”</td>
</tr>
</tbody>
</table>
Table 6 Cont.

<table>
<thead>
<tr>
<th>Subtheme</th>
<th>Quotes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alerts don’t reach the right people or are ignored</td>
<td>“we wind up overdosing the patient on morphine or Gabapentin, when things are changing... I don’t know that anybody on the medicine side gets pinged that this is dosed incorrectly now”</td>
</tr>
<tr>
<td></td>
<td>“alert fatigue is real... because you know, click right on, right past the creatinine clearance.”</td>
</tr>
</tbody>
</table>

Physician Information Needs & Recommendations: Analysis of the transcripts also revealed a list of specific information needs in managing the care of patients with altered mental status. This list is reported in Table 7. Physicians made the following 7 recommendations for improving access to information for managing patients with acute mental status changes. 1) “Have a mechanism to review things faster...mental status evaluations, decision making capacity in the past,” 2) “Assess functional capacity prior to hospitalization in older adults,” 3) “If something like a mini-mental status exam gets done ... have a separate place where you can record the raw score,” 4) “It would be nice if when you type in the term and search the notes, ‘the system’ not only pulled up the notes with that term, but the term was already highlighted,” 5) “The medical record is helpful, but it’s hard to know how much of that is just copied and pasted from visit to visit,” 6) “do the CAM [Confusion Assessment Method] daily for patient’s at risk of delirium,” and 7) “Automatically showing trends for common lab tests, hematocrit, platelets, creatinine, BUN, sodium graphic presentation could key us into a problem developing delirium.”

Table 7. Reported Information Needs of Physicians Caring for Elderly Patients with Altered Mental Status

<table>
<thead>
<tr>
<th>Baseline mental status &amp; history of cognitive impairment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fluctuation in mental status during hospitalization</td>
</tr>
<tr>
<td>Medications: recent changes, narcotics, drug-drug interaction</td>
</tr>
<tr>
<td>Alcohol use</td>
</tr>
<tr>
<td>Psychiatric history</td>
</tr>
<tr>
<td>Baseline functional status</td>
</tr>
<tr>
<td>Recent medical procedures</td>
</tr>
<tr>
<td>History of falls</td>
</tr>
<tr>
<td>Signs of agitation: crying out, fidgeting, pulling on lines</td>
</tr>
<tr>
<td>Electrolyte imbalance, metabolic imbalance, signs of dehydration</td>
</tr>
<tr>
<td>Signs of infection: fever, CBC, skin exam, chest x-ray, blood cultures, urinalysis, urine cultures, spinal fluid test/culture</td>
</tr>
<tr>
<td>Signs of organ failure: chem panel, electrolytes, oxygen saturation, blood gas, creatinine clearance</td>
</tr>
<tr>
<td>Signs of brain injury: imaging, neurologic test</td>
</tr>
</tbody>
</table>

Discussion
Delirium is a non-specific sign of a generalized failure of the body to maintain physiological homeostasis, the conditions necessary for normal cognitive function. It is a symptom that is extremely worrisome to patient’s families and is associated with patient safety events, increased healthcare utilization, and increased rates of death and higher rates of institutionalization at long-term care facilities. Our study revealed that physician’s may frequently assume dementia when managing the care of a cognitively impaired elderly patient whose baseline is unknown, creating diagnostic errors. This assumption is counter to the recommendations of assuming normal baseline mentation in the absence of information. This error in judgment may be explained by the uncertainty around mental status data and
the reported high “costs” of finding information combined with the low pay-off (delirium interventions do not produce observable differences in outcomes when compared with good care unless a specific treatable cause is identified).

Physicians in our study reported alerting features of their patients with delirium, including observable changes in personality, mood, appearance, cooperativeness, and cognitive skill that may be quickly recognized by family or caregivers familiar with the patient. The importance of obtaining mental status information from these people, when diagnosing delirium, has been previously reported.

Designing an effective computerized cognitive decision support will require creating several layers of organizing and displaying information, including better extraction of historical data across roles in order to view trends over time and a better integration of potential causes in order to support diagnosis and treatment. The distributed nature of mental status information makes creating such a design particularly difficult, because it must go beyond physicians alone. Integrating each of these steps into a coherent cognitive support design requires an innovative approach.

**Mental status documentation:** Providing support to clinicians in documenting mental status more precisely is an important step in providing CDS. Mental status information is often very “muddled” and clinicians often use terms that are imprecise and ambiguous. The result is that it is hard to determine when the mental status change occurred and subsequently difficult to identify a cause. The result may be a certain sense of “helplessness” leading to diffusion of responsibility. Any decision support intervention would benefit from improved precision in the identification and measurement of altered mental status. Because the largest benefit, in delirium, is achieved in prevention, providing physicians with information that enables them to reduce risk may improve outcomes the most. Awareness of the patient’s altered mental status, in the study cases always came from personal knowledge or verbal report of human sources who knew the patient well enough to judge that there had been an acute change from the patients normal baseline mentation. Currently, physicians report that they rarely read nurses’ notes, and the way that mental status information is documented by nurses is quite different than the way it is documented by physicians. Nurses almost universally report and measure Level of Consciousness (Alert and Oriented: A&O). This measure is not commonly displayed in a usable form nor tracked over time; although it reasonably correlates with other measures of delirium. Another step in improving delirium care and providing cognitive support to clinicians would be to improve communication about mental status across clinical roles.

Patient symptoms that prompted delirium investigation were visual or social/behavioral. Only a companion or caregiver is likely to notice these changes and because these types of changes are common in hospitalized people, reporting of the delirium is not likely to occur until it has existed for some time or is severe. At this point, damage to cerebral function may be difficult to reverse. Improving the communication process between patients, their families and providers is complex and requires integration of information across multiple roles (nurses, physicians, social workers, pharmacists, etc.) Methods of monitoring and earlier physician notification of delirium risk are important in this vulnerable population. Campaigns to teach the public early warning signs of heart attack and stroke have been quite successful. Perhaps it is time for a campaign to promote public awareness of delirium.

When delirium has not cleared prior to release from hospital the diagnosis often defaults to a dementia diagnosis, simply because there is no evidence that the delirium will clear. This erroneous dementia diagnosis may lead to inappropriate treatment that fails to support the conditions needed to correct the delirium. The need for a diagnosis of unresolved delirium rather than dementia may be essential for a patient to receive appropriate care.

**Attribution of Cause:** An attribution of cause is often required for humans to initiate treatment. This is especially true in delirium where diagnosis requires “evidence from the history, physical examination or laboratory findings that the disturbance is a direct physiological consequence of another medical condition, substance intoxication or withdrawal, or exposure to a toxin, or is due to multiple etiologies.” Providing support for physicians to collate and integrate the necessary information in a way that would support a causal attribution would be a core and important function of a decision support tool. Physicians do not often enter a diagnosis of delirium on their patient’s problem list. Some reasons that physicians may not diagnose delirium are: 1) Delirium is fluctuating by definition; it comes and goes. Physician’s who only see patient’s at the same time each day may not witness these fluctuations; and nurses commonly attribute evening changes in mental status to “sun-downing,” which they may not consider noteworthy; 2) Delirium is expected during the course of many hospitalizations and is not perceived as important; 3) Testing associated with a diagnosis of delirium may not be reimbursed; 4) Delirium is viewed as a symptom of a medical problem (such as over-medication, electrolyte imbalance, oxygen deficiency, infection, or dehydration) not a disease, per se, and 5) there is a tendency to diagnose acute altered mental status as dementia if it has not cleared prior to hospital discharge. Although these reasons might justify the decision not to add delirium to the problem list,
prior history of delirium imparts a higher risk of subsequent delirium to the patient. Knowledge of this risk may be important for future delirium risk prevention. Linking of the delirium incident and the attributions or associated patient conditions might be especially useful in subsequent medication and procedure decisions. Finally, delirium often extends past the hospitalization into other settings resulting in a diffusion of responsibility regarding identification of cause.

**Electronic Medical Record Design:** Mental status information is commonly documented in free-text notes using a wide variety of words. 25,30 Electronic medical records do not currently possess the capacity to support physician’s efforts in identifying mental status changes and fluctuations in an efficient manner. Design methods that (1) enable confirmation of accuracy of mental status information and (2) display mental status information in relevant groups and patterns may decrease cognitive load and support confidence in medical information for decision-making. There is a need for easy access to validated mental status information of older patients in the electronic medical record, prior to hospitalization. Trending display of fluctuations in mental status would support delirium workup. EMRs should alert clinicians of patient specific delirium risk factors and changes in patient mental status.

A concern in our findings is that physicians reported that alerts don’t reach the right people or are ignored, and medication guidelines and tools are rarely used, despite clinician knowledge of the existence of these tools. Presenting this information to providers without giving them a disruptive alert should be investigated.

Mental status information is sometimes lost as patients transition from setting to setting. Not only is the specific information lost, but the causal attribution or links of that information to other related information in the chart, such as when medications were started, is also lost. Interoperability tools and standards need continued development and promotion to improve access to records from external healthcare institutions. Improvements to search mechanisms are needed to improve efficiency in locating mental status data embedded in the medical record.

**Limitations:** This study was conducted in one location; therefore verification of themes is needed in other settings. Additional themes may be identified in future work. These results may only be applicable to physicians treating geriatric patients because all cases reported were older patients. Eight of nine cases were internal medicine; therefore the results may not be generalizable to other specialties. Physicians selected cases to report, and only reported cases that were memorable, difficult, and where the delirium was eventually identified, therefore, cases where the delirium was not identified is beyond the scope of this study.

**Conclusion**
Clinicians had significant information access and cognitive demands when dealing with a patient with delirium or acute onset mental status changes and substantial uncompensated effort was needed to try to uncover information. Uncertainty due to the lack of verifiable mental status information created problems with sense making for the clinical providers.

**References**
Facilitating Cohort Discovery by Enhancing Ontology Exploration, Query Management and Query Sharing for Large Clinical Data Repositories

Shiqiang Tao, PhD¹, Licong Cui, PhD², Xi Wu, MS², Guo-Qiang Zhang, PhD¹
¹Institute for Biomedical Informatics, University of Kentucky, Lexington, KY
²Computer Science Department, University of Kentucky, Lexington, KY

Abstract

To help researchers better access clinical data, we developed a prototype query engine called DataSphere for exploring large-scale integrated clinical data repositories. DataSphere expedites data importing using a NoSQL data management system and dynamically renders its user interface for concept-based querying tasks. DataSphere provides an interactive query-building interface together with query translation and optimization strategies, which enable users to build and execute queries effectively and efficiently. We successfully loaded a dataset of one million patients for University of Kentucky (UK) Healthcare into DataSphere with more than 300 million clinical data records. We evaluated DataSphere by comparing it with an instance of i2b2 deployed at UK Healthcare, demonstrating that DataSphere provides enhanced user experience for both query building and execution.

Introduction

In the last decade, electronic health records (EHR) technologies have been increasingly adopted across the United States. According to a data brief from the Office of National Coordinator for Health Information Technology, the adoption of basic EHR technology between 2008 and 2015 rose from 9.4 to 83.8 percent¹. This exponential growth directly resulted in a significant increase in the rise of the secondary use of clinical data. Secondary use of clinical data aims to reuse the data captured from clinical care for purposes of research or quality improvement ². A major source of valuable information for clinical and translational research resides in Enterprise Data Warehouses (EDWs) or integrated clinical data repositories (IDRs), as they are designed to integrate multiple facets of data generated from patient care. Surveys³, 4 show that many IDRs are built on the transaction-based clinical care systems in recent years. Previous research ⁵–⁸ also suggests that IDRs have facilitated researchers to conduct biomedical research in various areas such as monitoring infectious diseases, mining patterns between disease progression and management, and identifying eligible subjects for clinical trials.

A typical IDR query engine consists of 4 necessary components( see Figure 1): Ontology Concepts, Observational Facts, Query Interface, and Query Execution. The Ontology Concepts component manages terminology and knowledge information such as ICD-9-CM or ICD-10-CM diagnosis codes, RxNorm medication codes, and LONIC lab test codes. Observational Facts contain patient clinical information including but not limited to demographics, diagnoses, lab tests, procedures, and medications that are structured and coded by the concepts provided by the ontology concepts component. Query Interface is the component that directly interacts with end users providing them the interface to enter query terms and review query results. The Query Interface determines the effectiveness of an IDR directly. Behind the Query Interface is the Query Execution component, which translates queries composed in the Query Interface into query language of the backend database and retrieves query results.

Query exploration tools implemented as parts of query engines make IDRs easier and more practical to use. Existing query engines such as i2b2⁹ and Harvest¹⁰ provide direct support for data query needs. These systems incorporate powerful interfaces to allow researchers to explore clinical data, build queries, and execute queries without the need for the users to understand the underlying data models. Previous research shows that properly designed query interfaces significantly lower decision maker’s mental workload¹¹. However, the query interface is increasingly recognized as a bottleneck for the rate of return for investments and innovations in clinical research ¹²–¹⁴. Therefore, query interfaces need to continue to improve with the availability of advanced information technologies.

Figure 1. A typical IDR query engine structure.
A challenge of current popular IDR query engines is that they use relational database management system (RDBMS) such as MySQL, PostgreSQL, Oracle and Microsoft SQL Server to manage IDR data. The relational database has been the foundation of enterprise applications for decades, but with the explosion in both the volume and variety of data in recent years, researchers have found that the traditional RDBMS has become a performance bottleneck and NoSQL database management systems such as MongoDB outperformed it in management of large complex biomedical data\textsuperscript{15,16}.

Because of the rich variety of data sources integrated in an EDW, existing approaches to data integration leverages an amalgamation of controlled vocabularies (or ontologies) such as ICD, CPT, LOINC, and SNOMED CT. Too close of a coupling between the roles of these controlled vocabularies play, both for data coding/annotation, as well as for the interface for query construction, can limit the benefits of what informatics innovations can offer. The strategy of decoupling of the backend annotation and front-end query interface roles has been previously used successfully in three projects: VISAGE/PhysioMIMI\textsuperscript{17}, x-search for the NHLBI-funded National Sleep Research Resource\textsuperscript{18}, and the epilepsy-ontology-driven query interface for the NINDS-funded Center for SUDEP Research\textsuperscript{19}. Since 2010, we have been applying an ontology-driven methodology to the interface design of clinical data management tools. A series of tools have been created under this philosophy including VISAGE (2010) \textsuperscript{17}, OPIC (2012) \textsuperscript{20}, FEMI (2013)\textsuperscript{21}, MEDCIS (2014)\textsuperscript{22}, and Neuro3D (2016)\textsuperscript{23}. Using the role-decoupling strategy, we developed a prototype system called DataSphere, which incorporates ontologies in a streamlined interface with dynamic browsing, searching capabilities to facilitate the construction, management and sharing of queries for cohort discovery. We performed limited-scale comparative experiments both on usability and performance of DataSphere on UK Healthcare’s clinical data by porting our i2b2 data store to MongoDB to achieve enhanced flexibility and performance.

1 Background

1.1 Biomedical Ontologies

Biomedical ontologies play an important role in IDRs, which provide a standard and formalized way to specify entities (or concepts), their attributes and relationships among the entities. This enables integration, query, and exchange of heterogeneous biomedical data and interoperability between computer systems. The ontologies involved in this paper include International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) and International Classification of Diseases, Tenth Revision, Clinical Modification (ICD-10-CM) for diagnoses coding, Logical Observation Identifiers Names and Codes (LOINC) for lab tests coding, Current Procedural Terminology (CPT) for medical, surgical, and diagnostic procedures coding, and RxNorm for clinical drugs coding. DataSphere imports these ontology concepts into database and visualize them with its “Concept Finder.”

1.2 VISAGE

The VIStual AGgregator and Explorer (VISAGE) is a domain ontology-driven interface for a federated approach to data integration. Interface design features in VISAGE include auto-generated slider bar, selection boxes, and built-in charting. DataSphere’s user interface is an extension of VISAGE, with new features (Section 2.2) developed such as: Concept Finder with browse and search modes to locate ontology concepts, Widget Canvas with single and disjunctive concept containers, logical annotation between different concept containers and widgets, and collapsable organization of different interface modules. All these new features work together to make query building more intuitive and efficient.

1.3 MongoDB

As an open-source, cross-platform document-oriented database, MongoDB\textsuperscript{24} provides us with advantages such as flexible data model, scalability, and improved performance. MongoDB uses a different set of terms for database storage and operations. The terms involved in this paper include “Collection,” “Document,” and “MQL” similar to the concepts “Table,” “Row,” and “SQL” in the traditional relational database. Unlike the traditional relational database that stores data in tables and uses structured query language (SQL) for data access, MongoDB stores data in JSON-like documents that can vary in structure. Therefore, in MongoDB, related information can be stored together for fast access using MongoDB query language (MQL), which is the main motivation for choosing MongoDB as the backend database for DataSphere.
Table 1. Architecture difference between DataSphere and i2b2.

<table>
<thead>
<tr>
<th>Function Module</th>
<th>i2b2</th>
<th>DataSphere</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ontology Library</td>
<td>Pre-defined</td>
<td>Plug</td>
</tr>
<tr>
<td>Query Interface</td>
<td>Pre-defined</td>
<td>Dynamically Rendered</td>
</tr>
<tr>
<td>Patient Data Import</td>
<td>via ETL</td>
<td>via Files of Common Formats</td>
</tr>
<tr>
<td>Backend Database</td>
<td>RDBMS</td>
<td>MongoDB</td>
</tr>
<tr>
<td>Query Management (Save Load)</td>
<td>Limited</td>
<td>Comprehensive</td>
</tr>
<tr>
<td>Query Share</td>
<td>Not supported</td>
<td>Comprehensive</td>
</tr>
</tbody>
</table>

1.4 i2b2 and UK Healthcare IDR

i2b2 is an open source clinical data analytics platform originally developed with funding from the National Institutes of Health. i2b2 is widely used for querying patient data to address research questions. i2b2 aggregates a copy of the patient data from an EHR system and provide query services in parallel to the EHR system for research purposes. i2b2 provides a web client with which users can build customized queries and retrieve results. As an excellent query engine for IDRs, i2b2 has over 100 instances deployed internationally in various organizations including CTSA sites, academic health centers, HMOs, and companies.25

UK HealthCare is the hospitals and clinics of the University of Kentucky. It has captured clinical data for more than one million patients since 2006. The data include, but are not limited to, demographics, financial classification, provider level details, medical diagnoses, medical procedures, lab tests and results, medications, visit details, and vital signs. An i2b2 instance was deployed at UK Healthcare in 2016. It has more than 350 active users and has handled more than one thousand queries from UK Healthcare community. In this paper, we import UK Healthcare IDR data into DataSphere and evaluate it by comparing with UK Healthcare’s i2b2 instance.

2 Methods

2.1 Architecture Design

DataSphere follows the web-interface driven development methodology created in our Neuro3D work23 to involve the end users closely in the process of system design and implementation. As illustrated in Figure 2, DataSphere consists of four critical modules: Data Import, Query Interface, Query Translation and Optimization, and Query Management. First, DataSphere imports data from IDRs to its own MongoDB database; then based on the imported ontology concepts, it renders a query building interface with two critical components: Concept Finder and Widget Canvas. Following that, users can customize data queries using the building interface and the queries will be translated into MongoDB query language and optimized to run against the imported clinical data. At last, the queries can be saved for reuse or shared with other DataSphere users.

Table 1 shows the architecture differences between DataSphere and the traditional IDR i2b2. DataSphere’s allows a flexible ontology library and the ontology library dynamically determines the rendering of DataSphere’s query interface. Instead of relying on specific extract, transform, load (ETL) process, DataSphere import patient data from files of common formats such as csv and json, which provides the general applicability of DataSphere to various IDRs. DataSphere adopts MongoDB as the backed database to replace the traditional RDBMS to avoid potential data processing bottlenecks. For query management tasks (save, load, share), DataSphere provides comprehensive supports to encourage the collaborations between researchers. In the following section, we describe in detail how each module works.

2.2 IDR Data Import

The Data Import module handles the process of importing two kinds of data: data files and ontology concepts. Data files contain patient demographic (age, gender, race, etc) information and clinical (diagnoses, lab tests, procedures, medications, etc) information. Ontology concepts provide a standard way to describe patient data. Multiple ontologies are involved in IDRs. In our work, ICD-9-CM and ICD-10-CM are used for coding diagnoses, LOINC is used for coding lab tests, CPT is used for coding procedures, and RxNorm is used for coding clinical medications. DataSphere imports patient demographic information into a collection called “Patients,” clinical information into another collection called “Observational Facts” and ontology concepts into the third collection called “Concepts” with specific collection data schema. Figure 3 depicts the key fields of each collection and relationships among the collections.
2.3 Query Interface Design

DataSphere’s query interface consists of two critical components: Concept Finder and Widget Canvas.

Concept Finder – A Query of IDRs is built from the concepts that are used to code the patient data. Therefore, an appropriate way of locating or navigating concepts is a necessary component of the query interface. In DataSphere, we implement this component as “Concept Finder.” Concept Finder provides two modes: browsing and search. In browsing mode, concepts are organized in a hierarchical structure that users can navigate down level by level. Search mode performs keyword-based searches against concept names according to the user entry. These two modes are designed for users with different skill levels: more skilled users can search concepts directly by concept names within the searching mode, while less experienced users can navigate concepts level by level using the browsing mode. Designing the browsing mode for large-scale IDRs is a challenging task due to the large size of concepts. For example, UK Healthcare’s IDR contains about 200,000 concepts. Rendering the hierarchical structure of these concepts within a web page takes more than five minutes, which is not ideal for an interactive web application. Concept Finder overcomes this challenge by implementing a tailored concept hierarchy, which initially renders several top levels of concepts and displays the subsequent levels in real time with AJAX calls. In addition, the number of levels displayed initially is configurable. With appropriate configuration, the tailored hierarchy can tremendously reduce the size of concepts for initial display so that the rendering process is instant.

Widget Canvas – DataSphere users can add concepts to the Widget Canvas after they find their target concepts. Widget Canvas can render specific query widgets for concepts automatically according to their properties. For example, as illustrated in Figure 4(a), the concept “Age Today” is a numeric concept so that Widget Canvas renders a slider bar for it. Another example in Figure 4(a) is the concept “Gender” that is categorical with options “Male,” “Female,” and “Unknown.” Then Widget Canvas renders three checkboxes for it. Widget Canvas provides two kinds of concept containers: single concept container and disjunctive concept container. Intuitively, a single concept container only allows one concept in it while disjunctive concept container can consist of multiple concepts. The logical relationship among all concept containers is conjunctive while the relationship among all concepts within one disjunctive container is disjunctive. Semantically, the query displayed in Figure 4 is to find all adult (age ≥ 18) female patients with Parkinson’s disease or Dystonia.
2.4 Query Translation and Optimization

In DataSphere, all query widgets are eventually translated into MongoDB query statements and executed.

Query Translation – As illustrated in Figure 4(b), every concept container is translated to certain MongoDB query statements. The number of translated query statements is equal to the number of query widgets within each concept container since the translation from query widget to query statement is a “one to one” correspondence. The translation is directed by both concept properties and query widget details. In the first query container of Figure 4(a), the “Age Today” concept contains information that its target data collection is “Patients” and its matching column is “Age” And the query widget in the first concept container specifies that the age range to query is from 18 to 216. Therefore, the translated query (S1) is to scan patients collection with the condition as column age between 18 and 216. On the other hand, the last concept container in the figure is a disjunctive concept container with two diagnosis concepts in it: Parkinson’s Disease and Dystonia, with concept codes “ICD10:G20” and “ICD10:G24,” respectively. These two concepts state their target data collection is “Observational Facts” and matching column is “Concept Code” and there is no additional query widget information so that the translation results are two individual statements looking for documents in “Observational Facts” collection with column “Concept Code” as “ICD10:G20” and “ICD10:G24.” In summary, the query composed in Figure 4(a) is translated into query statements: S1, S2, and S3, within which S3 contains two individual MongoDB query statements. In DataSphere, we call S1, S2, and S3 together as a query statement workflow. Therefore, any semantic query in DataSphere is built into a set of query widgets and then translated into a query statement workflow.

Query Optimization – Without any optimization, a query statement workflow is naturally executed by its creation order or translation order. For the example in Figure 4, the query statement workflow should be executed in the order: S1 → S2 → S3. However, the order matters. Queries in DataSphere look for the number of unique patients with characteristics specified in the query widgets. Each statement returns a list of patient ids. The latter executing statement can query within this patient list to reduce its searching space. As stated in the above section, every concept itself defines its target data collection and matching column. We can pre-compute the number of documents within the target data collection match to a concept without additional query widget level restriction. This number is the maximum number of document scanning needed when executing an individual MongoDB query statement translated from the concept. In DataSphere, we call this number “Concept Selectivity.” Concept selectivity intuitively tells us how long it takes to execute a statement because statement execution time is proportional to the number of documents to scan. In addition, a translated query statement’s selectivity is defined as the sum of concept selectivities of all involved concepts. To be clear, the selectivity of a query statement translated from a single concept container is equal to the concept’s selectivity within the container, whereas the selectivity of a query statement translated from a disjunctive concept container is equal to the sum of all concepts within that disjunctive concept container. We optimize the statement execution order.
Table 2. Data import results from UK Healthcare.

<table>
<thead>
<tr>
<th>Data File Name</th>
<th>Format</th>
<th>Size</th>
<th>MongoDB Collection Name</th>
<th>Number of Documents</th>
</tr>
</thead>
<tbody>
<tr>
<td>Concept Dimension</td>
<td>CSV</td>
<td>160MB</td>
<td>Concepts</td>
<td>194,648</td>
</tr>
<tr>
<td>Patient Dimension</td>
<td>CSV</td>
<td>56MB</td>
<td>Patients</td>
<td>978,144</td>
</tr>
<tr>
<td>Visit Dimension</td>
<td>CSV</td>
<td>854MB</td>
<td>Visits</td>
<td>11,313,511</td>
</tr>
<tr>
<td>Facts</td>
<td>CSV</td>
<td>23.07G</td>
<td>Observational Facts</td>
<td>309,053,442</td>
</tr>
</tbody>
</table>

by evaluating the selectivities of statements. We pre-compute all concepts’ selectivities so that the evaluation process is instant. We order all statements by selectivity in ascending order and execute them one by one. Using the above query as an example, it takes 24.66 seconds to complete the query in order: \( S_1 \rightarrow S_2 \rightarrow S_3 \). After optimization, the execution order is: \( S_3 \rightarrow S_2 \rightarrow S_1 \) and the execution time is 0.08 second.

2.5 Query Management

DataSphere supports collaboration by providing researchers with the ability to save and share queries.

**Save a query** – After building and executing a query, DataSphere users can choose to save the query and give it a name and description. Saved queries can be found in each user’s own query library: “My Queries.” Users can search by name or description in their own library and click query name to load the query details in the Widget Canvas. Saving queries in DataSphere is very straightforward and similar to taking “screenshots” for Widget Canvas. As mentioned above, a query is built with several concept containers and every concept container can have multiple concept widgets. When saving queries, every concept container is saved as a “Query Concept Group” and each concept widget is saved as a “Query Concept” within one specific “Query Concept Group.”

**Share a query** – DataSphere provides two ways to share queries: public sharing and private sharing. Public sharing means users can make their saved queries public so that all DataSphere users can view these queries. Private sharing means users can select specific users to share their queries with. Within “My Queries”, each user can see all public queries and shared queries. Users can choose to save public queries and shared queries into their own libraries, in which case the query will not be affected by the owner’s further changes.

2.6 Evaluation Method

We implemented one instance of DataSphere for the IDR of UK HealthCare and evaluated two aspects of DataSphere: usability and performance. UKHealthCare is currently using an i2b2 instance to support patient cohort identification for clinical research. We evaluated the DataSphere instance by comparing it with the i2b2 instance. We selected 8 representative queries from i2b2 query logs, and interpreted them into the human readable language.

**Usability** – We invited 4 data scientists as evaluators to build these 8 queries with both the i2b2 web client and DataSphere. We asked each evaluator to record the query building time in seconds. System usability is considered better if users require less time to build the same query.

**Performance** – To evaluate the performance, we performed the 8 selected queries independently in two systems. We executed each query in each system three times and calculate the average execution time. The system performance is better if the execution time is shorter for the same query.

3 Results

3.1 Data Import

We loaded a total of 24GB de-identified data from UK Healthcare, which contained over one million patients data with more than 300 million observational fact records. We imported the data into DataSphere’s data collections and Table 2 shows the details of the data import results. The data loaded to DataSphere followed the same ETL process of UK Healthcare’s i2b2 ETL pipeline so that the two systems eventually contain the same data set.
3.2 Interactive Interface for Query Building

Figure 5(a) is a screenshot of the query interface of DataSphere consisting of four parts labeled 1 to 4. Label 1 shows Concept Finder’s browsing mode. Six top concepts are displayed in the figure. Hovering mouse over each concept will show the subsequent level of concepts. Each concept is clickable and clicking on the concept will add it to the right area labeled as 2. Label 2 denotes Widget Canvas. Widget Canvas contains a number of concept containers. These concept containers are connected with logical notation “&” indicating they are conjunctive. Concepts within a disjunctive container are connected with logical notation “|” indicating they are disjunctive within that container. Label 3 is the query control panel consisting of three buttons. The first query button triggers the query translation, optimization, and execution process. The second save button populates a form with fields name and description to save together with the query widgets. The last reset button removes all the query widgets in the current Widget Canvas. Saved queries can be found with the link “My Queries” labeled 4.

3.3 Query Management

Figure 5(b) shows “My Queries.” Three different sections are displayed in my queries: “My Own Queries,” “Queries Shared With Me,” and “Public Queries.” “My Own Queries” are queries created by users themselves; “Queries Shared with Me” are queries that are shared privately by other users; and “Public Queries” are queries that are made public by other users. At the end of each row in my own queries, the button “Share” allows users to make their own queries public or visible to specific users. At the end of each row in “queries shared with me” and “public queries” the button “save” allows users to save these queries to their own queries.

4 Evaluation

The evaluation was designed to assess DataSphere’s usability and performance by comparing it with the i2b2 instance deployed at UK Healthcare. To perform the evaluation, we chose some representative queries from the i2b2’s query logs with three goals: (1) these queries should at least cover multiple concepts; (2) all queries should cover all the categories of concepts, and (3) every query should have a clear clinical meaning. With these goals in mind, 8 queries were filtered as shown in Table 3. These 8 queries covered all categories of concepts we imported: Demographics, Diagnoses, Procedures, Lab Tests, Medications, and Visit Details. The smallest number of involved concepts in the queries was 3 and each query had clear clinical meaning. The general purpose of each query was to find the number of patients that satisfied the query’s requirements.

4.1 Deployment Environment

UK Healthcare’s i2b2 environment is deployed on a dedicated server with 72 CPU cores and 378 GB of random-access memory (RAM). The i2b2’s deployed version is 1.7.08 and the backend database is PostgreSQL of version 9. On the other hand, the DataSphere instance is deployed on a high-end research server provided by the Institute for Biomedical Informatics of University of Kentucky with 80 processors (10 CPU cores for each processor) and 1 Terabyte RAM. Although the physical environments of deployment are different, neither of the two systems detect bottlenecks about physical configuration so that we assume these two systems are running on sufficient configurations of RAM and processing power.
4.2 Usability Evaluation

As we have stated in the methods Section 2.5, the usability measurement was the time cost to compose queries. We invited four non-technical users to complete the query building task and counted the time cost. These four users have no development experience with either i2b2 or DataSphere. i2b2 web client also provides browsing and search modes to find concepts. Therefore, we asked the evaluators to build the 8 queries in browsing and search modes separately in the two systems. Table 4 shows the average time cost in seconds to build the 8 selected queries. To avoid potential bias introduced by the order of building these queries in the two different systems, two of our evaluators were requested to build queries first in DataSphere then in i2b2 and the other two evaluators completed the query construction in the opposite order. The results show that DataSphere’s query building efficiency is consistently better than i2b2 in both browsing and search modes for all 8 queries. On average, evaluators built same queries with 36.99% less time using DataSphere’s browse mode than using i2b2’s browse mode; evaluators composed the same queries with 37.14% less time using DataSphere’s search mode than using i2b2’s search mode. Two paired t-tests were conducted to compare query building time in DataSphere and i2b2 within browsing and search mode respectively. Test results show that: in browsing mode, there is significant difference in the scores for DataSphere (Mean=46, SD=10) and i2b2 (Mean=73, SD=13) with a p value < 0.0001; in search mode, the difference between DataSphere (Mean=44, SD=9) and i2b2 (Mean=70, SD=17) scores is also very significant with a p value 0.0016.

4.3 Execution Time

We ran the 8 queries in i2b2 and DataSphere and recorded the running statistics in Table 5. We ran each query in the two systems for 10 times respectively and Table 5 shows the average execution time for each query. For each of the evaluated queries, i2b2 web client took more than five seconds while DataSphere took less than one second. The last row shows that the i2b2 web client took 81.4 seconds to run all 8 queries which were more than 30 times of
Table 5. Performance statistics of query running of i2b2 web client and DataSphere.

<table>
<thead>
<tr>
<th>Query Number</th>
<th>i2b2 Running Time</th>
<th>DataSphere Running Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. 1</td>
<td>5.8 seconds</td>
<td>0.2 seconds</td>
</tr>
<tr>
<td>No. 2</td>
<td>6.1 seconds</td>
<td>0.1 seconds</td>
</tr>
<tr>
<td>No. 3</td>
<td>14.7 seconds</td>
<td>0.1 seconds</td>
</tr>
<tr>
<td>No. 4</td>
<td>23.3 seconds</td>
<td>0.2 seconds</td>
</tr>
<tr>
<td>No. 5</td>
<td>5.2 seconds</td>
<td>0.3 seconds</td>
</tr>
<tr>
<td>No. 6</td>
<td>7.5 seconds</td>
<td>0.7 seconds</td>
</tr>
<tr>
<td>No. 7</td>
<td>11.3 seconds</td>
<td>0.9 seconds</td>
</tr>
<tr>
<td>No. 8</td>
<td>7.5 seconds</td>
<td>0.1 seconds</td>
</tr>
<tr>
<td>Total</td>
<td>81.4 seconds</td>
<td>2.6 seconds</td>
</tr>
</tbody>
</table>

DataSphere’s running time. DataSphere showed much better performance for the selected evaluation queries. A paired t-test was conducted to compare query execution times in i2b2 and DataSphere. There was a significant difference in the scores for i2b2 (Mean=10.2, SD=6.1) and DataSphere (Mean=0.3, SD=0.3) with a p value 0.0029. These results suggest that DataSphere has a better query execution performance than i2b2.

5 Discussions

For a query engine for IDRs, effectiveness (or usability) and efficiency (or performance) are two important aspects to optimize for researchers to achieve better user experience.

Usability – For effectiveness of query engine DataSphere, we described two modes of concept finding in this paper: browse and search. We stated that search is for more skilled users and browse is for less experienced users. We expected search mode to be faster than browse mode in composing queries, but results suggested that also depended on specific queries. For example, in our comparative usability evaluation (Section 4.1), search mode outperformed browse mode for most queries except for query No. 7 when all evaluators reported that browse mode was faster. The reason for that is the evaluators spent a lot of time deciding on the proper search terms for concept “Malignant Neoplasms of Breast.” All combinations except the full name returned too many results, which made it difficult to find the correct concept. Therefore, an appropriate ranking algorithm is a key factor for a successful search function.

Performance – For the efficiency or performance of DataSphere, we observed significant improvement compared to i2b2’s web client. The query workflow optimization contributed substantially to the performance improvement. We verified that by running same queries with query workflow optimization disabled. Results showed that queries ran much slower. For example, query 6 took 19.5 seconds to complete in DataSphere without query workflow optimization, which is much slower than 0.7 seconds with optimization reported in Table 5.

General Applicability The simplicity of DataSphere’s data models makes it easy to integrate heterogenous clinical data from different data sources. Although we did not have a chance to import data from multiple IDRs into DataSphere, we expect it to do so in the future. One type of IDRs in mind are the ones following ETL process for OMOP Common Data Model.

Limitation – First, we were only able to have four evaluators complete the usability testing. Second, the selection process of the evaluation queries is subjective although we stated 3 goals to direct the query selection (Section 4). DataSphere is in a prototyping stage with additional functionalities to complete such as temporal queries and federated search across different data sources.

6 Conclusion

In this paper, we introduced an ontology-driven, enhanced, general query framework DataSphere for large scale integrated data repositories. It uses document-oriented NoSQL MongoDB as the backend database, creates an interactive query interface that renders hierarchical Concept Finder automatically from concepts imported from IDRs, and applies query optimization strategies to improve query execution performance. The comparative evaluations with i2b2 demonstrated superior usability and performance of DataSphere, revealing the potential of DataSphere to support queries for large-scale IDRs.
Acknowledgement

This work is supported by University of Kentucky Center for Clinical and Translational Science (Clinical and Translational Science Award UL1TR001998) and National Science Foundation MRI Award No.1626364.

References


A Scalable Privacy-preserving Data Generation Methodology for Exploratory Analysis

Jaideep Vaidya, Ph.D. 1, Basit Shafiq, Ph.D. 2, Muazzam Asani 2, Nabil Adam, Ph.D. 1, Xiaqian Jiang, Ph.D. 3, Lucila Ohno-Machado, M.D., Ph.D. 3

1 Rutgers University, Newark, NJ, USA; 2 Lahore University of Management Sciences, Lahore, Punjab, Pakistan; 3 University of California at San Diego, La Jolla, CA, USA

Abstract Big data coupled with precision medicine has the potential to significantly improve our understanding and treatment of complex disorders, such as cancer, diabetes, depression, etc. However, the essential problem is that data are stuck in silos, and it is difficult to precisely identify which data would be relevant and useful for any particular type of analysis. While the process to acquire and access biomedical data requires significant effort, in many cases the data may not provide much insight to the problem at hand. Therefore, there is a need to be able to measure the utility/relevance of additional datasets for a particular biomedical research task without direct access to the data. Towards this, in this paper, we develop a privacy-preserving approach to create synthetic data that can provide a first-order approximation of utility. We evaluate the proposed approach with several biomedical datasets in the context of regression and classification tasks and discuss how it can be incorporated into existing data management systems such as REDCap.

Introduction

Today, we have the unprecedented opportunity to gather genomic, transcriptomic, clinical, behavioral, and social data, in ways relevant to health. Analysis of this data can enable new pathways to discovery and can improve the understanding, prevention and treatment of complex disorders such as cancer, diabetes, depression, etc., which are significantly on the rise. Indeed, the Precision Medicine Initiative crucially depends on data-driven science and research.

However, unfettered use of data poses significant concerns regarding patient privacy and the abuse of access to sensitive data. Indeed, the digitization of health information, without appropriate controls, magnifies the risk to privacy, due to the ease of retrieval, analysis, and linkage. Privacy and confidentiality are critical to healthcare. Improving privacy protection encourages people and organizations to share data and realize hidden insights. For example, orphan diseases can be treated more effectively when more observations from different regions in the world are shared and aggregated. Similarly, personalized medicine can be targeted to individuals more accurately if more patients similar to the person of interest are observed and analyzed.

Preserving privacy is a non-trivial task because any protection scheme essentially involves a tradeoff with data utility. Simple strategies may lead to private but uninformative data disclosure, and vice versa. This is especially challenging if the data required for a study resides at several different institutions. Note that in the early stages of any study, the researcher needs to explore the data to understand its utility for the study purpose. Often, the researcher needs to iteratively access the data across the different sources. Even though data exploration may not require fine grained access to all of the data, researchers have to spend an inordinate amount of time trying to get access to relevant data. In many cases, they may not even know whether the data would actually help their particular analysis. For example, consider a researcher at Rutgers Biomedical Health Sciences (RBHS), Newark, carrying out a survival analysis of pancreatic cancer patients using the clinical trials data collected in the local REDCap instance. Since the Newark population is overwhelmingly African-American, who typically have higher incidence rates and suffer from more aggressive forms of cancer, the researcher would like to carry out a comparative analysis of survival after adjuvant radiotherapy and/or chemotherapy with respect to the general population. To do this, the research requires access to linked cancer registry data from the Surveillance, Epidemiology, and End Results (SEER) Program 1 to get information about the cancer site, stage, grade, etc., as well as access to the corresponding Medicare/Medicaid data from the Centers for Medicare & Medicaid Services (CMS) to get information of the specific therapy. If the researcher would like to study the health costs associated with patients having other comorbidities such as diabetes and contrast them with patients who do not, the researcher would also need access to data from the Healthcare Cost and Utilization Project (HCUP) data provided by AHRQ 2, and/or to clinical data such as DMITR1 from the National Center for integrating Data for analysis, Anonymization, and Sharing (iDASH) 3. The researcher would like to confirm which of
these data sources would be relevant to his/her study, and find out which specific data files from these sources should be retrieved. Note that, at this stage, it is sufficient for the researcher to get this confirmation from an approximate analysis without extensive data integration for each different source, and to then retrieve only the relevant data.

Our goal is to precisely enable such exploratory analysis by providing an assessment of the data with respect to utility while preserving its privacy. Towards this, we develop a technique to generate sample datasets that preserve the structure and semantics of the original data, but not exact values, thus preserving its privacy. We evaluate the proposed approach with several different biomedical datasets in the context of classification and regression and demonstrate its effectiveness. While our approach for generating synthetic data does not directly provide a measure quantifying the utility of the dataset for a specific type of analysis (e.g., regression), our evaluation does show that the area under the curve (AUC) / root mean squared error (RMSE) of models built from synthetic data are comparable to that of models built from real data. Thus, a researcher can get a first-order approximation of the overall usefulness of the real data. For example, suppose that a researcher is interested in exploring the relationship between Vitamin D levels and Cancer diagnosis, and is interested in looking at three different datasets which have been collected at institutions in different geographic regions in the country. Using the corresponding synthetic data, the researcher might be able to get an estimate of where such a relationship might exist, or whether the strength of this relationship is more significant in a particular region of the country, or whether the combined data shows the same relationship or not.

An overview of random decision trees

The proposed data generation approach is based on Random Decision Trees (RDT), developed by Fan et al.⁴ The Random Decision Trees algorithm builds multiple (or m) iso-depth (i.e., equal depth) random decision trees. As opposed to typical decision tree learning, the trees built in RDTs are completely random in structure – i.e., the structure of a random tree is constructed completely independent of the training data. However, the statistics recorded for each node are computed based on the training data. Thus, the RDT training phase consists of building the trees (BuildTreeStructure) and populating the nodes with training instance data (UpdateStatistics). It is assumed that the number of attributes is known based on the training dataset. The depth of each tree is decided based on a heuristic – Fan et al.⁴ show that when the depth of the tree is equal to half of the total number of features present in the data, the most diversity is achieved, preserving the advantage of random modeling.

The process for generating a tree is as follows. First, we start with a list of features (attributes) from the dataset. We then generate a tree by randomly choosing one of the features without using any training data. The tree stops growing once the height limit is reached. Then, use the training data to update the statistics of each node. Note that only the leaf nodes need to record the number of examples of different classes that are classified through the nodes in the tree. The training data are scanned exactly once to update the statistics in multiple random trees. Based on the statistical analysis as well as conducted experiments, from 10 to at most 30 trees are sufficient for most applications⁴–⁶. When used for classification, the new instance is evaluated with respect to all of the random trees constructed and the average class probability distribution is output. In our case, we do not need to worry about the classification phase, and after constructing the RDT model, we can use them for synthetic data generation as discussed in the following section.

An important question is why choose RDTs as a way to encode the underlying training data. There are several reasons for this. First, an important property of RDT is that the same code can be used for multiple data mining tasks: classification, regression, ranking and multiple classification⁴–⁶. As shown previously, the random decision tree is an efficient implementation of Bayes Optimal Classifier (BOC)⁴, effective non-parametric density estimation⁶, and can be explained via high order statistics such as moments⁷.

While the use of Random Decision Trees may seem counterintuitive, there are any many benefits in terms of performance and accuracy, that are gained by using this method versus traditional algorithms. The use of the multiple random decision trees in various learning tasks offers many benefits over other traditional classification/tree building techniques, because its structure and progression lends itself to modification for distributed/parallel tasks⁸. At the same time, RDT outperforms other models in terms of computational speed, due to the inherent properties of random partitioning used in tree construction. Indeed, one of the key advantages to RDT is the efficiency gained by the way the model is trained, as well as its minimal memory requirement⁹. Finally, RDTs are also extremely well suited from the privacy perspective since:
1) Randomness in structure rather than simple perturbation of input/output is more effective – perturbing the input or output from a database to achieve privacy works, but the utility of the information garnered from data mining can be diminished if the perturbations are not carefully controlled, or conversely, information can be leaked if the information is not perturbed enough. Instead, we can exploit the design properties of RDT to generate trees that are random in structure, providing us with a similar end effect as perturbation without the associated pitfalls. A random structure provides security against leveraging a priori information to discover the entire classification model or instances.

2) Purely cryptographic approaches are often too slow to be practical and can become computationally expensive as the size of the dataset increases and intercommunications between different parties increase. RDT provides a convenient escape from this paradigm thanks to its structural properties, and furthermore can be computed in a distributed fashion thus preserving privacy even if the data are distributed.

3) An additional security advantage of RDTs are that they are amenable to meeting the privacy requirements of the differential privacy model. As shown by Jagannathan et. al, the node statistics can be viewed as queries over the training data. Therefore, standard techniques can be used to return differentially private results, without significant loss of accuracy.

Proposed Approach

The proposed approach follows a two-step process for generation of synthetic data as shown in Figure 1. In the first step, we generate multiple parameterized RDTs using the original dataset. In the second step, we perform random walk over the RDTs to recreate instances.

The procedure for generation of RDTs is given in steps 1-13 of Algorithm 1. The input to Algorithm 1 is the original dataset, D, and number of RDTs to be generated. The dataset D may include both nominal and numeric attributes. For discretization of numeric attributes, the algorithm takes two discretization parameters as input: i) f is a function that gives the number of splits for each numeric attribute – for example, the numeric attribute Age with range [0, 100] can be split into four discrete values by setting f(Age) = 3; the function g gives the discretization split points for each numeric attribute – for example, g(Age) = (30, 50, 70) discretizes the Age into 4 non-uniform age ranges [0, 30), [31, 50), [51, 70), [71, 100].

After discretization, we build the structure of the required k RDTs by following the standard BuildTreeStructure procedure of RDTs. Next, we use the given dataset D to compute the conditional probability of visiting any node in the RDT. Note that this is the conditional probability of visiting a node given that we are currently visiting that node’s parent, and is given by:

\[
pr(n) = \frac{\text{# instances in } D \text{ reaching node } n}{\text{# instances in } D \text{ reaching node } parent(n)}
\] (1)

For the root of the tree, the probability is 1, since we start from the root. We illustrate the process of creating the RDT through the following example. Consider, as mentioned before, a cancer study. The data set records the cancer site, stage, grade, and other features along with the survival months. Table 1 shows a small dataset sample with 5
Algorithm 1 Synthetic data generation using RDTs

Require: Original Dataset $D$
Require: Number of random decision trees to be generated, $k$
Require: Degree of discretization for each numeric attribute, $f : A \rightarrow \mathbb{Z}^+$
Require: Discretization split points for a numeric attribute, $g : A \rightarrow \mathbb{R}^{f(A)}$
Require: Number of synthetic instances to be generated, $ns$

Ensure: Synthetic Dataset $D'$ such that $|D'| = ns$, and schema($D'$) = schema($D$)

1: $D' \leftarrow \phi$
2: for each numeric attribute $A \in D$ do
3: $splits \leftarrow g(A)$
4: for all instances $i \in D$ do
5: Discretize $i_A$ using $splits$
6: end for
7: end for
8: Generate $k$ random decision trees $RDT_1, \ldots, RDT_k$
9: for each tree $RDT_i$ do
10: for each node $n \in RDT_i$ do
11: $pr(n) \leftarrow (\# \text{ instances in } D \text{ reaching node } n) / (\# \text{ instances in } D \text{ reaching the parent node of } n)$ \{Record the conditional probability of reaching that node\}
12: end for
13: end for
14: for $i = 1 \ldots ns$ do
15: $inst \leftarrow \phi$
16: while $inst$ is not completely generated do
17: Randomly choose a random decision tree $RDT_j$
18: \{Do a random walk over $RDT_j$ to generate the instance values\}
19: $currnode \leftarrow \text{root}(RDT_j)$
20: while $currnode$ is not a leaf node in $RDT_j$ do
21: Choose a child node $c$ to random walk to using the node probabilities computed above
22: if $inst$ does not have a value for current node attribute then
23: Generate the attribute value for $inst$ based on the child node $c$ that is chosen
24: end if
25: $currnode \leftarrow c$
26: end while
27: end while
28: end for
Table 1: Example of Cancer Dataset

<table>
<thead>
<tr>
<th>Cancer Site</th>
<th>Stage</th>
<th>Grade</th>
<th>Age</th>
<th>Diabetic</th>
<th>Total Charges</th>
<th>Survival in Months</th>
</tr>
</thead>
<tbody>
<tr>
<td>Colon</td>
<td>1</td>
<td>2</td>
<td>62</td>
<td>Yes</td>
<td>$70,000</td>
<td>90</td>
</tr>
<tr>
<td>Pancreas</td>
<td>3</td>
<td>3</td>
<td>46</td>
<td>Yes</td>
<td>$45,000</td>
<td>5</td>
</tr>
<tr>
<td>Colon</td>
<td>4</td>
<td>3</td>
<td>53</td>
<td>No</td>
<td>$47,000</td>
<td>39</td>
</tr>
<tr>
<td>Pancreas</td>
<td>2</td>
<td>3</td>
<td>47</td>
<td>No</td>
<td>$52,000</td>
<td>9</td>
</tr>
<tr>
<td>Colon</td>
<td>3</td>
<td>3</td>
<td>72</td>
<td>No</td>
<td>$58,000</td>
<td>46</td>
</tr>
</tbody>
</table>

Figure 2: Example RDT

instances. Figure 2 shows one possible RDT that can be learned from the data in Table 1. Since this tree serves as a partial summarization of the data, multiple such RDTs are created. In this RDT, as mentioned before, the conditional probability of reaching a node is computed as stated in Equation 1. Thus, for example, the probability of reaching the node Diabetic? is \( \frac{2}{5} \) since only 2 out of the 5 instances have stage 1 or 2. Similarly, the probability of reaching the Cancer Site node in the left subtree is \( \frac{1}{2} \) since only 1 of the 2 instances with Stage 1 or 2 cancer is diabetic.

Synthetic data is generated instance by instance by performing random walk over RDTs. For each instance, we randomly choose an RDT and traverse through it using the node probabilities. The attribute values of the instance are assigned based on the path taken during the random walk. After reaching the leaf node, the next RDT is randomly chosen for assignment of remaining attribute values. We continue this process until all the attribute values of the instance are assigned or all the \( k \) trees have been traversed. Traversal of multiple RDTs for any instance may result in different branches for the same attribute. We keep the attribute value that was assigned first and ignore subsequent values. In case, an instance has missing attribute values after traversal of all \( k \) RDTs, we discard all the assigned values for that instance and restart the random walk for that instance. However, this happens very rarely in practice. The specific procedure for creating synthetic data is listed in steps 14-28 of Algorithm 1.

As depicted in Figure 1, the proposed synthetic data generation approach can generate different data views corresponding to the privacy levels of different classes of users. For example, physicians may require accessing data at the finest granularity without any suppression. Therefore, the data view generated for physicians is generated without any generalization of nominal attribute values and splitting of numeric attributes at the finest granularity level. However, for student researchers access may need to be provided at a coarser granularity. Therefore, the data view generated for student researchers is generated with generalization/splitting of attributes at coarser granularity level.

Integration with REDCap

We are currently working on integration of the proposed synthetic data generation approach into a data collection and sharing infrastructure REDCap\(^{12}\), which is widely used in the medical community for building and managing online surveys and databases. The overall architecture of the extended REDCap system for privacy-preserving analysis and sharing of data is depicted in Figure 3. The system will enable different types of users to access external data from different sources through the REDCap application. The access to data will be provided based on user authorizations defined in the access control policy of the data owner. The system allows data owners to specify authorizations for different classes of users in their access control policies. Accordingly, different views
of synthetic data are generated as shown in Figure 3. The system also includes a distributed data access module that enables retrieval and integration of data from multiple sources in response to a user query or exploratory data analysis task. For exploratory data analysis, the system allows a user to integrate data from multiple data sources and assess its utility.

Note that within our architecture, the data owner is still responsible for generating the synthetic dataset. This is not different from the current practice in multi-level secure databases where different views of the data are generated for different users. The idea is that the data owner creates multiple versions of the synthetic data based on his/her access control policy and the privacy requirements of the data to be shared. The synthetic data can be generated offline and the appropriate version be made available to a researcher based on the authorization of the researcher requesting access. This access will be mediated by the data access module. Furthermore, the synthetic data looks just like the real data (i.e., the records have precisely the same structure). Therefore, it can be processed the same way as the real data, and the researcher can run any standard query in REDCap pertaining to that dataset, though the exact answer to the query may vary from the answer computed over the original data.

Since the synthetic data is only an approximation of the real data, the models built from the synthetic data and the real data are likely to be different to some extent. It would be useful for a researcher to know the degree of discrepancy in results based on use of either real or synthetic data. This can be easily incorporated into the REDCap architecture, wherein the utility assessment module can measure the difference between the results of the posed query on both the real and synthetic data and provide a measure of the difference.

**Experimental Evaluation**

We now evaluate the efficacy of the RDT framework in creating synthetic data that can give an approximation of the utility of the original data. For this, we use three different real datasets obtained from the UCI Machine Learning Repository: 1) the Breast Cancer Wisconsin (Original) Data Set\(^{13}\) 2) the Parkinsons Telemonitoring Data Set\(^{14}\) 3) the Diabetes 130-US hospitals for years 1999-2008 Data Set\(^{15}\). We processed the data to remove the instances with missing values. We also removed attributes where the overwhelming majority of instances have the same single value. We also discretized all of the numeric attributes into 10 discrete ranges using equal width discretization. Table 2 gives the characteristics of each individual dataset after preprocessing. The Breast Cancer dataset and the Diabetes dataset were used for classification (since they have a single categorical variable to be predicted). On the other hand, the Parkinsons Telemonitoring dataset had two independent numeric response variables. Therefore, linear regression was independently performed with both response variables. Note that the original RDT paper recommends that the depth
Table 2: Data Set Characteristics

<table>
<thead>
<tr>
<th>Name</th>
<th># instances</th>
<th># Attributes</th>
<th>RDT Depth</th>
<th># Response Variables</th>
<th>Task</th>
</tr>
</thead>
<tbody>
<tr>
<td>Breast Cancer</td>
<td>683</td>
<td>10</td>
<td>5 [10/2]</td>
<td>1</td>
<td>Classification</td>
</tr>
<tr>
<td>Parkinson’s Telemonitoring</td>
<td>3178</td>
<td>20</td>
<td>5 [20/4]</td>
<td>2</td>
<td>Regression</td>
</tr>
<tr>
<td>Diabetes</td>
<td>98042</td>
<td>37</td>
<td>6 [37/7]</td>
<td>1</td>
<td>Classification</td>
</tr>
</tbody>
</table>

of each RDT be set to \(n/2\) where \(n\) is the number of attributes, and that at least 10 RDTs be generated. Following this, for the classification tasks, we created 20 RDTs, where as for the regression task, 15 RDTs were generated. However, unlike the datasets considered in the original RDT paper, the biomedical datasets considered here have several nominal attributes with many values. For example, each of the three diagnosis attributes in the Diabetes dataset has over 700 distinct values, which leads to a significant fan-out. Note that the depth and the fan-out of the tree together determine the size of the tree and thus the memory requirements of the RDT. For Diabetes, since the fan-out of the three diagnosis attributes is extremely high, this tremendously increases the memory requirements, and therefore, we had to limit the depth of the RDT. As such, based on the characteristics of the different datasets, different depths were chosen as noted in Table 2.

The experiments were run on a 24-core Intel(R) Xeon(R) server with the CPU E5-2670 v3 2.30GHz and 264 GB RAM. However, only one of the cores (i.e., no parallelization) was used and the memory heap size was restricted to only 48GB. We now discuss the results obtained. For each dataset, 10-fold cross-validation was carried out. Essentially, the data was split into 10 folds with 9 folds used for training and the remaining fold used for testing. When generating the synthetic data, we built the RDTs from the folds of data kept for training, and then generated synthetic data from the RDTs. This synthetic data was used for training the model, and the model built from the synthetic data was tested over the fold of real data held out earlier for testing. We then compare the accuracy of the model built from the real training data to the accuracy of the model built from the synthetic data. Since we are performing 10-fold cross validation, the accuracy compared is the average accuracy over all of the iterations. For classification, since the classes are not unbalanced, we directly compare the overall accuracy of the classifier using the AUC (area under the curve) metric. For regression, we compare the root mean squared error (RMSE) obtained. One point to note is that once the RDTs are generated, any amount of synthetic data can be regenerated. Therefore, we tested with generating the same number of instances as in the original training set, ten times more instances, and 100,000 instances (note that for the diabetes dataset, the training set size was itself close to 100,000, therefore we did not oversample this dataset). Table 3 gives the results obtained for classification with Naïve Bayes classifier while Table 4 gives the results obtained for linear regression. Note that we did try other classifiers such as decision trees as well and the results obtained were similar. The results show that for classification, the model generated from synthetic data achieves almost the same accuracy as the model generated from the original data in terms of the AUC. Indeed, the AUC of Breast Cancer with synthetic data is 0.9937 even with no oversampling while the AUC with the original data is only marginally higher at 0.9945. Similarly, the reduction in AUC for diabetes is also nominal. For Parkinsons, the RMSE with original data for motor_UPDRS is 6.51. Instead we get an RMSE of 7.05 with the synthetic data with no oversampling. Similarly, the RMSE for total_UPDRS is 8.14 for the original data, whereas it is 8.93 for the synthetic data with no oversampling. Since, lower is better for RMSE, it can be seen that the performance of linear regression is slightly worse with the synthetic data but is still very comparable. Furthermore, in all cases, increasing the degree of oversampling tends to improve the results. Though the accuracy for Diabetes is low with synthetic data, it is also quite low for the original data, since this dataset is quite difficult in terms of classification. The critical point is that from the perspective of exploratory analysis, the models generated from synthetic data do give a similar view of the data as compared to the models generated from the original data. For example, there was significant overlap in the variables identified as significant in the regression model built from the training data and that built from the synthetic data, though the p-values varied.

One point worth noting is that the synthetic data generation process is extremely efficient. Once the RDTs are built, it takes very little time to generate new synthetic instances. For example, generating 100,000 instances only took a few minutes. The process for building the RDTs is memory intensive. However, if the memory required is capped by limiting the depth of the tree, then this process is also extremely efficient. i.e., in our case, it only took a few seconds...
Table 3: Experimental Results for Classification

<table>
<thead>
<tr>
<th>Data Set</th>
<th># Classes</th>
<th>AUC with Original Data</th>
<th>AUC with Synthetic Data (no oversampling)</th>
<th>AUC with Synthetic Data (10-fold oversampling)</th>
<th>AUC with 100,000 synthetic instances</th>
</tr>
</thead>
<tbody>
<tr>
<td>Breast Cancer</td>
<td>2</td>
<td>0.9945</td>
<td>0.9937</td>
<td>0.9933</td>
<td>0.9938</td>
</tr>
<tr>
<td>Diabetes</td>
<td>3</td>
<td>0.6534</td>
<td>0.6134</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 4: Experimental Results for Regression

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Response Variable</th>
<th>RMSE with Original Data</th>
<th>RMSE with Synthetic Data (no oversampling)</th>
<th>RMSE with Synthetic Data (10-fold oversampling)</th>
<th>RMSE with 100,000 synthetic instances</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parkinsons</td>
<td>motor_UPDRS</td>
<td>6.51</td>
<td>7.05</td>
<td>7.04</td>
<td>7.03</td>
</tr>
<tr>
<td>Parkinsons</td>
<td>total_UPDRS</td>
<td>8.14</td>
<td>8.93</td>
<td>8.86</td>
<td>8.87</td>
</tr>
</tbody>
</table>

to generate a single RDT. In general the time taken to build an RDT increases linearly with the size of the training data, which is extremely efficient.

Related Work

There is a lot of prior work on privacy-preserving distributed analytics\(^{16}\). For example, specific analytics tasks such as classification\(^{17,18}\), clustering\(^{19,20}\), association analysis\(^{21,22}\), and outlier detection\(^{23}\) can be carried out in a distributed manner. Privacy-preserving analytics has also been applied in the biomedical domain, for example for Grid LOGIstic REgression (GLORE)\(^{24,25}\), Distributed Cox proportional hazards model\(^{26}\), and Distributed Privacy Preserving Support Vector Machine (DPP-SVM)\(^{27}\) to allow accurate model construction while respecting individual institutions’ data privacy policy, and medical data integration\(^{28}\). All of the above techniques are designed for specific data analytics tasks. The proposed synthetic data generation approach enables exploratory analysis of the data, and can be used for any of the above data analytics tasks. There is some existing work on anonymizing data or for generating synthetic data. For example, techniques for k-anonymity\(^{29–31}\) will give a k-anonymized version of the dataset. However, this can lose significant data utility, and selecting an appropriate value for k is difficult. Aggarwal and Yu\(^{32}\) present an alternative condensation approach that clusters the data and regenerates synthetic data for each cluster – however, this does not keep track of combinations of attributes and may not fully specify all of the various combinations in the existing data.

Conclusion

Exploratory analysis is crucial to scientific innovation and often requires iterative exploration of the data. When the data are distributed across different organizational boundaries, this may lead to significant privacy concerns. Our aim in this paper is to enable exploratory analysis for the biomedical domain that is both privacy-preserving and scalable. Towards this, we develop a synthetic data generation technique using the framework of random decision trees to preserve the structure and semantics of the original data, while protecting privacy. We evaluate the proposed approach with several real biomedical datasets and typical tasks such as classification and linear regression. Our evaluation shows that the proposed approach is effective and can provide a sense of the utility of the data while protecting privacy.

Note that since our goal was simply to provide a first-order approximation of utility, we did not use standard HIPAA de-identification methods at this point. Since RDTs are amenable to meeting the privacy requirements of the differential privacy model, in the future we plan to compare and contrast the synthetic data generation approach with Safe Harbor/statistically de-identified data in more detail. We also plan to explore how other analytics tasks can be carried out and provide more accurate estimates of utility for specific tasks, and measure the uniqueness of a specific dataset for a particular task.
Acknowledgments

Research reported in this publication was supported by the National Institutes of Health under awards U54HL108460, R01GM118574, R01HG008802, R01GM114612, R21LM012060, R01GM118609, U01EB023685, by the National Science Foundation under awards CNS-1422501 and CNS-1564034 and by the Patient-Centered Outcomes Research Institute (PCORI) under Contract CDRN-1306-04819. The work of Shafiq is supported by Pakistan’s Higher Education Commission’s NRPU grant. The content is solely the responsibility of the authors and does not necessarily represent the official views of the agencies funding the research. We would also like to acknowledge the work done by Daniyal Jangda and Zain Sattar of LUMS in preparing and processing the datasets.

References

2. NIS HNIS. Healthcare cost and utilization project (HCUP). 2011;.


ProvCaRe Semantic Provenance Knowledgebase: Evaluating Scientific Reproducibility of Research Studies

Joshua Valdez, MS¹, Matthew Kim, MD², Michael Rueschman, MPH², Vimig Socrates BS¹, Susan Redline, MD, MPH², Satya S. Sahoo, PhD¹

¹Division of Medical Informatics, School of Medicine, Case Western Reserve University, Cleveland, OH
²Department of Medicine, Brigham and Women’s Hospital and Beth Israel Deaconess Medical Center, Harvard Medical School, Harvard University Boston, MA

Abstract
Scientific reproducibility is critical for biomedical research as it enables us to advance science by building on previous results, helps ensure the success of increasingly expensive drug trials, and allows funding agencies to make informed decisions. However, there is a growing “crisis” of reproducibility as evidenced by a recent Nature journal survey of more than 1500 researchers that found that 70% of researchers were not able to replicate results from other research groups and more than 50% of researchers were not able reproduce their own research results. In 2016, the National Institutes of Health (NIH) announced the “Rigor and Reproducibility” guidelines to support reproducibility in biomedical research. A key component of the NIH Rigor and Reproducibility guidelines is the recording and analysis of “provenance” information, which describes the origin or history of data and plays a central role in ensuring scientific reproducibility. As part of the NIH Big Data to Knowledge (BD2K)-funded data provenance project, we have developed a new informatics framework called Provenance for Clinical and Healthcare Research (ProvCaRe) to extract, model, and analyze provenance information from published literature describing research studies. Using sleep medicine research studies that have made their data available through the National Sleep Research Resource (NSRR), we have developed an automated pipeline to identify and extract provenance metadata from published literature that is made available for analysis in the ProvCaRe knowledgebase. NSRR is the largest repository of sleep data from over 40,000 studies involving 36,000 participants and we used 75 published articles describing 6 research studies to populate the ProvCaRe knowledgebase. We evaluated the ProvCaRe knowledgebase with 28,474 “provenance triples” using hypothesis-driven queries to identify and rank research studies based on the provenance information extracted from published articles.

Introduction
Reproducibility of biomedical study results is essential for advancing science as it enables researchers to build on sound results and allows patients as well as their families to have confidence in the resulting drugs, diagnosis tools and related aspects of patient care (1-3). However, there is growing concern in the research community regarding the decreasing reproducibility of published results. For example, a recent Nature journal survey of 1500 scientists found that 70% of researchers were not able to replicate results from another research group and more significantly about 50% of researchers were not able to replicate their own research results (1). This lack of reproducibility has a major impact on multiple aspects of biomedical research and patient care, including cost of clinical trials that are not based on sound study design, and diversion of funding resources from suitable projects to research studies that cannot be replicated (2, 4). In addition, the inability of researchers to replicate published studied affects the confidence of peer researchers and the wider public in results, such as drugs and diagnostics, that are generated from experiments (5). To address this growing “crisis” of reproducibility in biomedical research, there is increasing focus on developing guidelines together with suitable data management tools to support scientific reproducibility. For example, the National Institutes of Health (NIH) released the “Rigor and Reproducibility” guidelines in 2016 to improve existing reporting standards in biomedical research, which focused on rigorous statistical analysis plan, transparency in reporting, and facilitating material sharing (6).

An important component of the NIH “Rigor and Reproducibility” guidelines is the systematic collection, processing, and analysis of provenance metadata, which describes the source or origin of data and underpins scientific reproducibility (10). Provenance information is widely used in computer science research to trace the origin of data and validate data quality, for example database provenance research has focused on provenance of database entries (7, 8). Similarly, workflow provenance records the various steps involved in processing and analyzing data in scientific workflow system to support replicability of results (9). In 2013, the World Wide Web Consortium (W3C), which is the Web technology standards organization, recommended PROV specifications for standardized modeling and accessibility of provenance information (10). The PROV specifications include the PROV-Ontology (PROV-O),
which uses Web Ontology Language (OWL2) to develop a formal model of the provenance information using a common minimum set of terms (11). The W3C PROV-O and related components of the PROV specifications are being widely used in a variety of domains to model and analyze provenance information. However, there has been limited work in the biomedical research domain to adapt and extend the PROV specifications to manage provenance information and support scientific reproducibility. As part of the NIH Big Data to Knowledge (BD2K)-funded data provenance project, we are developing the Provenance for Clinical and Healthcare Research (ProvCaRe) framework to: (a) define a provenance model for biomedical research; (b) evaluate the quality of provenance metadata in published biomedical literature; and (c) enable users to develop rigorous study designs using provenance information of existing studies.

The provenance metadata associated with clinical research studies include study design, statistical analysis model, raw material used in an experiment, inclusion and exclusion criteria, and intervention techniques used in a research study (2, 3). Other metadata are relevant to basic science, such as testing reagents, assays, cell lines, and model organisms. The growing focus on data sharing through NIH-funded initiatives, such as The Cancer Genome Atlas (TCGA) (12), the Human Connectome Project (13), and the National Sleep Research Resource (NSRR) (14), provides opportunities to evaluate the quality of provenance metadata available to support replicability. In particular, there is a critical need to address two essential aspects of scientific reproducibility in biomedical research:

1. **What provenance metadata terms are required to enhance replicability of research results if original study data is already available (e.g., through TCGA or NSRR)?**
2. **How to leverage the provenance metadata available from previous research studies to adapt “best practices” in study designs?**

To address these two important issues, the ProvCaRe project has created a knowledgebase of semantic provenance metadata, which was automatically extracted from published literature describing research studies that have made their data available through the NSRR project. NSRR is a NIH-funded project that is building the largest repository of sleep research data from more than 40,000 sleep studies involving more than 36,000 participants (14). The NSRR project enables a researcher to access large amounts of study data after receiving an Institutional Review Board (IRB) approval and completing a Data User Agreement (DUA) with NSRR.

The development of the ProvCaRe knowledgebase involved development of: (1) a ProvCaRe ontology for modeling provenance metadata terms; (2) a Natural Language Processing (NLP) pipeline that extends the clinical Text Analysis Knowledge Extraction System (cTAKES) (15) with a new “provenance-triple” generator module and dynamic ontology-lookup tool, and (3) an intuitive user interface using the Django Web framework based on the Model View Template (MVT) software architecture pattern. In the following section, we provide additional details regarding the resources and technologies used in the development of the ProvCaRe knowledgebase as well as work related to use of provenance metadata for reproducibility in scientific research.

**Background**

**NIH Guidelines for Rigor and Reproducibility.** The NIH “Rigor and Reproducibility” guidelines released in January 2016 describe the key components of experimental design, transparent reporting, and data management to support scientific reproducibility (3). The guidelines encourage clear description of the statistical analysis methods used in a research study, adoption of best practices for reporting of study results including: (1) use of common reporting standards recommended by the community, (2) repeated performance of experiments, (3) use of randomization as well as blinding techniques, and (4) description of inclusion and exclusion criteria used to select participants, among other aspects related to the conduct of research studies (2). The “Rigor and Reproducibility” guidelines also encourage the establishment of best practices for management of data as well as biological resources used in experiments to ensure data quality, for example antibodies, cell lines, and animals (6). The guidelines also recommend sharing of datasets to evaluate reproducibility of reported results. In this paper, we build on “Rigor and Reproducibility” guidelines to develop the ProvCaRe knowledgebase using a new ontology to formally model provenance metadata designed to enhance efforts at replicating biomedical research studies.

**National Sleep Research Resource (NSRR).** The National Sleep Research Resource (NSRR) is creating one of the largest repositories of de-identified physiological signals and clinical data elements that were collected from multiple research cohorts and clinical trials (14). NSRR aims to provide free access to over 40,000 sleep studies collected from about 36,000 participants, which include demographic, physiological, clinical, and raw physiological signal data as well as annotation files from polysomnograms with scored events (14). Researchers can use the curated and well-defined sleep data from multiple NIH studies to develop large cohort studies with more diverse population that was
not possible with a single study cohort. This is expected to significantly accelerate the pace of research in role of sleep and sleep disorders in pathogenesis of chronic illness such as cardiovascular and metabolic diseases.

**W3C PROV Ontology.** The PROV Ontology (PROV-O) models common provenance metadata terms and properties using the description logic-based Web Ontology Language (OWL2) (16) defined in the W3C PROV specifications. The three foundational provenance terms modeled in PROV-O are Entity, Activity, and Agent, which are linked to each other using seven properties, namely wasDerivedFrom, wasAttributedTo, wasGeneratedBy, wasAssociatedWith, actedOnBehalfOf, wasInformedBy, and used (11). The PROV-O is designed as an upper-level reference ontology that can be used as a common foundation to build interoperable domain-specific provenance applications such as the ontology for ProvCaRe knowledgebase. The use of PROV specifications in the ProvCaRe knowledgebase allows us to leverage a large number of Semantic Web technology tools.

**cTAKES** The ProvCaRe knowledgebase is built using a NLP pipeline that extends the cTAKES tool, which combines rule-based and machine learning approaches for knowledge extraction from clinical free text. The cTAKES builds on the Apache Unstructured Information Management Architecture (UIMA) framework (17) and OpenNLP toolkit. cTAKES consists of six components, namely (1) Sentence boundary detector, (2) Tokenizer, (3) Normalizer, (4) Part-of-speech (POS) Tagger, (5) Shallow parser, and (6) Named Entity Recognition (NER) annotator, which are invoked sequentially to create an annotated dataset from unstructured text. The downloadable version of cTAKES was trained on GENIA, Penn TreeBank (PTB), and a corpus derived from the Mayo Clinic Electronic Medical Record (EMR) repository consisting of 273 clinical notes (15). cTAKES is built using an extensible architecture and this feature has been used in the ProvCaRe framework to customize the NLP pipeline for identifying and extracting provenance metadata from published literature.

**Related Work.** Reproducibility has been a cornerstone of scientific advancement, however there has been growing focus on the lack of replicability of scientific studies as summarized by a review of published studies (4, 18) and survey of researchers (1). The survey of 1500 researchers by Baker et al. suggested that poor reporting and statistical analysis techniques are responsible for irreproducible results (1). To address some of the underlying factors for limited reproducibility in biomedical research, Landis et al. recommended a core set of standards for reporting scientific experiments to support evaluation of results (2). These reporting standards focused on rigorous study design and included: (1) randomization techniques used to assign tasks in a research group, (2) blinding techniques to allow concealment of allocation, (3) specification of statistical methods for estimating sample size; (4) predefined rules for inclusion and exclusion of data; (5) consistency in addressing the issue of missing data, and (6) repetition under different conditions (2). These recommendations form a core component of the NIH Rigor and Reproducibility guidelines. As we discussed in the previous section, the practical implementation of the NIH Rigor and Reproducibility guidelines requires the development of an integrated platform to extract, model, and analyze provenance information to evaluate the reliability of previous findings. To the best of our knowledge, the ProvCaRe framework is the first medical informatics project that is implementing the “Rigor and Reproducibility” guidelines using the W3C PROV specification and provenance metadata analysis for reproducibility. In the next section, we describe the three-stage development of ProvCaRe knowledgebase, which is the first publicly available repository of provenance metadata extracted from published articles describing biomedical research studies.

**Method**

The ProvCaRe knowledgebase has two primary objectives:

1. Develop a repository of semantic provenance metadata extracted from published research studies that can be queried and analyzed to evaluate research studies; and
2. Enable researchers to perform hypothesis-driven search of previous research studies (with associated provenance information) and facilitate development of rigorously designed experiments for reproducible results.

These two objectives closely align with the aim of the NIH Rigor and Reproducibility guidelines (2, 3). We implemented the ProvCaRe semantic provenance knowledgebase in three stages (Figure 1 illustrates the ProvCaRe architecture). In the first stage, we developed the ProvCaRe S3 Model to represent provenance metadata for scientific reproducibility in biomedical research. In the second stage, we implemented a provenance focused NLP pipeline for entity identification, extraction, and generation of “provenance-triple” from published articles. In the third stage, we implemented the ProvCaRe semantic provenance knowledgebase with a powerful hypothesis-driven search and query functionality together with a new ranking algorithm that evaluates the provenance of research studies. We describe the details of each stage in the following sections.

1690
Stage I: Development of the ProvCaRe S3 Model for Biomedical Research. Many clinical researchers follow the Population, Intervention, Comparison, and Outcome (PICO) model to frame research questions and the PICO framework is an integral component of evidence-based medicine (EBM) (19). Similarly, the Ontology for Clinical Research (OCRe) project has characterized different phases of clinical research and formalized the relevant terms in the OCRe ontology (20). The ProvCaRe S3 model builds on these templates and defines three categories of provenance metadata, namely **Study Method**, **Study Tools**, and **Study Data**, which are important for scientific reproducibility. The three top-level S3 model terms can be mapped to the W3C PROV Ontology terms of prov:Activity (prov corresponds to the W3C PROV Ontology namespace http://www.w3.org/TR/prov-o/), prov:Agent, and prov:Entity. These three terms represent multiple subcategories of provenance metadata as follows:

1. **Study Method**: represents provenance metadata describing the protocol used to conduct the experiment (e.g., observational study or interventional study), the method used for recording of experimental data and the statistical analysis plan used for data analysis (e.g., covariance analysis or correlation analysis). Provenance information modeled by Study Method and its sub categories are important components in assessing reproducibility, as it helps identify potential differences in methods used to conduct research studies.

2. **Study Tools**: represents information describing the tools used for both data recording and data analysis. For example, electrocardiogram (ECG) data may be recorded using a bipolar, one lead, three lead, or a twelve lead electrocardiogram, which represents important contextual information required to ensure correct interpretation.
and analysis of the ECG data. The resolution of imaging instruments, recording parameters associated with an instrument, and related tool-specific metadata is essential for transparent reporting of experiment studies and evaluating the reliability of results.

3. **Study Data:** represents information describing the context of study data collections, for example timestamp of the recorded data, the different categories of study variables used (e.g., continuous variables or categorical variables), how data were collected (e.g., sitting, supine), and valid range of values for a study variable. The provenance metadata represented by Study Data and subcategories of this term allows researchers to accurately identify and extract the relevant subsets of study variables for replicability.

The ProvCaRe S3 model is formalized in the ProvCaRe ontology by extending the W3C PROV Ontology (PROV-O), which uses the description logic-based W3C Web Ontology Language (OWL2) (16). The ProvCaRe ontology (available at: [http://provcare.case.edu/](http://provcare.case.edu/)) extends the three core PROV-O classes to model the terms constituting the S3 model. The `prov:Activity` class models `provcare:StudyMethod`, `provcare:DataAnalysisMethod`, `provcare:DataCollectionMethod`, and `provcare:Intervention` as its subclasses (provcare refers to the ProvCaRe ontology namespace: [http://www.case.edu/provcare](http://www.case.edu/provcare)). These ontology classes are further extended to model additional provenance information, for example different categories of statistical data analysis techniques including `provcare:DescriptiveStatisticalAnalysis` and `provcare:InferentialStatisticalAnalysis` classes. Similarly, the W3C PROV-O `prov:Agent` class is extended to model `provcare:StudyInstrument`, and `provcare:StudyParticipant`. The `provcareStudyInstrument` models many different categories of data recording instruments including `provcare:Actigraph`, `provcare:Audiometer`, `provcare:Thermistor`, `provcare:ImagingInstrument` among other (Figure 2 illustrates a segment of the ProvCaRe ontology class hierarchy). To model the `provcare:StudyData` class, we extended the `prov:Entity` class and defined `provcare:StudyOutcome`, `provcare:StudyPopulationData`, and `provcare:TemporalEntity` classes.

It is important to note that given the extremely large number of provenance metadata terms used in biomedical research, it is not practical to model all provenance-related terms in the ProvCaRe ontology as precoordinated class expressions (21). Therefore, we have extended the postcoordination compositional grammar syntax defined for the Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) (21) to create new provenance-specific ontology class expressions in the ProvCaRe ontology for use across different domains. The new provenance-specific compositional grammar syntax enables us to re-use ontology classes from more than 500 existing biomedical ontologies listed at the National Center for Biomedical Ontologies (NCBO) (22) in the ProvCaRe ontology. For example, the expression `[Electroencephalogram]: [hadStudyInstrument] = [ScalpElectrode], [wasAssessedUsing] = [EpworthSleepinessScale]` represents provenance of electrophysiological signal recording in neurology and `[DayTimeSleepiness]: [wasScoredAt] = [CentralReadingCenter]` represents provenance of sleep medicine data measurement, which are defined using our ProvCaRe compositional grammar syntax. A detailed description of the ProvCaRe postcoordination expression syntax is provided in our paper submitted for review to AMIA 2017 titled “An Extensible Ontology Modeling Approach using Post Coordinated Expressions for Provenance Metadata”. The ProvCaRe ontology is used as the reference knowledge model to identify and extract provenance information from published research and generate structured provenance information.

**Stage 2. Automated Provenance Metadata Extraction from Research Study Articles.** Using the S3 model as reference, the ProvCaRe NLP pipeline extracts provenance information from published research studies and generates “provenance triples” for populating the ProvCaRe knowledgebase. The “provenance triples” conform to the provenance model used in the W3C PROV specification and allows search as well as query functionalities. For example, provenance information extracted from a research study that analyzed the association between sleep-disordered breathing and hypertension (23) can be represented as “DataAnalysis → wasPerformedUsing → GeneralizedEstimatingEquation(GEE)” and “SleepData→wasScoredAt→CentralReadingCenter”. These provenance
triples describe the provenance of data analysis as well as processing techniques used in the research study and can be aggregated to form a “provenance graph” as described in the W3C PROV specifications (10). The two end terms (e.g., DataAnalysis and GeneralizedEstimatingEquation (GEE)) represent the nodes of a graph and the intermediate term (e.g., wasPerformedUsing) corresponds to a relationship or an edge in the graph (10).

The identification and extraction of both the provenance terms and relations from unstructured biomedical text is a difficult task (automated extraction of structured information from text is a focus of extensive research in biomedical informatics (24)), therefore we used a combination of techniques to address this challenge. First, we extended cTAKES with additional ontology-lookup and provenance triple generation modules to process the published articles (15). In addition to cTAKES, we used the MetaMap tool, which uses the Unified Medical Language System (UMLS) as reference knowledge model (25), for named entity recognition (NER) task. In addition to MetaMap and cTAKES, the ProvCaRe-NLP tool uses the National Center for Biomedical Ontologies (NCBO) Open Biomedical Annotator tool (26) for NER task also (in previous section, Figure 1 shows the architecture of the ProvCaRe-NLP pipeline). We developed the ontology lookup module (OLM) to allow use of multiple OWL ontologies for provenance-specific NER tasks. The OLM uses the open source OWL Application Programming Interface (OWLAPI) to parse and index classes and properties in an OWL ontology (27).

### Table 1: Distribution of provenance terms identified by the NER modules of the three tools used together in the ProvCaRe-NLP pipeline

<table>
<thead>
<tr>
<th>Provenance Terms Identified by NER Modules of NLP Tools</th>
<th>ProvCaRe-cTAKES</th>
<th>NCBO Open Biomedical Annotator</th>
<th>MetaMap-UMLS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Study Method</td>
<td>425</td>
<td>115</td>
<td>169</td>
</tr>
<tr>
<td>Data Collection Method</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data Analysis Method</td>
<td>2209</td>
<td>31</td>
<td>42</td>
</tr>
<tr>
<td>Intervention</td>
<td>53</td>
<td>2</td>
<td>7</td>
</tr>
<tr>
<td>Comparison</td>
<td>79</td>
<td>162</td>
<td>0</td>
</tr>
<tr>
<td>Time Point</td>
<td>80</td>
<td>0</td>
<td>467</td>
</tr>
<tr>
<td>Study Data</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Study Variables</td>
<td>1453</td>
<td>3764</td>
<td>4490</td>
</tr>
<tr>
<td>Study Data Preprocessing</td>
<td>677</td>
<td>15</td>
<td>9</td>
</tr>
<tr>
<td>Cohort Variables</td>
<td>837</td>
<td>1119</td>
<td>1240</td>
</tr>
<tr>
<td>Study Tools</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Software Tools</td>
<td>77</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Statistical Model</td>
<td>957</td>
<td>63</td>
<td>5</td>
</tr>
<tr>
<td>Measurement/Recording Instrument</td>
<td>113</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1 describes the distribution of the 28,474 provenance triples extracted from the 75 published articles describing sleep research studies. Two experts working in the NSRR project selected 75 published articles describing 6 NIH-funded sleep medicine research studies that have made their data available through NSRR: (1) Sleep Heart Health Study (SHHS1) Study: SHHS1 is a multi-center cohort study to determine the consequences of sleep-disordered breathing on risk for coronary heart disease, stroke, all causes of mortality, and hypertension; (2) Multi-Ethnic Study of Atherosclerosis (MESA) Study: The MESA study is a multi-center project involving 6 field centers to study the characteristics of subclinical cardiovascular disease (CVD) and risk factors that predict progression to clinical CVD; (3) Childhood Adenotonsillectomy Trial (CHAT) Study: The CHAT study is a multi-center, single blind, randomized controlled trial to test whether children aged between 5 to 9.9 years with mild to moderate obstructive sleep apnea randomized to early adenotonsillectomy (eAT) show greater neurocognitive functioning as compared to children randomized to watchful waiting plus supportive care after a 7 month observation period; (4) Heart Biomarker Evaluation in Apnea Treatment (HeartBEAT) Study: The HeartBEAT study is a multi-center Phase II randomized controlled trial to evaluate the effects of supplemental nocturnal oxygen or Positive Airway Pressure (PAP) therapy compared to optimal medical preventive therapy over 3 months intervention period; Osteoporotic Fractures in Men (MrOS) Study: The MrOS study is a large observational study of the determinants of fracture in older men involving 5994 participants living in communities with age of 65 years or older across 6 clinical centers; and (5) Cleveland Family Study (CFS): CFS is a family-based study of sleep apnea worldwide 2,284 individual enrolled from 361 families and studied on 4 occasions over a period of 16 years designed to identify genetic and non-genetic risk factors for sleep apnea and related traits.
Algorithm for Generating Provenance Triples. We have developed a new algorithm to construct provenance triples from cTAKES output using the annotation functionalities of the clearNLP semantic role labeler (SRL) module (15). Specifically, the SRL module generates a “predicate-argument” structure for each sentence in the text based on a transition-based SRL approach (28). For example, two terms *measurements* and *parent* occurring in a sentence “…measurements of height and the circumference of waist and hips [argument] were obtained from [predicate] parent study [argument]…” are linked by the predicate *were obtained from*. Using a set of heuristics-based rules, the provenance triple generator algorithm uses the position of the arguments in the sentence with respect to the predicate to generate triple structure *Measurements → wereObtainedFrom → ParentStudy*. After processing the 75 published articles related to the six sleep research studies, the algorithm generated a total of 28,474 provenance triples, which are stored in the ProvCaRe knowledge for query and analysis.

Stage 3. ProvCaRe Knowledgebase: Hypothesis-driven Search and Query of Provenance Metadata. To meet the two key objectives of the ProvCaRe project, that is: (a) allow researchers to search for research studies related to a hypothesis based on their provenance information; and (b) enable researchers to develop rigorously designed experiments, we implemented a Web-based user interface for querying the ProvCaRe knowledgebase (Figure 3 shows the different features of the user interface, which is available at: [http://provcare.case.edu](http://provcare.case.edu)). The ProvCaRe knowledgebase consists of the provenance triples generated during Stage 2 (described above) that are indexed by their source text articles, which allows users to query for provenance metadata of individual research studies. The ProvCaRe interface supports hypothesis-driven search functionality that allows users to form query expressions corresponding to their research hypothesis. To help users to formulate accurate query expressions, the user interface supports an “auto complete” feature that is similar to functionalities offered in Google or Bing search engines (Figure 3 illustrates the auto complete feature of the user interface). The auto complete feature is implemented using a powerful open source text-indexing tool called Apache Solr, which supports search queries over unstructured text using efficient inverted indexing structures (29).

Using the high-performance Apache Solr indexing system enables the ProvCaRe knowledgebase to quickly identify relevant publications for a given set of keywords in a hypothesis. After a relevant set of publications is identified for a user-specified research hypothesis, a new provenance-based ranking algorithm ranks the result of the user query. The provenance-based ranking algorithm incorporates the frequency of provenance terms extracted from the published studies and ranks studies with a greater number of provenance triples higher in the search results. The weight assigned to the provenance triples extracted from each study represents the potential for scientific reproducibility of a research study. The default setting of the ranking algorithm assigns equal weight to the three components of the S3 model namely, Study Method, Study Tools, and Study Data. However, different weights can be assigned to different terms of the S3 model based on user preferences, which will allow users to find research studies that have more provenance information about their experiment protocol (Study Method), instrument details (Study Tools), or dataset (Study Data). In the next section, we discuss additional results and an evaluation of the ProvCaRe knowledgebase search and query feature.

Results and Evaluation

Provenance Metadata Extracted from Published Research Studies. Table 2 shows the distribution of the provenance terms extracted from the 75 published articles corresponding to the three components of S3 model. We note that all 75 published articles include a description of “Study Variables”, whereas only 43 articles include
provenance information related to temporal information (“Time Point”). Similarly, all 75 publications include provenance metadata describing the method used for data analysis and statistical model, which is encouraging as both the details of the data analysis techniques and the statistical model are critical for ensuring transparency in reporting of studies as well as reproducibility.

Table 2: Distribution of provenance terms and their occurrence in the 75 published articles used to create the ProvCaRe knowledgebase

<table>
<thead>
<tr>
<th>ProvCaRe S3 Model</th>
<th>S3 Model Subcategories</th>
<th>Instances of Provenance Terms</th>
<th>Number of Provenance Terms in Published Articles</th>
</tr>
</thead>
<tbody>
<tr>
<td>Study Method</td>
<td>Data Collection Method</td>
<td>425</td>
<td>6%</td>
</tr>
<tr>
<td></td>
<td>Data Analysis Method</td>
<td>2209</td>
<td>32%</td>
</tr>
<tr>
<td></td>
<td>Intervention</td>
<td>53</td>
<td>&gt;1%</td>
</tr>
<tr>
<td></td>
<td>Comparison</td>
<td>79</td>
<td>1%</td>
</tr>
<tr>
<td></td>
<td>Time Point</td>
<td>80</td>
<td>1%</td>
</tr>
<tr>
<td>Study Data</td>
<td>Study Variables</td>
<td>1453</td>
<td>21%</td>
</tr>
<tr>
<td></td>
<td>Study Data Preprocessing</td>
<td>677</td>
<td>10%</td>
</tr>
<tr>
<td></td>
<td>Cohort Variables</td>
<td>837</td>
<td>12%</td>
</tr>
<tr>
<td>Study Tools</td>
<td>Software Tools</td>
<td>77</td>
<td>1%</td>
</tr>
<tr>
<td></td>
<td>Statistical Model</td>
<td>957</td>
<td>14%</td>
</tr>
<tr>
<td></td>
<td>Measurement/Recording Instrument</td>
<td>113</td>
<td>2%</td>
</tr>
</tbody>
</table>

Table 3 shows the 15 most frequent provenance terms in the ProvCaRe knowledgebase and the corresponding ProvCaRe ontology term. A comparison with Table 2 shows that provenance terms corresponding to the data analysis techniques and statistical models used in research studies match the ontology classes modeled in the ProvCaRe ontology. The provenance term Study describing the experiment constitute a little less than one fourth (22%) of the provenance terms in the ProvCaRe knowledgebase. It is interesting to note that although 71 of the 75 articles include provenance information about the software tools used in a study (Table 2), the frequency of occurrence of provenance term describing instrument (including hardware instruments) is only 2% (Table 3).

Table 3: The top 15 provenance terms, their frequency, and mapping to the S3 Model categories, and representative ontology terms

<table>
<thead>
<tr>
<th>Provenance Term</th>
<th>Frequency</th>
<th>S3 Model Category</th>
<th>Representative Ontology Term</th>
</tr>
</thead>
<tbody>
<tr>
<td>Study</td>
<td>22%</td>
<td>Research Study</td>
<td>ResearchStudy</td>
</tr>
<tr>
<td>Population</td>
<td>8%</td>
<td>Population</td>
<td>StudyPopulation</td>
</tr>
<tr>
<td>Outcome</td>
<td>8%</td>
<td>Outcome</td>
<td>StudyOutcome</td>
</tr>
<tr>
<td>Activity</td>
<td>6%</td>
<td>Data Collection/Analysis Method</td>
<td>Activity</td>
</tr>
<tr>
<td>Average</td>
<td>6%</td>
<td>Statistical Model</td>
<td>Statistical Mean</td>
</tr>
<tr>
<td>Data</td>
<td>5%</td>
<td>Study Data</td>
<td>StudyData</td>
</tr>
<tr>
<td>Method</td>
<td>4%</td>
<td>Study Method</td>
<td>StudyMethod</td>
</tr>
<tr>
<td>Prevalence</td>
<td>4%</td>
<td>Statistical Model</td>
<td>StatisticalMethod</td>
</tr>
<tr>
<td>(Study) Design</td>
<td>4%</td>
<td>Data Collection Method</td>
<td>StudyDesign</td>
</tr>
<tr>
<td>Association</td>
<td>3%</td>
<td>Data Analysis Method</td>
<td>AssociationAnalysis</td>
</tr>
<tr>
<td>Median</td>
<td>3%</td>
<td>Statistical Model</td>
<td>Median</td>
</tr>
<tr>
<td>Estimate</td>
<td>2%</td>
<td>Statistical Model</td>
<td>EstimationAnalysis</td>
</tr>
<tr>
<td>Collection</td>
<td>2%</td>
<td>Data Collection Method</td>
<td>Collection</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>2%</td>
<td>Statistical Model</td>
<td>StandardDeviationMeasure</td>
</tr>
<tr>
<td>Instrument</td>
<td>2%</td>
<td>Study Tool</td>
<td>StudyInstrument</td>
</tr>
</tbody>
</table>

Hypothesis-driven Study Search and Query Performance. To evaluate the hypothesis-driven search feature of the ProvCaRe interface, we used a set of five hypotheses defined by a clinical domain expert:

**Hypothesis 1**: Sleep disordered breathing is associated with an increased incidence of cardiac arrhythmias.

**Hypothesis 2**: Early adenotonsillectomy improves neurocognitive outcomes in children with obstructive sleep apnea.

**Hypothesis 3**: Specific genetic polymorphisms are associated with an increased risk of sleep disordered breathing;
Hypothesis 4: Sleep disordered breathing is associated with an increased risk of developing the metabolic syndrome characterized by changes in markers of systemic inflammation; and

Hypothesis 5: Short sleep duration is associated with increased cardiovascular morbidity and mortality

We performed hypothesis-based search in the ProvCaRe platform using the above listed hypothesis and evaluated the performance of the ProvCaRe knowledgebase with respect to: (a) time taken to query and retrieve relevant research studies, and (b) time taken to rank the research studies using the provenance information extracted from these research studies. Figure 4 shows the performance of the provenance-ranking module, which clearly shows that the additional overhead for provenance-based ranking of the research studies is less than the time required to search for research studies that are relevant to a given hypothesis. This demonstrates that the additional functionality of the ProvCaRe knowledgebase to allow users to search and query for previous research studies and rank the results based on their provenance metadata is practical and does not lead to additional computing costs. We also note that query performance across the five research hypotheses does not vary significantly due to the effectiveness of the inverted indexing feature of Apache Solr. This is important for ensuring the scalability of the ProvCaRe knowledgebase as additional studies are processed and added in the future.

Discussion and Conclusion

The Resource Description Framework (RDF) is a well known and widely used graph-based data structure for ontology-based datasets (30). The RDF/XML syntax is often used to represent provenance metadata, which naturally forms a graph structure (10). Therefore, as part of our ongoing work, we plan to transform the “provenance triples” currently stored in the ProvCaRe knowledgebase into RDF triples, which will allow us to implement the SPARQL query language for advanced query and retrieval of provenance information. In addition to the use of RDF, we are using the different components of the ProvCaRe project to process and extract provenance information from all publications related to sleep medicine research studies. We are developing an automated pipeline to retrieve new publications available in both NCBI PubMed and PubMed Central, which will be processed using the ProvCaRe-NLP pipeline. This will allow users to search and query for provenance information of newly published research studies. In addition to extending the ProvCaRe knowledgebase, we are developing new query algorithms for provenance graphs to efficiently compare the provenance information of multiple research studies.

In conclusion, the ProvCaRe knowledgebase described in this paper is one of the first publicly available repositories of provenance metadata extracted from published biomedical research studies. The ProvCaRe S3 Model serves as key provenance knowledge model for representing provenance terms and the ProvCaRe ontology enables the identification of provenance information necessary to support scientific reproducibility in biomedical research. Using the ProvCaRe-NLP pipeline, we demonstrate an efficient and scalable approach to automatically identify and extract provenance metadata information from 75 published articles corresponding to sleep medicine research studies that have made their data available through the NSRR. Users can search and query the provenance metadata of research studies using an intuitive Web-based ProvCaRe interface that uses a provenance-based ranking algorithm to search and retrieve relevant study related articles. The various components of the ProvCaRe project are allowing us to support the objectives of the NIH Rigor and Reproducibility guidelines by developing and implementing a set of practical tools to model, extract, and analyze provenance information associated with research studies.

Acknowledgement. This work is supported in part by the NIH-NIBIB Big Data to Knowledge (BD2K) 1U01EB020955 grant, NSF grant# 1636850, and the NIH-NHLBI R24HL114473 grant.

References
Using a community-engaged health informatics approach to develop a web analytics research platform for sharing data with community stakeholders

Karen H. Wang, MD MHS¹, Luis Marenco, MD¹,³ Johanna Elumn Madera, PhD ² Jenerius A. Aminawung, MD, MPH¹, Emily A. Wang, MD MAS¹, Kei-Hoi Cheung, PhD¹,³

¹Yale University School of Medicine, New Haven, CT ²Yale University School of Public Health, New Haven, CT ³VA Connecticut Healthcare System, West Haven, CT.

Abstract
Marginalized communities are rarely included in the planning of research relevant to their own health or have access to research data. We implemented a community-based participatory research approach in developing a new health informatics system, WARP (Web Analytics Research Platform), to enable stakeholders to access and analyze research data. We leveraged data from a cohort study of 751 patients in the Transitions Clinic Network (TCN), a network of clinical programs serving patients with a history of incarceration. WARP holds de-identified, patient data, streamlines data processing (i.e. transformation, archival, and partitioning), and has a web analytic tool for users to perform statistical analyses. We used feedback from focus groups of patients with a history of incarceration and workshops with TCN research teams, including patients, community health workers and policymakers, to develop WARP. Our approach advances mechanisms to engage stakeholders in research. Future work will evaluate its effect on community-engagement in research.
Patients are rarely engaged in research design or planning of studies affecting their own health. This lack of engagement in research delays the translation of research into actual improvements in health for patients, since the dissemination of evidence-based information often takes 10 to 25 years. Community-based participatory research (CBPR), which includes community members in every step of the research process, promises to facilitate the more efficacious adoption of findings into action, especially among marginalized communities. Yet, CBPR’s potential in analysis and dissemination is largely untapped. Even CBPR projects often fail to engage all stakeholders in analysis or disseminate study findings because limited statistical, analytic, or technologic experience among stakeholders. In addition, researchers using a CBPR approach do share data with community partners or participants until analyses have been completed.

The CBPR approach in informatics is in its nascency. One recent manuscript highlighted the projects that utilized both CBPR and informatics approaches to engage community stakeholders. The authors summarized that informatics researchers have primarily applied CBPR to use technology to explore and understand health-related problems in marginalized communities, the technologic practices in a specific community, or to develop context-appropriate technologic interventions. In these studies, there was increased participation of community stakeholders in the design and development of informatics platforms to address the health needs of communities. In one study, researchers shared data with participants, but only after analyses have been completed, e.g. working with communities in the development of infographics to display health data.

To improve engagement of patients in research, we use a CBPR approach to create a web-based platform that enables patients to both access and analyze de-identified data. Our use case population to improve engagement is among the most marginalized populations, i.e. those with a history of incarceration, who are rarely included in the planning of research relevant to their own health or have access to research data. We developed and designed, with optimal input from patients with a history of incarceration, a functional prototype of a web-based platform that enables exploration of a population’s health at a local and national level. We leveraged data from the Transitions Clinic Network (TCN), a network of clinics providing specialized services to individuals returning home from prison. In this paper, we describe our research setting and approach and the web-based platform that enables access to and analysis of research data.

**Methods**

**The Setting**

The Transitions Clinic Network (TCN): Founded in 2006, the TCN is the largest national collaboration of patients, clinician researchers, and policymakers whose primary focus is improving the health of individuals returning home from prison. The TCN provides healthcare based on a three-pronged approach: (1) community health workers with a history of incarceration as care coordinators; (2) early coordination of care that begins prior to release from prison to identify patients with chronic medical conditions and facilitate improved care coordination; and 3) culturally-competent care specific to people returning to the community from prison to reduce stigma and distrust of medical professionals. TCN is guided by local community advisory boards comprised mostly of individuals with a history of incarceration and a National Advisory Board whose expertise includes leadership in the Association of State Correctional Administrators and civil rights and advocacy organizations for formerly incarcerated individuals, like All of Us or None and JUSTUSA.

**TCN Cohort Study:** The TCN has established the largest prospective cohort of individuals with chronic medical conditions leaving prison as part of a Health Care Innovation Award from the Center for Medicare & Medicaid Innovation. Data were collected from participants using surveys (collected by tools such as SalesForce© and Qualtrics© prior to our web prototype development), and electronic medical records at 13 clinics and community health centers nationwide, located in: New Haven, CT; San Francisco, CA; Richmond, CA; San Jose, CA; Boston, MA; New York City, NY; Bronx, NY; Rochester NY; Baltimore, MD; Birmingham, AL; and Caguas, PR. Data collected include: patient sociodemographics, housing, employment, incarceration history, past medical diagnoses, health outcomes (including physical, mental health, and substance misuse), health care utilization, and re-incarceration in three-month increments. To be eligible for the TCN cohort, a person has to have been released from prison within 6 months, have a chronic medical condition (including chronic physical, mental health, or substance use conditions) which warrants longitudinal primary care or be older than 50 years of age, speak English or Spanish, and not plan to leave the area for at least 12 months. The TCN Cohort Study enrolled 751 participants, some of whom were followed for more than 12 months.
The SHARE project, funded by the Patient-Centered Outcomes Research Institute (http://www.pcori.org/research-results/2015/share-project) (PCORI) is a sub-study within the TCN whose main goals are to integrate the principles of CBPR and patient centered outcomes research to improve patient outcomes by engaging patients in the research process. One prong of the SHARE project is the development of a web-based platform to share de-identified TCN cohort data that is accessible to patients and other stakeholders. We used baseline data from the TCN cohort study of returning individuals to pilot the development of the web-based platform prototype.

The Web Analytics Research Platform (WARP)

Development Approach and Methods:

Community Participation: The SHARE project created a study advisory board of patients with a history of incarceration and a community health worker from the New Haven Transitions Clinic program. They were engaged in the development of the idea and submission of the PCORI grant. With grant funding, they were then involved in brainstorming the functionalities of WARP, gave input on the display of the data, and participated in curriculum development by participating in focus groups.

These advisory board members were selected to represent the diversity of the TCN patients. There were seven individuals identified by the TCN clinic staff who were available and willing to participate and are reflective of the diverse TCN population. There were 3 women and 4 men, whose self-identified ethnicity/race categories were Black, Latino/a, and white. Focus group participants’ employment ranged from unemployed to health professionals, their ages ranged from 32-67, and they reported having limited to proficient computer usage.

Data: We used the TCN Cohort Study data (Table 1) to develop and design the web-based platform prototype to share de-identified data with end-user stakeholder groups (patients with a history of incarceration, researchers, providers, and policy makers). Table 1 shows some of the data types in the study.

<table>
<thead>
<tr>
<th>Table 1. Example of Data Elements from TCN</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Domain</strong></td>
</tr>
<tr>
<td><strong>Components/method of assessment</strong></td>
</tr>
<tr>
<td><strong>When Assessed</strong></td>
</tr>
<tr>
<td><strong>Sociodemographics</strong></td>
</tr>
<tr>
<td>Age, sex, race/ethnicity, level of education, housing, employment, income; self-report</td>
</tr>
<tr>
<td><strong>Source of referral</strong></td>
</tr>
<tr>
<td>Prison/community health worker/community based organization/self; self-report</td>
</tr>
<tr>
<td><strong>Healthcare utilization outcomes</strong></td>
</tr>
<tr>
<td>Primary care utilization</td>
</tr>
<tr>
<td>Hospitalization (reason)</td>
</tr>
<tr>
<td><strong>Criminal Justice outcomes</strong></td>
</tr>
<tr>
<td>Re-incarceration</td>
</tr>
<tr>
<td><strong>Health-related data</strong></td>
</tr>
<tr>
<td>Self-reported health</td>
</tr>
<tr>
<td>Health Literacy</td>
</tr>
<tr>
<td>History of solitary confinement</td>
</tr>
<tr>
<td>Report of discrimination</td>
</tr>
<tr>
<td>Depression symptoms</td>
</tr>
<tr>
<td>Smoking</td>
</tr>
<tr>
<td>Alcohol use</td>
</tr>
</tbody>
</table>

Methods: We used the rapid cycle prototyping approach and adapted the principles of participatory design to understand anticipated user requirements and research needs so that the needed functionality would be implemented rapidly and iteratively.15-17
From March to September 2016, we convened three focus groups of four to seven advisory board members for end-user input and feedback. Topics of discussion included the design of the interface, user-friendliness, and their understanding of the data summaries (e.g. prevalence and correlations) and relationships between two variables.

The focus groups lasted two and a half hours. For the first focus group, participants reviewed different types of images for displaying data (e.g. icon arrays, graphs, pie charts) and chose the ones that were easiest to understand and the types of images they preferred. The second and third focus groups introduced the latest versions of the web-based platform. For each focus group, we had an interview guide to elicit participant responses about the platform. The interview guide had general questions to elicit initial thoughts about the platform, such as “what do you think about the way the website looks,” “what do you think about the way the website functions,” “what are some things that you would like to see changed,” and “is there anything you would like to see added to the site that is not there now?” Participants were asked to demonstrate how they would find specific information, such as finding information about people in the cohort who are male, have diabetes, or length of prison term. Then they were asked to provide an answer to a specific question about the information they found. For example, one question was “how would you look at the length of prison term and those who screened positive for depression” followed by “which group had the most people who screened positive for depression.” A social worker trained in qualitative research methods whose own research focuses on criminal justice-involved populations conducted each of the focus groups. The social worker took detailed notes from the focus groups.

The web-platform was changed based on feedback from each focus group to ensure that people with varying degrees of computer and data literacy could use it. In addition to the focus groups, four TCN teams made up of patients and community health workers with a history of incarceration, policymakers, and researchers provided feedback about the web-based prototype and used it to begin to develop a research question for future work.

Description of System

![System Overview](image)

**Figure 1. System Overview**
Figure 1 depicts the main components of our data and analytics system. It consists of the following components:

1. **Data conversion tool.** For the TCN cohort study, patients’ data (e.g., surveys) had been collected by each TCN site using tools such as SalesForce© and Qualtrics©. We then integrated all TCN study data collected from these tools into the TCN database and used this to develop WARP. The TCN datasets were exported using SPSS format, a metadata-rich format for statistical data analysis. We developed a data conversion tool to transform and load SPSS datasets into the EAV/CR (Entity Attribute Value with Classes and Relationships) data model. EAV/CR is derived from the triple-store model used in clinical trials and was enhanced to deal with domain variability. EAV/CR data model allowed for creating several virtual databases with their own set of administrative users. Each study or version of it (by cut time or partition) was stored as a separate EAV/CR database. During transformation, each instrument was converted into a separate EAV/CR Class. Elements on each Class were related to others using internal object IDs from the Subject (pivot) Class. Classes had attributes that were related to questions in the surveys and supported strong data types. The EAV/CR database enabled capturing of all questions’ metadata (data type, labels, option text, value, boundaries, etc.). Once the data was transferred, as part of the data transformation process, we performed data harmonization by mapping local variable names/types to common variable names/types.

2. **Centralized database.** Our EAV/CR database was implemented into a SQL Server 2016 relational database. It provided all system required data capabilities for: a) data warehouse, b) system administration, c) security, and d) online analytic processing.

3. **Metadata-driven framework and special enhancements to enable WARP capabilities.** Our prototype required extensible data and system needs that can evolve over time. For these capabilities, we leveraged and extended the flexible metadata driven approach within EAV/CR programming framework. This framework provided extensible metadata description of datasets including their structure, mappings (between local data elements and common data elements) and metadata-driven user interface configuration. The system provided a web-based dashboard for the system developer to view/modify the metadata. EAV/CR enhancements to support special WARP capabilities include but are not limited to:
   a. Vertical security: Data elements can be made visible depending on user role. We used metadata to configure which data fields are included as part of a query and to display query output based on user roles (e.g., patient role vs. researcher role vs. policymaker role). However, for this project, this functionality was removed as the research teams wanted equitable access to the data.
   b. HIPAA values: Data that could potentially be identifiable cannot be queried. We defined metadata to hide questions and used annotations to limit or mask values that will make information identifiable. For example, no results were provided when the total number of individuals with a particular characteristic were less than ten or when a TCN site had less than 100 individuals.
   c. Data type conversions: We developed a tool to convert attribute data types. For example, this tool can convert values from questions that are provided as a string into an integer or a date.
   d. Question groups: For improved data display and analysis, we clustered questions together. A new type of attribute was created (grouping attribute) to enable this functionality.
   e. Metadata analysis tools: We developed tools to analyze questions for misspelled values, completeness, and data type validation.
   f. Metadata curation and data normalization tools: We developed tools to allow attribute data curation using global value replacements (used to correct invalid or misspelled values) and normalization (to spawn an attribute into a new class).

4. **Web interface for data queries and visualization.** We created the “WARP Navigator.” This interface is used to display the data and statistical findings for a study. The interface consists of a variable list on the left panel and a graphical dashboard on the right panel (See Figures 2 and 3). The variables list shows all questions using an accordion control, which collapses and expands content domains within the survey. Some questions are hierarchically arranged and placed in groups. The graphical dashboard provides controls to choose: a) type of graph (bar/line/pie), b) type of values (numbers/percentages), and c) style: color/pattern. In addition, bivariate analysis is available to compare two variables. The WARP Navigator is dynamically driven and configured based on the metadata given by the study protocol and security constrains on user roles. For graphics, we used the JS chart component by amCharts (https://www.amcharts.com/).
Figures 2 and 3 illustrate the basic components of WARPs web-interface. Users can select from the list of variables on the left side of the display to describe the population, and the right side of the display depicts the format in which the data are presented. The image on the left is prior to end-user input about the display and interactive capabilities. The image on the right is after incorporating user input, demonstrating different graphics and more functionalities. WARP allows the display of data as bar graphs or pie charts and has functionalities for users to change the display (e.g. view the data as counts or percentages). Users can also change the font size and color of the display, and there is a screen reader for those who prefer to listen to the text.

Figure 2. Data display pre-and post-user feedback

Figure 3 demonstrates an interactive function in which an end-user can conduct a bivariate analysis. For example, an end-user who was interested in looking at the relationships between the number of years a person is incarcerated to the number of chronic diseases would see the display in Figure 3.

Figure 3. Data analysis functionality in WARP
Discussion

Our work highlights 1) a novel mechanism to advance the engagement and authentic participation of communities in research by providing a user-centered platform in which consumers can directly interact with (i.e. visualize and analyze) research data and 2) the importance of incorporating stakeholders’ feedback and perspectives into the design and development of the user interface early in the process.

We believe through the integration of CBPR principles in the development of this web-based platform, we have engaged patients who have traditionally been marginalized from the research process, increased the capacity of end-users to use the data, and have shared data back to end-user groups who will be instrumental in catalyzing improvements in health outcomes for this population. We believe this platform enables TCN patients to become more engaged in data collection and participate equitably in the analysis and interpretation of data. Although more research is needed, this approach and the web prototype initiates a mechanism to enrich our understanding of how to improve the health of this population and other marginalized communities. WARP allows for transformation of research data from survey instruments and other formats that have been processed by software, such as SPSS, to be accessible to end-users for analysis. This functionality can be utilized to support the transformation of other types of data from the TCN and other research studies. In addition, WARP Navigator allows end-users to conduct bivariate analysis via the web-platform. WARP will be used by TCN sites and their patients to analyze, interpret, and generate new hypotheses and identify strategies or potential targets to improve the health outcomes of patients with a history of incarceration. This type of data access and interactive capabilities customized to patient end-users shortens the time between data collection and dissemination and may enable more rapid translation of research data into meaningful changes for communities.

One of the main lessons we learned as a part of this work was developing a mechanism for a common understanding and communication between our interdisciplinary team members consisting of patients and community health workers with a history of incarceration, clinician researchers, informatics researchers, and a systems architect. In the initial development of the interactive platform and the data visualization, researchers proposed the use of iconographs to display data based on existing literature, however the patient end-users preferred basic graphs. Through this work, our feedback loop from patients to researchers to systems architect was continuous and frequent to ensure patient feedback was incorporated accurately. Another lesson was the need to transform research data in its different formats for availability via the web-platform. While we attempted to harmonize the data elements (i.e., variables including their names and value sets) within the TCN, there is a broader common data element effort taking place to harmonize variables used by different research projects across different networks and research communities. For example, the NIH Common Data Element (CDE) resource portal (https://www.nlm.nih.gov/cde/) consists of a CDE repository that provides access to CDEs and forms that group certain CDE’s together for research questions. Our project can potentially benefit from and contribute to the CDE effort.

We have already experienced and anticipate several challenges with the refinement and dissemination of the platform. One challenge is and will continue to be balancing the needs of diverse platform user groups—from patients to policy makers to other researchers. As the goal of the platform is to serve as a tool to produce patient-centered research by enabling full interaction of patients with the platform, we prioritized the perspective of patients with a history of incarceration and accommodated their specific needs. Therefore, all users have access to the same data; and the visualization and interactive capacities are targeted to address their needs and not the researchers or policymakers. Another challenge will be continued integration of new TCN data into the platform and ensuring the continued use and improvement of the platform.

This team approach and platform prototype are only part of the initial stages to meet our objective to build collective capacity among patients, community health workers with a history of incarceration, policymakers, and clinician researchers to improve the health of people returning to the community from prison. We have outlined several more stages in this project in Figure 4, including the integration of publicly available data with the patient-level data.
Interested end users with limited health, data, or research literacy, and other stakeholders such as correctional and community health policymakers

Enhanced CBPR to activate patients and engage communities

Rich, rigorous PCOR with outcomes that matter to hard-to-reach populations

Better understanding of how to improve the health of former prisoners

Integration of patient-level data with publicly available health data, relevant to individuals with history of incarceration

Data platform that incorporates local and national health data of TCN patients with publicly available health data and allows users to view aggregate and individual patient data

Functional prototype to enable exploration of a population's health at local and national level

Workshop on CBPR, data and use of data

• Organizing data from prospective cohort study of the Transitions Clinic Network (TCN)
• Visualization of data based on patient needs and preferences

• Fully developed sharable WARP system in other populations
• Model for using a CBPR in health informatics research to engage patients
• CHW who are data informed

Current Stage

Figure 4. Community-based participatory research merged with health informatics

Since the finalization of the prototype, we have conducted a three-day in-person training session about CBPR and patient centered outcomes research principles, research design, and partnership development for the TCN teams, with curriculum adapted from Data & Democracy. In addition, one half-day session was focused on the importance and use of data in research and the use of WARP. We also held four webinars focusing on research design, creation of patient-centered research proposals, community-engagement, and using research for advocacy. With more training, we anticipate increased use of the system and an increased capacity for collaboration that will enable new patient-centered outcomes research ideas. Our immediate plans for the WARP navigator include: adding drill-down filtering of results, user annotations, and sharing of graphical queries. While the current web interface supports data queries only, it can potentially be expanded to support data entry as part of the collection process in the future. We are also in the process of linking WARP to the main webpage for the TCN. This platform will be accessible to the TCN network only with user log-ins. Once available on the TCN webpage, our plan is to conduct follow-up studies to evaluate the use of this system among the TCN population.

Conclusion

Data sharing is occurring in multiple different forms in the biomedical sciences, mostly among researchers with limited consideration of patient stakeholders, especially from marginalized communities. The development of WARP with basic functionalities has the potential to allow researchers to share and feedback data to the communities who contributed to the data. Moreover, WARP is a tool that can potentially engage patients in research by providing a mechanism to access and analyze the data. Alongside the curriculum to increase research capacity, these capabilities may help accelerate the translation of research into actual improvements in health for patients. We believe this work advances the work at the intersection of community-engaged health informatics and CBPR and will have broader applicability to researchers employing a CBPR approach or those employing methods of patient-centered research outcomes. We will explore the utility of this web-based platform for other research programs and conceive of future functionalities, including a mobile platform and a prospective data entry platform for transparency and rapid dissemination.
References

Drug repositioning for prostate cancer: using a data-driven approach to gain new insights

QuanQiu Wang 1, PhD, Rong Xu 2
1 ThinTek, LLC, Palo Alto, CA 94306
2 Department of Epidemiology and Biostatistics, School of Medicine, Case Western Reserve University, Cleveland OH 44106

Abstract

Prostate cancer (PC) is the most common cancer and the third leading cause of cancer death in men worldwide. Despite its high incidence and mortality, the likelihood of a cure is low for late-stages of PC. There is an unmet need for more effective agents for treating PC. Here, we present a drug repositioning system, GenoPredict, for finding innovative drug candidates for treating PC. GenoPredict leverages upon a large amount of disease genomics data and a large-scale drug treatment knowledge base (TreatKB) that we recently constructed. We first constructed a genetic disease network (GDN) that comprised of 882 nodes and 200,758 edges and applied a network-based ranking algorithm to find diseases from GDN that are genetically related to PC. We developed a drug prioritization algorithm to reposition drugs from PC-related diseases to treat PC. When evaluated in a de-novo prediction setting using 27 FDA-approved PC drugs, GenoPredict found 25 of 27 FDA-approved PC drugs and ranked them highly (recall: 0.925, mean ranking: 27.3%, median ranking: 15.6%). When compared to PREDICT, a comprehensive drug repositioning system, in novel predictions, GenoPredict performed better than PREDICT across two evaluation datasets. GenoPredict achieved a mean average precision (MAP) of 0.447 when evaluated with 172 PC drugs extracted from 172,888 clinical trial reports, representing a 164.5% improvement as compared to a MAP of 0.169 for PREDICT. When evaluated with 72 PC drugs extracted from 43,811 ongoing clinical trial reports, GenoPredict achieved a MAP of 0.278, representing a 231.1% improvement as compared to a MAP of 0.084 for PREDICT.

The data is publicly available at: http://nlp.case.edu/public/data/PC_GenoPredict and http://nlp.case.edu/public/data/treatKB.

Introduction

Prostate cancer (PC) is the most common cancer and the third leading cause of cancer death in men worldwide. In the United States alone, an estimated 238,590 men were diagnosed with PC in 2013 and 29,720 died from their disease [1]. Despite its high incidence and mortality, the likelihood of a cure is low for patients diagnosed with late-stage PC [2]. Docetaxel, the standard treatment for metastatic PC, provides a median overall survival improvement of roughly 3 months. While the PC market is expected to grow to $8.2 billion in 2019, there is an unmet need for more effective drug treatments for PC [2].

Computation-based drug repositioning approaches that assimilate vast amounts of genetic, genomic, chemical, and phenotypic data for thousands of drugs and diseases can greatly speed up the traditional drug discovery process [3]. Here we present a computational drug repositioning system, GenoPredict, that capitalizes on comprehensive disease genetic data generated by Genome-Wide Association Studies (GWAS) [4] and a unique large-scale drug treatment database that we recently constructed [5-7]. Our study is based on the premise that the genetic inter-relationships among diseases often reflect pathophysiological relevance. Though the majority of such shared pathophysiological features remain unknown, treatment insights from one disease may be used to inform our knowledge of others and potentiate their treatments. Existing approaches have directly inferred drug targets from disease genetics generated by GWAS studies [8-11]. Here, we present a complementary strategy to indirectly infer repositioned drug candidates from disease genomics data and demonstrated its utility in PC drug repositioning: diseases that share high genetics with PC were used as a starting point for both discovering repositioned drug candidates and gaining insights into the common mechanisms of action underlying identified candidates.

In order to systematically reposition drug treatments from one disease to another, it is critical to have a comprehensive drug treatment knowledge base. We have recently constructed a large-scale drug-disease treatment knowledge base
(TreatKB) from multiple heterogeneous and complementary data resources [5-7]. We previously showed that TreatKB is critical for our computational drug repositioning tasks [12-13]. We evaluated GenoPredict in identifying and prioritizing FDA-approved PC drugs and compared GenoPredict to PREDICT, a comprehensive drug repositioning system [14], in novel predictions for PC. We demonstrated that GenoPredict is effective in prioritizing FDA-approved PC drugs and achieved better performance than PREDICT in novel predictions.

Materials and methods

Data

Drug-disease treatment relationship knowledge bases (TreatKBs) TreatKB contains 208,330 unique drug-disease treatment pairs, representing 2,484 drugs and 24,511 unique disease concepts [5-7]. TreatKB includes 9,216 drug-disease treatment pairs extracted from all FDA drug labels, 111,862 pairs extracted from the FDA Adverse Event Reporting System (FAERS), a database supporting the FDA’s post-marketing drug safety surveillance; 34,306 pairs extracted from 21 million published biomedical literature abstracts, and 69,724 pairs extracted from 172,888 clinical trials. The data is publicly available at http://nlp.case.edu/public/data/treatKB/.

Disease genetics data from the GWAS catalog Disease-gene association data from the Catalog of Published Genome-Wide Association Studies (the GWAS catalog) was used to construct disease networks and to identify PC-related diseases. The GWAS catalog is an exhaustive source containing the description of disease- and trait-associated single nucleotide polymorphisms (SNPs) from published GWAS data and currently contains 22,470 disease-gene pairs for 882 diseases and 8,689 genes [4]. Recent studies have shown that disease genomics data in the GWAS catalog is a rich resource for drug discovery [8-11].

ICD10 for disease classification We identified diseases that are genetically related to PC (PC-related diseases) and analyzed PC-related diseases using the 10th revision of the International Statistical Classification of Diseases and Related Health Problems (ICD10) [15]. The 22 highest-level disease classes (e.g., “Neoplasms” and “Diseases of the nervous system”) were used in categorizing PC-related diseases.

The Anatomical Therapeutic Chemical (ATC) classification system for drug classification The ATC system was used to classify and analyze repositioned drugs. The ATC classification system consists of 13 first-level codes, 94 second-level codes, 267 third-level codes, 882 fourth-level codes, and 4580 fifth-level codes, which are individual drugs [16].

Drug target genes and genetic pathways data Drugs’ gene targets were obtained from DrugBank [17], a comprehensive database containing information on drugs and drug targets. Gene-associated pathways were obtained from the Molecular Signatures Database (MSigDB), which contains 10,295 annotated genetic pathways [18].

Methods

The experiment framework of GenoPredict consists of the following steps: (1) we constructed a genetic disease network (GDN) and applied a network-based ranking algorithm to find diseases that are genetically related to PC; (2) we developed a drug prioritization algorithm to identify candidate drugs; GenoPredict was tested using 27 FDA-approved PC drugs and was compared to PREDICT in novel predictions; and (3) we performed drug classification and pathway enrichment analysis to analyze top ranked repositioning drug candidates.

Construct genetic disease network (GDN) and find PC-related diseases from GDN

Construct GDN We constructed GDN using disease-gene association data from the GWAS catalog. On GDN, two diseases were connected if their associated genes overlapped. The edge weights were determined by the cosine similarity coefficients of disease-associated genes [19]. GDN comprised of 882 disease nodes and 200,758 edges.

Apply network-based ranking algorithm to find diseases that share high genetic relevance with PC (PC-related diseases) We have recently applied a commonly used network-based ranking algorithm [20] for disease genetics dis-
covery [21-23], human gut microbial metabolite prediction for diseases [24-25], and drug repositioning [12-13]. In this study, we directly used the network-based ranking algorithm to identify PC-related diseases from GDN. The iterative ranking algorithm is defined as: 

$$p_{t+1} = (1 - r)Mp_t + rp_0$$

wherein $M$ is the column-normalized adjacency matrix of GDN, $\gamma$ is a preset probability of restarting from the initial seed node ($\gamma=0.1$ in this study), and $p^t$ is a vector in which the $i_{th}$ element holds the normalized ranking score of disease $i$ at $t_{th}$ iteration. The initial probability vector $p^0$ contains normalized probability values for input. In our study, $p^0$ contains PC, with a probability of 1.0. Diseases were ranked according to values in the steady-state probability vector, which was obtained by iterating the algorithm until the change between $p_{t+1}$ and $p_t$ was less than $10^{-6}$.

We expect that top-ranked PC-related diseases should be enriched for cancers since cancers are known to share genetics with each other. To evaluate the network-based disease ranking approach and to systematically study PC-related diseases, we classified PC-related diseases based on the ICD10 classification scheme[15], and examined distributions of 22 disease classes at ten ranking cutoffs (top 10%, 20%, . . . 100%).

**Drug prioritization algorithm**

**Drug repositioning algorithm** We developed an approach to systematically reposition drugs from PC-related diseases to treat PC. We prioritized drugs based on the number of PC-related diseases that they could treat as well as the ranking scores of these diseases. For example, if drug X treats 25 top-ranked PC-related diseases, it will rank higher than drug Y, which treats only a few lower-ranked diseases. The drug prioritization algorithm is defined as: 

$$R_{drug} = \sum_{i=1}^{n} R_{disease,i}$$

wherein $n$ is the number of PC-related diseases that a drug can treat and $R_{disease,i}$ is the disease ranking score (output from the network-based disease ranking algorithm).

**De-novo validation using FDA-approved PC drugs** We evaluated GenoPredict using 27 FDA-approved PC drugs. The drug prioritization algorithm used only PC-related diseases and their drug treatments. PC and its known drug treatments were not used, therefore the evaluation was a de-novo validation. We calculated recall, mean and median rankings of these 27 FDA-approved drugs among all drugs.

**Compare GenoPredict to PREDICT in novel predictions** We compared GenoPredict with PREDICT using two evaluation datasets constructed from TreatKB: (1) 172 PC drugs extracted from 172,888 clinical trials; and (2) 72 PC drugs extracted from 43,811 ongoing clinical trials. The 27 FDA-approved PC drugs were removed from the evaluation datasets. The output from GenoPredict is a ranked list of 2484 drugs with the 27 FDA-approved PC drugs removed. In PREDICT, 48 drugs were classified as positives for PC among a total of 593 drugs. These 48 drugs along with their corresponding probabilities (ranging from 0.543-0.994) are publicly available [14]. For the remaining 545 drugs that were predicted as negatives by PREDICT, we assigned to each drug a value that was randomly picked from 0.0 to 0.5 (the random process was repeated 10 times).

Precision-Recall (PR) curves were used to evaluate and compare GenoPredict to PREDICT. Studies have shown that in domains where the number of negatives greatly exceeds the number of positives such as in drug repositioning and many other biomedical classification problems, PR curves are better than the more commonly used receiver operating characteristic (ROC) curves in avoiding overly optimistic view of an algorithm’s performance [26]. Using each of the two evaluation datasets as gold standard, we calculated precisions at 11 different recall cutoffs (0.0, 0.1, 0.2, ... 1.0) for both GenoPredict and PREDICT and plotted the PR curves. Mean Average Precision (MAP), which approximates the area under the precision-recall curve [27], was used to quantitatively compare GenoPredict to PREDICT.

**Analyze top-ranked repositioned drug candidates**

**Analyze drug classes for top-ranked drug candidates** To understand the commonalities shared by top-ranked drug candidates, we examined class distributions of top 124 (top 5%) repositioned candidate drugs. Drugs were classified based on the ATC classification system [16]. For each ATC code, we assessed its probability of being associated with these 124 drugs (e.g., the code “immunosuppressants” is associated with 30 of the 124 drugs) as compared to its probability of being associated with the same number of randomly selected drugs (e.g., the same code is on average
Cancers and metabolic diseases are enriched among top-ranked PC-related diseases

It is known that cancers often share genetics. We analyzed the distributions of cancers among diseases at different ranking cutoffs to evaluate the network construction and disease-ranking algorithms. As shown in Fig.1, cancers are enriched among top-ranked diseases. For example, at the 100% cut-off (all 822 retrieved diseases), 7.24% of the diseases are “Neoplasms”. At top 10% cutoff (top 82 diseases), 16.28% are “Neoplasms,” representing a 124.8% increase as compared to that for all diseases. The enrichment of cancers among top-ranked diseases demonstrated the validity of both network construction and ranking algorithms. Interestingly, the disease class “Endocrine, nutritional and metabolic diseases” was also enriched among top-ranked diseases. Recent studies demonstrated that certain metabolic disorders - including obesity, high blood pressure and high cholesterol - can impact the development and progression of prostate cancer [28]. Other disease classes were not enriched among top-ranked diseases.
GenoPredict found 25 of the 27 FDA-approved PC drugs and ranked them highly

We validated GenoPredict in predicting known PC drugs using 27 FDA-approved PC drugs. GenoPredict achieved a recall of 0.925, an average ranking of 27.3%, and a median ranking of 15.6%. There is a big difference between the mean and median rankings, demonstrating a skewed ranking distribution of these FDA-approved drugs (Fig. 2). While 16 of the 25 PC drugs were ranked within top 20%, some drugs were ranked very low, including degarelix at top 70.49%, abarelix at 86.76%, triptorelin at 89.21%, and histrelin at 91.22%.

Compare GenoPredict to PREDICT in novel predictions

We plotted PR curves for both GenoPredict and PREDICT using 172 novel PC drugs extracted from 172,888 clinical trial reports as the evaluation dataset. As shown in Fig. 3, the PR curve for GenoPredict clearly dominates that for PREDICT. The mean average precision (MAP) is 0.447 for GenoPredict and 0.169 for PREDICT, representing a significant 164.5% improvement.

Fig. 4 shows the PR curves when 72 novel PC drugs extracted from ongoing clinical trials were used as the evaluation set. While the MAPs for both algorithms were lower than previous ones, the curve for GenoPredict dominates that for PREDICT. The MAP is 0.278 for GenoPredict and 0.084 for PREDICT, representing a significant 231.1% improvement.

Analyze top-ranked drug candidates: drug classification and genetic pathway analysis

We examined drug classes enriched among top-ranked drugs. The highly enriched drug classes may offer insight into the mechanisms of action underlying repositioned drug candidates. Fig. 5 shows the top 15 drug classes. A total of 18.8% of top 124 drug candidates belong to the class “other antineoplastic agents” as compared to 3.9% based on random expectation. The other top-ranked drug classes include “immunosuppressants” and “corticosteroids for systemic use, plain.” We performed the same enrichment analysis for top 248 (top 10%) and top 372 (top 15%) drug
Figure 3: The Precision-Recall curves for GenoPredict and PREDICT. The evaluation data consists of 172 PC drugs extracted from 172,888 clinical trial reports.

Figure 4: The Precision-Recall curves for GenoPredict and PREDICT. The evaluation data consists of 72 novel PC drugs extracted from 43,811 ongoing clinical trials.
This study focused on integrating disease genetics with drug treatment data for drug repositioning. To build a more resource of human disease genetics is computation-based candidate disease genetics discovery. We recently showed genetics from rare Mendelian disorders represents another valuable source of novel drug targets [31]. Another rich catalog, additional drug repositioning opportunities will arise through human genetic analysis. Additionally, disease TreatKB includes 24,511 diseases and their drug treatments. With new studies being continually added to the GWAS catalog, even though the limited number of diseases (882 diseases) in the GWAS catalog, even though TreatKB includes 24,511 diseases and their drug treatments. With new studies being continually added to the GWAS catalog, additional drug repositioning opportunities will arise through human genetic analysis. Additionally, disease genetics from rare Mendelian disorders represents another valuable source of novel drug targets [31]. Another rich resource of human disease genetics is computation-based candidate disease genetics discovery. We recently showed that computationally-predicted disease genetics can lead to novel drug discovery [21-22].

This study focused on integrating disease genetics with drug treatment data for drug repositioning. To build a more candidates and obtained similar results (data not shown).

We then examined genetic pathways that are predominantly targeted by top-ranked (top 124) drug candidates. As shown in Fig. 6, the pathway “pathways in cancer” was targeted by 19.05% of the top 124 candidate drugs as compared to 7.92% based on random expectation. The pathway “prostate cancer” was also ranked at top, demonstrating the validity of our pathway analysis. In addition, four of the top fifteen pathways are related to immune systems. Currently, Immunotherapy holds great promise for cancer treatments. For example, the drug ipilimumab targets a checkpoint molecule called CTLA-4 on certain immune cells. This drug is now being tested in men with advanced PC, with early results showing treatment benefits [29-30]. Consistent with these published clinical studies, GenoPredict ranked ipilimumab at top 15% among 2484 drugs.

Discussion

GenoPredict exploited genetic and drug treatment connections among tens of thousands of diseases. We applied GenoPredict to identify repositioning drug candidates for PC and performed in-depth analysis of top repositioned drug candidates. GenoPredict found 25 of 27 FDA-approved PC drugs and ranked them highly. GenoPredict performed significantly better than PREDICT, a comprehensive drug repositioning system, in novel predictions for PC. However, the performance of GenoPredict is not optimal as seen in its PR curves for drug repositioning for PC. We are further improving GenoPredict by incorporating more comprehensive disease genetics data and will evaluate it in broad types of diseases (e.g., all diseases in the GWAS catalog).

Our current study is restricted by the limited number of diseases (882 diseases) in the GWAS catalog, even though TreatKB includes 24,511 diseases and their drug treatments. With new studies being continually added to the GWAS catalog, additional drug repositioning opportunities will arise through human genetic analysis. Additionally, disease genetics from rare Mendelian disorders represents another valuable source of novel drug targets [31]. Another rich resource of human disease genetics is computation-based candidate disease genetics discovery. We recently showed that computationally-predicted disease genetics can lead to novel drug discovery [21-22].
Top genetic pathways targeted by repositioned candidate drugs

Figure 6: Top fifteen pathways targeted by top 124 (top 5%) repositioned candidate drugs.

A comprehensive prediction system, additional data resources such as disease phenotypes, gene expression, drug side effects, and chemical structures can be incorporated into GenoPredict to further improve the performance.

Author’s contributions

Xu and Wang have jointly conceived the idea, designed and implemented the algorithms and prepared the manuscript. All authors read and approved the final manuscript.

Acknowledgements

RX is funded by Case Western Reserve University/Cleveland Clinic CTSA Grant (UL1 RR024989), the NIH Director’s New Innovator Award under the Eunice Kennedy Shriver National Institute Of Child Health & Human Development of the National Institutes of Health (DP2HD084068, Xu), American Cancer Society Research Scholar Grant (RSG-16-049-01 - MPC, Xu), the Landon Foundation-AACR INNOVATOR Award for Cancer Prevention Research (15-20-27-XU), Mary Kay Foundation Grant (057-15, Xu), and Pfizer 2015 ASPIRE Rheumatology and Dermatology Research Award (WI206753, Xu). QuanQiu’s effort on this project is funded by DP2HD084068.

References

5. Xu, R., & Wang, Q. (2013). Large-scale extraction of accurate drug-disease treatment pairs from biomedical

Combining mechanism-based prediction with patient-based profiling for psoriasis metabolomics biomarker discovery

QuanQiu Wang¹, MS, Thomas S. McCormick³, PhD, Nicole L. Ward², PhD, Kevin D. Cooper², MD, Ruzica Conic, MD, Rong Xu³*, PhD

¹ ThinTek, LLC, Palo Alto, CA, USA
² Department of Dermatology, School of Medicine, Case Western Reserve University, Cleveland, OH, USA
³ Department of Epidemiology and Biostatistics, Institute of Computational Biology, School of Medicine, Case Western Reserve University, Cleveland, OH, USA

Abstract

Psoriasis is a chronic, debilitating skin condition that affects approximately 125 million individuals worldwide. The cause of psoriasis appears multifactorial, and no unified mitigating signal or single antigenic target has been identified to date. Metabolomic studies hold great potential for explaining disease mechanism, facilitating early diagnosis, and identifying potential therapeutic areas. Here, we present an integrated disease metabolomic biomarker discovery strategy that combines mechanism-based biomarker discovery with clinical sample-based metabolomic profiling. We applied this strategy in identifying and understanding metabolite biomarkers for psoriasis. The key innovation of our strategy is a novel mechanism-based metabolite prediction system, mmPredict, which assimilates vast amounts of existing knowledge of diseases and metabolites. mmPredict first constructed a psoriasis-specific mouse mutational phenotype profile. It then constructed phenotype profiles for a total of 259,170 chemicals/metabolites using known chemical genetics and human metabolomic data. Metabolites were then prioritized based on the phenotypic similarities between disease- and metabolites. We evaluated mmPredict using 150 metabolites identified using our in-house metabolome profiling study of psoriasis patient samples. mmPredict found 96 of the 150 metabolites and ranked them highly (recall: 0.64, mean ranking: 8.73%, median ranking: 2.33%, p-value: 4.75E-44). These results show that mmPredict is consistent with, as well as a complement to, traditional human metabolomic profiling studies. We then developed a strategy to combine outputs from both systems and found that the oxidative product of linoleic acid, 13(S)-hydroxy-9Z,11E-octadecadienoic acid (13-HODE), ranked highly by both mmPredict and our in-house experiments. Our integrated analysis indicates that 13-HODE may be a mechanistic link between psoriasis and cardiovascular comorbidities associated with psoriasis. In summary, we developed an integrated metabolomic prediction system that combines both human metabolomic studies and mechanism-based prediction and demonstrated its application in the skin disease psoriasis. Our system is highly general and can be applied to other diseases when patient-based metabolomic profiling data becomes more increasingly available.

Data is publicly available at: http://nlp.case.edu/public/data/mmPredict_PSO

1. Introduction

The prevalence of skin disease exceeds those of obesity, hypertension or cancer. One in three Americans suffers from a skin disease, and ~2-3% of these individuals have psoriasis. Costs of treating psoriasis patients are believed to exceed $1B, including over $350M in prescriptions alone [1-4]. The pathophysiology of psoriasis is complex, and no unified mitigating signal or single antigenic target has been identified.

Human metabolomics is the comprehensive characterization of all metabolites found in the human body. Human metabolomics has great potential for explaining disease mechanism, facilitating early diagnosis, and identifying potential therapeutic areas. Common technologies for metabolomics include mass spectrometry (MS) and nuclear magnetic resonance spectroscopy (NMR) [5-6]. Human metabolites are highly heterogeneous. The Human Metabolome Database (HMDB), a comprehensive database of small molecule metabolites found in the human body, contains 42,032 metabolites of lipids, small peptides, amino acids, organic acids, vitamins, carbohydrates, nucleic acids, as well as metabolites derived from drugs, environmental contaminants, food additives, toxins, cosmetics, and other xenobiotics such as microbial or fungal symbionts [7]. Profiling the human metabolome is difficult owing to the challenges of reproducibility and knowledge generalization, as the human metabolome is affected by not only intrinsic but also many extrinsic factors such as sample collection, storage, processing and data analysis [8]. The NIH Common Fund Metabolomics Program was launched in 2012 with a primary goal to increase the national capacity to conduct metabolomics research [9]. Currently, the NIH Metabolomics Workbench Metabolite Database
We evaluated mmPredict using 150 metabolites identified by comparing involved tissue from psoriasis patients to components and five data resources. mmPredict: a combined genome- and phenome-wide metabolite prediction system, including three

Fig 1: (MWMD) contains metabolomics studies for a very limited number of diseases [10]. While current metabolomics profiling strategies can identify clinically significant metabolite biomarkers, it is limited in understanding the underlying mechanistic links between identified metabolites and diseases.

Here, we present a novel mechanism-based metabolomic biomarker discovery system, mmPredict, to complement current patient-based metabolomic profiling studies. We developed an approach to combine outputs from these two complementary strategies and demonstrate its utility in the clinical skin disease psoriasis. The output of this integrated system is a ranked list of metabolomic biomarkers that incorporate both clinical significance and interpretable molecular mechanisms. mmPredict integrates large amounts of data from human disease genetics, chemical genetics, mouse mutational phenomics, human metabolomics, and genetic pathways in order to predict disease metabolomic biomarkers. The underlying rationale for mmPredict is that, if the changes in metabolite-associated genes cause many phenotypes (reflected in mouse models) that are also involved in any given disease (psoriasis in this study), then the metabolite is likely to be involved mechanistically in the disease. In order to test the validity of mmPredict to traditional metabolomic profiling, we compared the output of mmPredict to our in-house metabolomic profile. Our in-house psoriasis metabolome was identified using a traditional metabolomics comparison study of skin samples obtained from the involved plaques of psoriasis patients (n=12) compared to skin samples obtained from age-matched healthy control subjects (n=9) and identified 150 significantly altered metabolites. mmPredict highly ranked psoriasis metabolites identified from our in-house metabolomic profiling study, indicating that these two strategies are consistent with, as well as complementary to, each other. We then developed a strategy to combine outputs from these two complementary strategies and identified psoriasis-associated metabolite biomarkers likely to have both mechanistic and clinical significance.

mmPredict performs both genome- and phenome-wide matching between metabolites and diseases. We compared mmPredict to a genome-wide- strategy that we recently developed to identify human gut microbial metabolite biomarkers for colorectal cancer (CRC) [11] as well as Alzheimer’s disease [12]. We demonstrated in this study that mmPredict is more effective in identifying clinically relevant metabolite biomarker for psoriasis than the genome-wide approach.

2. Data and methods

2.0 The overview of mmPredict

mmPredict matches psoriasis metabolites based on both genetic and phenotypic relevance. mmPredict consists of the following components: (1) mmPredict constructs mouse mutational phenotype profiles for a given disease (psoriasis in this study) using publicly available disease genetics and genomics databases; (2) it constructs mouse mutational phenotype profiles for a total of 259,170 chemicals/metabolites; (3) mmPredict prioritizes metabolites for a given disease based on the phenotype profile similarities between the disease and metabolites (Fig. 1).

![Fig 1: mmPredict: a combined genome- and phenome-wide metabolite prediction system, including three components and five data resources.](image)

We evaluated mmPredict using 150 metabolites identified by comparing involved tissue from psoriasis patients to healthy control skin in our in-house metabolomic study. We developed an approach to combine predictions from mmPredict and from our in-house metabolomic study to identify metabolites likely to have both clinical and mechanistic significance.
For comparison, we implemented a genome-wide metabolite prediction system mmPredict_Gen. The underlying assumption is that if a metabolite is associated with many psoriasis-associated genes, then the metabolite may be associated with the mechanism(s) causing psoriasis. We have previously applied this strategy to identify human gut microbial metabolites associated with colorectal cancer [11] as well as Alzheimer’s disease [12]. The difference between mmPredict and mmPredict_Gen is that mmPredict utilizes both genome- and phenome-wide information, while mmPredict_Gen prioritizes metabolites for diseases based on genetic profile similarities (Fig 2). We demonstrated in our study that mmPredict performed consistently better than mmPredict_Gen.

Fig 2: mmPredict_Gen: a genome-wide metabolite prediction system (five data resources are shown).

2.1 Data
We used the disease phenotype knowledge bases that we recently constructed, as well as publicly available data from the human metabolome, disease genetics, chemical genetics, functional protein interactions and signaling pathway databases to identify metabolites associated with psoriasis and its’ comorbidities.

2.1.1 Disease genetics and genomics data
We used two complementary databases to obtain genes associated with psoriasis. We identified 49 genes associated with common complex forms of psoriasis from the Catalog of Published Genome-Wide Association Studies (GWAS catalog), an exhaustive source containing descriptions of disease/trait-associated single nucleotide polymorphisms (SNPs) from published GWAS studies [13]. We identified 35 genes associated with rare Mendelian forms of psoriasis from the Online Mendelian Inheritance in Man (OMIM) database, the most comprehensive source of disease genetics for Mendelian disorders [14]. We used these two complementary resources of disease genetics to demonstrate the robustness of our algorithms and findings.

2.1.1 The Human Metabolome Database (HMDB)
HMDB is intended for applications in metabolomics, biomarker discovery and other applications and contains detailed information regarding small molecule metabolites found in the human body [7]. Currently, HMDB is the most comprehensive collection of metabolite data, containing 42,032 metabolites. We used HMDB to obtain a list of metabolites found in the human body.

2.1.2 Chemical genetics data
We used the STITCH (Search Tool for Interactions of Chemicals) database [15] to obtain chemical/metabolite-gene associations. We used chemical-gene associations found in the human body (1,466,636 chemical-gene pairs for 259,171 chemicals and 15,620 genes) and metabolites from HMDB to link human metabolites to human genes.

2.1.4 Genome-wide mutational phenotypes in experimental mouse models
We used gene-phenotype associations from the Mous e Genome Database (MGD) [16] to assess the phenotypic effects of metabolites. The Phenotypes/Alleles project in the MGD provides access to spontaneous, induced, and genetically engineered mutations and their specific phenotypic outcomes. Currently, MGD contains 278,553 gene-phenotype associations, representing 41,905 mutant alleles and 10,744 phenotypes. For example, the mutation of 13-HODE-associated gene myeloperoxidase (MPO) in mouse models is associated with 12 phenotypes, including ‘decreased inflammatory response,’ ‘atherosclerotic lesions’ and ‘increased monocyte cell number.’ The mutation of psoriasis-associated gene TRAF3 interacting protein 2 (TRAF3IP2) is associated with 80 phenotypes including ‘decreased inflammatory response,’ ‘increased pruritus,’ and ‘increased monocyte cell number.’ We have recently shown that systematic approaches to interrogate human genes to their mouse mutational phenotypes in MGD have
great potential in understanding both disease mechanisms and drug effects [17-19]. In this study, we leveraged the large number of gene-phenotype associations from MGD and performed genome- and phenome-wide analysis to infer phenotypic effects of metabolites on disease.

2.2 Methods

2.2.1 Construct mouse mutation phenotype profiles for input disease

We identified 35 psoriasis-associated genes from the OMIM database and 49 psoriasis-associated genes from the GWAS Catalog. We mapped these genes to their corresponding mouse gene homologs (e.g., TRAF3IP2 => Traf3ip2) using human-mouse homolog mapping data from the Mouse Genome Database (MGD) [16]. The mapped mouse genes were then linked to their corresponding mutational phenotypes (e.g., knockout) in mouse models (e.g., Traf3ip2 => 'increased pruritus') using gene-phenotype association data from MGD. For each mapped phenotype, we assessed its probability of being associated with the given set of genes (e.g., the phenotype “increased pruritus” is associated with 10 of the 49 psoriasis genes) as compared to its probability associated with the same number of randomly selected genes (e.g., the phenotype “increased pruritus” is on average associated with 0.5 of 49 randomly selected genes). The random process is repeated 1000 times and a t-test was used to assess the statistical significance.

2.2.2 Construct mutational phenotype profiles for chemicals

We built phenotype profiles for 259,170 chemicals/metabolites from the STITCH database in the same manner as described above (e.g., 13-HODE => MPO =>Mpo => 'increased monocyte cell number').

2.2.3 Prioritize metabolites for psoriasis and test the performance using in-house metabolomics study

Metabolite prioritization

We calculated the phenotypic similarity of psoriasis and metabolites. We compared four different approaches in calculating the similarities, including cosine similarity, overlap, Jaccard similarity [20], and ontology-based semantic similarity [17-19, 21]. We showed that ranking based on Jaccard similarity performed consistently better than the other three similarity measures used in this study. Therefore, throughout this study, the metabolite ranking values are based on the Jaccard similarity.

Evaluation of recall and ranking

We evaluated mmPredict in identifying and prioritizing metabolite biomarkers for psoriasis using our in-house metabolome study. We identified 150 significantly altered metabolites by comparing global metabolic profiles in skin samples from normal control subjects (n=9) and age matched psoriasis patient involved psoriasis plaques (n=12). We evaluated the algorithm for both identifying and ranking known metabolites using recall, mean ranking, and median rankings. Random expectation that the metabolites would have an average ranking of 50% was used to evaluate the significance of these rankings.

Evaluation of prioritization (precision-recall curve)

We evaluate mmPredict using an 11-point interpolated average precision measure, which is commonly used to evaluate retrieved ranked lists for search engines [22]. For each ranked list, the interpolated precision was measured at the 11 recall levels of 0.0, 0.1, 0.2, ..., 1.0. At each recall level, we calculated the arithmetic mean of the interpolated precision. A composite precision-recall curve showing 11 points was then graphed.

2.2.4 Develop a strategy to combine outputs from mmPredict and patient-based metabolomics profiling

We developed an approach to combine rankings from mmPredict and fold changes from the clinical metabolomic study. The output of mmPredict is a list of metabolites ranked by their percent ranking (“ranking_m”). For example, 13-HODE ranked at the top 0.23% among 259,170 chemicals. We then ranked the 150 in-house metabolites by their fold changes and calculated their percent ranking (“ranking_c”) among all metabolites. For example, 13-HODE was ranked at the top 0.67% (top one) among the 150 in-house generated metabolites. The combined ranking is a balance measure of rankings from both studies and is defined as: ranking_combined = 2*(ranking_m* ranking_c)/(ranking_m + ranking_c). We studied the top one ranked metabolite by performing both genetic pathway enrichment and mutational phenotype enrichment analyses.

3. Results

3.1. Understanding psoriasis genetics, genomics, and phenomics

We used mmPredict to perform both genome- and phenome-wide analyses to prioritize metabolites for psoriasis. To understand both genetic and phenotypic implications of psoriasis-associations genes, we began our study using Ingenuity Pathway Analysis (IPA) [23] to examine the top canonical pathways generated by searching for either rare
Mendelian psoriasis genes or common complex psoriasis genes. As shown in Table 1, the top diseases and biological functions identified dermatological diseases and conditions, with immunological and inflammatory response also being identified. Common cellular response pathways and cells (e.g., dendritic cells and janus activated kinase (JAK) signaling) as well as other skin and immunologically prominent diseases (e.g., graft versus host disease (GVHD), Rheumatoid Arthritis (RA) and Diabetes Mellitus (DM)) were also identified.

<table>
<thead>
<tr>
<th>Top Canonical Pathways associated with OMIM PSO genes</th>
<th>Top Diseases and Bio Functions associated with OMIM PSO genes</th>
<th>Top Canonical Pathways associated with GWAS PSO genes</th>
<th>Top Diseases and Bio Functions associated with GWAS PSO genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dendritic Cell Maturation</td>
<td>Dermatological Diseases and Conditions</td>
<td>Type I Diabetes Mellitus Signaling</td>
<td>Dermatological Diseases and Conditions</td>
</tr>
<tr>
<td>Graft-versus-Host Disease Signaling</td>
<td>Inflammatory Response</td>
<td>Role of JAK1, JAK2 and TYK2 in Interferon Signaling</td>
<td>Immunological Disease</td>
</tr>
<tr>
<td>Role of Macrophages, Fibroblasts and Endothelial Cells in Rheumatoid Arthritis</td>
<td>Infectious Diseases</td>
<td>Interferon Signaling</td>
<td>Gastrointestinal Disease</td>
</tr>
<tr>
<td>Altered T Cell and B Cell Signaling in Rheumatoid Arthritis</td>
<td>Neurological Disease</td>
<td>Dendritic Cell Maturation</td>
<td>Inflammatory Disease</td>
</tr>
<tr>
<td>Communication between Innate and Adaptive Immune Cells</td>
<td>Psychological Disorders</td>
<td>T Helper Cell Differentiation</td>
<td>Inflammatory Response</td>
</tr>
</tbody>
</table>

Table 1. Top 5 genetic pathways, diseases, and biological functions enriched for psoriasis-associated genes. Skin-specific diseases and pathways are shown in pink and Immune-related diseases and pathways are shown in green.

Table 2 outlines the top 10 phenotypes associated with psoriasis genes identified through either the OMIM or GWAS databases. As shown in the table, the top phenotypes are enriched for immune-response and skin-related phenotypes, confirming the observational and experimental data supporting psoriasis as a predominantly immune-mediated autoimmune disorder. Both genetic pathway enrichment and mutational phenotype enrichment analyses provided the rationale underlying mmPredict’s combined genome- and phenome-wide analysis. As shown in the following sections, genome-wide prioritization is not as effective as the combined approach.

<table>
<thead>
<tr>
<th>Top phenotypes associated with OMIM PSO genes</th>
<th>Top phenotypes associated with GWAS PSO genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abnormal immune system morphology</td>
<td>Premature death</td>
</tr>
<tr>
<td>Increased double-negative T cell number</td>
<td>Complete embryonic lethality during organogenesis</td>
</tr>
<tr>
<td>Decreased IgG level</td>
<td>Thick epidermis</td>
</tr>
<tr>
<td>Increased leukocyte cell number</td>
<td>Abnormal CD4-positive, alpha-beta T cell physiology</td>
</tr>
<tr>
<td>Increased neutrophil cell number</td>
<td>Increased inflammatory response</td>
</tr>
<tr>
<td>Abnormal cytokine secretion</td>
<td>Decreased IgG level</td>
</tr>
<tr>
<td>Increased pruritus</td>
<td>Increased IgM level</td>
</tr>
<tr>
<td>Dermatitis</td>
<td>Decreased interferon-gamma secretion</td>
</tr>
<tr>
<td>Altered susceptibility to infection</td>
<td>Abnormal cytokine secretion</td>
</tr>
<tr>
<td>Decreased transitional stage B cell number</td>
<td>Abnormal macrophage physiology</td>
</tr>
</tbody>
</table>

Table 2. Top 10 phenotypes enriched for psoriasis genes from the OMIM and the GWAS catalog. Skin-specific diseases and pathways are shown in pink and Immune-related diseases and pathways are shown in green.

3.2. mmPredict ranked clinically derived psoriasis-associated metabolites highly

The output of mmPredict is a ranked list of 259,170 chemicals. We also filtered the list of chemicals from HMDB by their origins. In HMDB, metabolites are classified based on their origins, such as “endogenous”, “food”, “microbial”, “drug”, “plant”, and “toxin/pollutant”. Only 4,057 of the 259,170 chemicals appeared in HMDB, including 2,067 endogenous metabolites. We evaluated the performance of mmPredict using the 150 clinically derived metabolites. As shown in Table 3, mmPredict found 96 of these 150 metabolites (recall: 0.64) and consistently ranked them highly for both the Mendelian form of psoriasis (mean ranking based on Jaccard similarity: 8.73%, median ranking based on Jaccard similarity: 2.33%, p-value: 4.75E-44) and common complex form of psoriasis (mean ranking based on Jaccard similarity: 9.08%, median ranking based on Jaccard similarity: 2.39%, p-
value: 1.35E-42). After filtering the chemicals by metabolites from HMDB, the rankings are still significantly higher than random expectation. However, the recall are lower, indicating that the list of chemicals from the STITCH database represents a more complete list of metabolites found in the human body than the list of metabolites from HMDB.

| Metabolite source | Psoriasis genetics | Recall | Mean Ranking | Median Ranking | P-value  
|-------------------|--------------------|--------|--------------|---------------|----------
| STITCH (259,170 chemicals) | OMIM | 0.64 | 8.73% | 2.33% | 4.75E-44 |
|                   | GWAS | 0.64 | 9.08% | 2.39% | 1.35E-42 |
| HMDB (4057 metabolites) | OMIM | 0.41 | 22.36% | 14.03% | 8.54E-13 |
|                   | GWAS | 0.41 | 23.25% | 14.56% | 8.54E-13 |
| HMDB (2067 endogeneous metabolites) | OMIM | 0.34 | 19.27% | 11.67% | 2.64E-14 |
|                   | GWAS | 0.34 | 20.42% | 12.63% | 2.16E-13 |

Table 3. Evaluation of mmPredict with 150 clinically derived psoriasis metabolites. The ranking was based on Jaccard similarity.

For comparison, we show in Table 4 that genome-wide prioritization alone is less effective than the combined genome- and phenome-wide implementation of mmPredict. mmPredict has consistently better mean and median rankings than mmPredict_Gene across three metabolites data resources.

| Metabolite source | Psoriasis genetics | Recall | Mean Ranking | Median Ranking | P-value  
|-------------------|--------------------|--------|--------------|---------------|----------
| STITCH (259,170 chemicals) | OMIM | 0.64 | 46.55% | 53.55% | 0.309 |
|                   | GWAS | 0.64 | 42.50% | 37.92% | 0.039 |
| HMDB (4057 metabolites) | OMIM | 0.40 | 44.24% | 45.42% | 0.184 |
|                   | GWAS | 0.40 | 39.21% | 40.46% | 0.008 |
| HMDB (2067 endogeneous metabolites) | OMIM | 0.34 | 39.84% | 36.64% | 0.024 |
|                   | GWAS | 0.34 | 35.09% | 30.93% | 0.001 |

Table 4. Evaluation of mmPredict_Gen with 150 psoriasis metabolites. The ranking was based on Jaccard similarity.

We further show in the plotted 11-point interpolated precision-recall curve that mmPredict is effective in enriching clinically derived psoriatic metabolites at 11 recall cutoffs (Fig. 3). At a recall level of 0.1, the mean precision of the metabolites is 0.156, which represents a 524% enrichment as compared to the precision of 0.025 for all metabolites (recall = 1.0). These results indicate that mmPredict enriched clinically identified metabolites among the top and its predictions are consistent with patient-based metabolomic profiling. The precision was calculated using the clinically derived metabolites and does not represent the true precision measure of mmPredict. The low values of the precision (e.g., 0.156 at recall of 0.1) indicate that many more metabolites are to be discovered and that mmPredict can complement patient-based profiling in identifying metabolites not captured in our in-house metabolomic study.
3.2. 13-HODE represents a metabolite biomarker mechanistically linking psoriasis to its cardiovascular comorbidities

mmPredict represents a mechanism-based approach to identify metabolites for psoriasis. Our in-house metabolomics study represents a standard approach to identify altered metabolites in psoriasis patients. Our combined scores from these two complementary studies identified 13-HODE as the highest ranked metabolite, indicating that this metabolite may be involved in psoriasis at both mechanistic and clinical levels. To understand how 13-HODE is mechanistically linked to psoriasis, we identified a total of 78 13-HODE-associated genes from the STITCH database. We performed pathway enrichment analysis using Ingenuity Pathway Analysis (IPA) to examine the top canonical pathways associated with 13-HODE. The top five canonical pathways identified by IPA are Atherosclerosis Signaling, Eicosanoid Signaling, Glutathione Redox Reactions I, Bupropion Degradation, and Acetone degradation I. The top five diseases and biofunctions are Organismal Injury and Abnormalities, Cardiovascular Disease, Gastrointestinal Disease, Inflammatory Disease, and Inflammatory Response. The pathway enrichment analysis shows that 13-HODE is involved in inflammatory response and cardiovascular diseases, common linkages associated with human psoriasis.

We then investigated the phenotypic effects of 13-HODE-associated genes by interrogating the associations of mouse phenotypes with mutations of HODE-associated genes. We identified a total of 1,214 phenotypes associated with mutations of the 78 13-HODE-associated genes. As shown in Table 5, the top ranked phenotypes associated with mutations of 13-HODE associated genes are clearly related to cardiovascular disease, immune response and abnormal skin morphology and function. These results indicate that 13-HODE may be mechanistically involved in both psoriasis and its' cardiovascular comorbidities.

<table>
<thead>
<tr>
<th>Phenotype Category</th>
<th>Phenotype</th>
<th>Rank (top%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CVD phenotypes</td>
<td>Congestive heart failure</td>
<td>0.41%</td>
</tr>
<tr>
<td></td>
<td>Cardiac fibrosis</td>
<td>0.82%</td>
</tr>
<tr>
<td></td>
<td>Cardiac hypertrophy</td>
<td>1.57%</td>
</tr>
<tr>
<td></td>
<td>Atherosclerosis</td>
<td>2.22%</td>
</tr>
<tr>
<td>Immune phenotypes</td>
<td>Abnormal neutrophil physiology</td>
<td>1.15%</td>
</tr>
<tr>
<td></td>
<td>Abnormal macrophage physiology</td>
<td>1.81%</td>
</tr>
<tr>
<td></td>
<td>Abnormal immune system physiology</td>
<td>1.98%</td>
</tr>
<tr>
<td></td>
<td>Increased monocyte cell number</td>
<td>3.13%</td>
</tr>
<tr>
<td>Skin phenotypes</td>
<td>Abnormal cutaneous collagen fibril morphology</td>
<td>3.79%</td>
</tr>
<tr>
<td></td>
<td>Abnormal epidermis stratum corneum morphology</td>
<td>5.60%</td>
</tr>
<tr>
<td></td>
<td>Impaired skin barrier function</td>
<td>6.43%</td>
</tr>
<tr>
<td></td>
<td>Decreased skin tensile strength</td>
<td>7.50%</td>
</tr>
</tbody>
</table>

Table 5. Top-ranked phenotypes related to CVD, immune system dysfunction, and abnormal skin morphology and function, and their rankings among all 1,214 HODE mutation-associated phenotypes.
We also categorized the 13-HODE-associated phenotypes (1,214 phenotypes were classified into 54 classes) and analyzed which phenotype categories are highly associated with HODE (the top 11 HODE-associated phenotype classes are shown in Fig. 4). Classes we found relevant to psoriasis and comorbidities include: 1) abnormal hematopoietic system (9.2% of all phenotypes), 2) abnormal skin morphology (5.0% of all phenotypes), 3) abnormal immune system (5.0% of all phenotypes), 4) abnormal cardiovascular system morphology (5.0%), and 5) abnormal cardiovascular system physiology (4.1%). Another skin-related class, abnormal skin adnexa morphology also ranked highly (3.3%). These results suggest that HODE may be a biomarker linking remote inflammation observed in the skin with cardiovascular diseases, immune system activation, and psoriasis.

Fig. 4: Top-ranked phenotype classes related to CVD (green), immune system dysfunction (blue), and abnormal skin morphology and function (azure).

Discussion and conclusions
We present a novel mechanism-based metabolomic biomarker discovery system, mmPredict, to complement current patient-based metabolomics profiling studies. We developed an approach to combine outputs from these two complementary strategies and demonstrated its utility in the human clinical disease psoriasis. The identification of metabolomic biomarkers and the understanding of their roles in psoriasis and psoriasis-related co-morbidities may provide insight into the basic mechanisms of psoriasis pathogenesis and enable new possibilities for psoriasis diagnosis, prevention, and treatment.

Both mmPredict and our strategy of combining these two complementary strategies for use in studying human metabolomics are general. We demonstrated its utility in psoriasis since we have performed in-house human metabolomic profiling on psoriasis patient and healthy control skin samples. We expect that mmPredict would be equally effective in identifying metabolite biomarkers for other diseases, albeit alternative sources of protein metabolites, such as serum or plasma may be necessary. However, to fully evaluate mmPredict and our integrated strategy in other diseases will require the gold standard of known disease-associated metabolite biomarkers, which are currently lacking due to the paucity of published disease specific metabolomic data. With the vast amounts of knowledge built into mmPredict, the only input needed for the system is a list of genes, therefore we expect that mmPredict can be applied to identify metabolite signatures unique for disease subtypes, disease progression, as well as treatment response given that the involved genes are available.

Computational algorithm design critically depends on the data incorporated into the system. In this study, we integrated disease genetics data with human metabolomics, chemical genetics, genetic pathways, and mouse mutational phenotypes. Our study shows that combined genome-phenome approach performed significantly better than the genome approach alone, indicating interrogating human genes to their functional effects on disease-specific phenotypes has major contribution in the improved performance. Currently, we are further improving mmPredict by incorporating other types of data, including higher-level disease and drug phenotypic data observed in humans.

Author’s contributions
QW and RX have jointly conceived, designed and implemented the algorithms, performed data analysis and algorithm evaluation, and prepared the manuscript. The metabolomics study was performed by TSM, KDC and NW. All authors read, edited and approved the final manuscript.

Acknowledgements
RX is funded by Case Western Reserve University/Cleveland Clinic CTSA Grant (UL1 RR024989), the NIH Director's New Innovator Award under the Eunice Kennedy Shriver National Institute Of Child Health & Human Development of the National Institutes of Health (DP2HD084068, Xu), American Cancer Society Research Scholar Grant (RSG-16-049-01 - MPC, Xu), the Landon Foundation-AACR INNOVATOR Award for Cancer Prevention Research (15-20-27-XU), Mary Kay Foundation Grant (057-15, Xu), and Pfizer 2015 ASPIRE Rheumatology and Dermatology Research Award (WI206753, Xu). QW’s effort on this project is partially supported by DP2HD084068. TSM was supported by NIH P30 AR039750-24, R01DE018276-05, R01 AR063437-03, R01 AR062546-02; KDC was supported by P30, AR039750; NW was supported by NIH R21 AR063852, R01 AR062546, R01 AR063437, and P30 AR39750; RC was supported by T32 AR007569-20.

References


SCOTCH: Secure Counting Of encrypTed genomiC data  
using a Hybrid approach  
Chenghong Wang\textsuperscript{1,2*} and Yichen Jiang\textsuperscript{1,2*}, Noman Mohammed\textsuperscript{3}, Feng Chen\textsuperscript{1}, Xiaoqian Jiang\textsuperscript{1}, Md Momin Al Aziz\textsuperscript{2}, Md Nazmus Sada\textsuperscript{1}, Shuang Wang\textsuperscript{1}  
\textsuperscript{1}Dept. of Biomedical Informatics, University of California San Diego, La Jolla, CA, USA  
\textsuperscript{2}Dept. of Computer Science, Syracuse University, Syracuse, NY, USA  
\textsuperscript{3}Dept. of Computer Science, University of Manitoba, Winnipeg, MB R3T 2N2, Canada  
* Both authors contribute equally to this paper.

Abstract  
As genomic data are usually at large scale and highly sensitive, it is essential to enable both efficient and secure analysis, by which the data owner can securely delegate both computation and storage on untrusted public cloud. Counting query of genotypes is a basic function for many downstream applications in biomedical research (e.g., computing allele frequency, calculating chi-squared statistics, etc.). Previous solutions show promise on secure counting of outsourced data but the efficiency is still a big limitation for real world applications. In this paper, we propose a novel hybrid solution to combine a rigorous theoretical model (homomorphic encryption) and the latest hardware-based infrastructure (i.e., Software Guard Extensions) to speed up the computation while preserving the privacy of both data owners and data users. Our results demonstrated efficiency by using the real data from the personal genome project.

Introduction  
Due to the advancement of genome sequencing technology, the cost of generating genomic data has been reduced significantly. The increasing collection of genomic data might contribute to the research in various ways, e.g., discovering innovative treatments, improving healthcare, and facilitating resources for biomedical research\cite{1}. For example, genome-wide association studies (GWASs)\cite{2} aims to identify the association relationship between single-nucleotide variants or polymorphisms (SNPs) with some specific disease/phenotype. On the other hand, the scale and size of genomic data that are being collected also raise challenges in storing and processing them\cite{3}. The cloud computing\cite{4} provides an elastic solution with on-demand computation ability and flexible storage. But the sensitivity of genomic data make the privacy a major concern to outsource genome data on a remote cloud service\cite{5}.

Existing studies have shown that it is possible to recover sensitive individual information, such as personal identity\cite{6,7}, disease condition\cite{8-10}, facial appearance\cite{11} through human genome data. As we know, genomic data are also shared among blood relatives, there will be extended impacts on the whole family\cite{12,13}, if the genomic data of a single family member are disclosed inappropriately. According to Lin and Altman’s study\cite{14}, as few as 75 independent SNPs would be sufficient to re-identify an individual. Sweeney\textit{et al.}\cite{7} published a study to show that the demographic information of a large percent of participants in the Personal Genome Project (PGP) can be linked to publicly available records. Another study by Gymrek showed that even the surname of participants in a “de-identified” database (1000 genome project) could be inferred using a genealogy database and some short tandem repeats on Y-chromosome\cite{8}. Even aggregated statistics\cite{9,15,16} like allele frequency and test statistics can leak sensitive personal information. Homer’s attack\cite{9} and the log likelihood ratio attack\cite{15} are two well-known attack models, which shows statistical power to infer the membership of individuals in a case group (i.e., with certain disease). Most recently, Shringarpure and Bustamante\cite{17} showed the vulnerability of the genomic data sharing Beacon project\cite{17}, which is designed to share only the existence of variants across multiple databases. More specifically, through accumulated queries of a Beacon database, Shringarpure et al.\cite{17} designed a method to learn if a vulnerable patient may be included in the database. Given the knowledge of rare allele, Raisaro\textit{et al.}\cite{18} further improve the attack power for Beacon database, where the enhanced attack requires much less number of queries than that of the Shringarpure’s model.

Previous studies have demonstrated many risk factors in directly sharing human genomic data. To mitigate the privacy/security dilemma when outsourcing or disseminating genomic data for research studies, several protection technologies been proposed\cite{19}, such as differential privacy (DP)\cite{20,21}, Homomorphic encryption (HME)\cite{22} and Secure multiparty computation (SMC)\cite{23,24}. DP protects sensitive information through adding perturbation on the private data. Existing studies showed that DP based protection methods\cite{21,25} did not scale well to tackle high dimensional genomic data. HME method\cite{22} is an innovative cryptographic solution, which allows certain computations (e.g., addition and multiplication) to be carried on encrypted data directly and also generates encrypted outcomes. After decryption, the corresponding plaintexts of encrypted results will exactly match the results operated directly on original inputs.
without encryption. Built upon HME, researchers can delegate secure computation on an untrusted environment\textsuperscript{26} (e.g., public cloud services), where they can produce encrypted computation with encrypted results, without revealing their sensitive inputs. During the past few years, the efficiency and capability of HME have been significantly improved\textsuperscript{27-30}. Secure multiparty computation (SMC)\textsuperscript{23,31-33} is another widely used privacy-preserving computing techniques which establish interfaces for multiple parties to jointly compute functions over their own inputs without knowing others’. SMC is widely adopted to safeguard federated studies with sensitive data across multiple institutions\textsuperscript{34,35}. Many HME-based or SMC based privacy-preserving computing applications been proposed, for instance safeguarding machine learning on HME scheme\textsuperscript{36,37}, Homomorphic Computation of Edit Distance\textsuperscript{38,39}, Private predictive analysis on encrypted medical data\textsuperscript{37}, exact logistic regression for GWAS research\textsuperscript{30}, secure multiparty GWAS\textsuperscript{34,35,40-42}, secure distributed regression model learning \textsuperscript{33} etc.. However, current HME and SMC techniques still have challenges with sharply increased computation complexity when the dataset increases in size. For the large-scale genomic dataset, the performance overhead is not negligible and it is the critical problem limiting the design and implement for secure and privacy-preserving cloud genomic computing platform.

Another solution for secure outsourcing computation called “Software Guard Extension (SGX)\textsuperscript{43}” is made available very recently. There is a hardware protected secure region (so called “Enclave”) on Skylake or higher generation CPUs released by the Intel to provide security and integrity on an untrusted host, where privileged modules (e.g., operating system, virtual machine scheduler, etc.) might be potentially malicious\textsuperscript{44}. This approach provides an alternative way to protect private secrets on untrusted hosts through the “enclave” architecture\textsuperscript{45}. Private codes, sensitive data, and other secrets can be stored within enclaves and sealed outside enclave at any time. The access of secrets within enclaves are highly restricted by hardware access control (i.e., using a different instruction set from the main CPU) so that only authorized code has the privilege to access secrets within enclaves. This has attracted the attentions of many researchers recently, who conducted several related works on privacy-preserving cloud computation\textsuperscript{45-47}, multi-party collaboration for disease analysis\textsuperscript{48}, ancestry analysis\textsuperscript{49} secure logger for medical devices \textsuperscript{50} Those research outcomes demonstrated the feasibility of this alternative solution for outsourcing secure genomic computation. Thanks to its optimized hardware design, SGX can provide efficient computation and shows good promise on dealing with large scale genomic datasets.

But SGX still has many limitations as a new technology, which takes time to become mature and gain wider acceptance. One practical problem is that it only provides limited enclave space, large data cannot fit into that will have to be split and sealed outside enclave (encrypted and stored outside) to be processed in batches. Depending on the nature of the task, the process might need careful engineering to ensure optimal performance and meet high security guarantee. Another potential problem is that there are concerns about information leakage in SGX due to the side channel attack\textsuperscript{51}, cache timing attack\textsuperscript{42}, etc. Therefore, additional protections on top of SGX are still required to mitigate above concerns. Another issue is that most healthcare centers have not deployed SGX-enabled servers and models requiring mutual hardware trust are not yet practical.

Inspired by the idea of secret sharing\textsuperscript{52}, we developed a hybrid framework in this paper, which can distribute sensitive information between SGX and HME to achieve a better protection in an untrusted environment. More specifically, we mainly target on the secure count query of genomic data on untrusted hosts, which takes advantages of both HME and SGX techniques. The main contributions of our proposed hybrid methods are as follows:

- We present the first design of hybrid privacy-preserving framework that combines SGX and HME together to facilitate outsourced secure and efficient genomic count query operations.
- We proposed an innovative data provisioning method that split single genomic dataset into an encrypted index tree structure data within SGX enclave and a sealed record data of tree node counts outside the enclave. The query execution will be carried out within enclave through traversing the index tree structure and checking whether a query matches some specific branches. The addition for aggregating actual counts of the query is done outside enclave over the homomorphically encrypted count records via the Paillier cryptosystem\textsuperscript{53}.
- We proposed the dynamical HME key pair generation and exchange schemes incorporate with SGX techniques, which guaranteed the security and integrity of private paillier keys.
- Our proposed hybrid platform has been evaluated over real data from the Personal Genome Project \textsuperscript{54}, and experimental results demonstrated efficiency advantage of our proposed HME and SGX hybrid framework over the existing solution \textsuperscript{55}.

**Methods**

In this section, we present the proposed SCOTCH framework for securely executing count query on genomic data by
using the combined model of HME and Intel SGX. Figure 1 provides an overview of the proposed SCOTCH framework. Our framework is optimized for the dual objective of security and efficiency. Our secure model consists of four roles, including, Data Owner (DO), Authentication Service Provider (ASP), Cloud Service Provider (CSP) and authorized researchers (Researchers). Each interactor is responsible for performing different transactions to contribute the overall platform secure and functional. The interactors are discussed as follow:

**Figure 1:** The overview of the proposed SCOTCH framework

- **Authentication Service Provider (ASP):** A trusted entity, which can authenticate (1) data owner to securely outsource their sensitive and private data to a public cloud, (2) authorized researchers to initiate data analysis query and to securely receive the final results, and (3) the secure enclave that is running on the SGX-enabled hardware hosted by an untrusted Cloud Service Provider.

- **Cloud Service Provider (CSP):** An untrusted entity, which can communicate with all the other entities. It provides securely outsourced storage of encrypted genomic data from the data owner. Based a hybrid strategy using HME and SGX, the CSP will securely evaluate the encrypted query from an authorized researcher against the encrypted data from the data owner to obtain the encrypted final results.

- **Authorized Researchers (Researchers):** Individuals who have been granted the permission by the ASP to execute a query (i.e., the counts of multiple alleles of a genomic database). An authorized researcher sends the query to the CSP and receive encrypted results based on data from certain data owner.

- **Data Owners (DO):** Institutions or hospitals that possess databases upon which would like to outsource both storage and computation on the public cloud and allow authorized researcher to perform queries.

In our design, the framework workflow (see Figure 2) contains three major steps: Key Distribution, Data Provisioning, and Secure Query Execution. Key distribution is responsible for exchanging the HME encryption keys among CSP, DO and authorized researchers, DO will secure pack their private genomic data and upload to CSP through data provisioning, and the Query Execution presents the detail flow of how researchers execute secure query on CSP and count the results. In the remaining part of this section, we will elaborate our approach in detail.

**Step 1. Key Distribution.** A remote attestation procedure is required between ASP and CSP to create a trusted enclave environment. Using remote attestation, CSP can provide the evidences to other entities to prove its integrity and authenticities through the ASP. Only when CSP confirmed its identity and verified that it is running on a genuine SGX-enabled platform, DO or researcher can safely provisioning data or issue genetic query execution on CSP, respectively. After the trusted enclave is created, CSP generates pair of paillier cryptosystem keys within the enclave, the public key $k_{pub}$ will be distributed to DO and Researcher for further encryption of genomic data and queries. For the corresponding private key $k_{priv}$, it will be kept within the enclave, or could be sealed outside the enclave for long term storage purpose. The private key $k_{priv}$ can only be accessed by authorized codes within enclave. The hardware instructions in SGX will prevent malicious attempts from accessing $k_{priv}$.

**Step 2. Data Provisioning.** After the Key Distribution Stage, the CSP has been attested by ASP, and the public key $k_{pub}$ for encryption has already been disseminated to DO and Researcher. Then we can step to the data provisioning stage, which is used for preparing secure genetic dataset on CSP and enabling secure access interface to its users and owners. We assume that the DO hold the private genetic dataset in Single Nucleotide Polymorphism (SNP) sequences
format (see Table 1 as an example). Let us denote by $G = (s_1, s_2, s_3, ..., s_n)$ the genomic dataset held by a DO, where $s_i \in G$ represents the SNP sequence for the $i$-th individual in the dataset, denoted by $s_i = (a_1, a_2, a_3, ..., a_k, p)$, where an element $a_k$ represents a genotype at the $k$-th location, and the last element $p$ indicates whether the individual is associated to a specific disease, like cancer, as shown in Table 1, where each row corresponding to $s_i \in G$, and the columns in table corresponds to each element of $s_i$.

**Figure 2.** Total allergy alerts, overridden alerts, or drug order cancelled.

**Table 1.** Data Presentation for Private Genomic Dataset

<table>
<thead>
<tr>
<th>Individual</th>
<th>SNP_1</th>
<th>SNP_2</th>
<th>SNP_3</th>
<th>SNP_4</th>
<th>SNP_5</th>
<th>SNP_6</th>
<th>SNP_7</th>
<th>...</th>
<th>Cancer</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>AG</td>
<td>CC</td>
<td>CG</td>
<td>AG</td>
<td>AG</td>
<td>CC</td>
<td>GG</td>
<td>...</td>
<td>-</td>
</tr>
<tr>
<td>2</td>
<td>AA</td>
<td>CC</td>
<td>GG</td>
<td>CC</td>
<td>GG</td>
<td>CC</td>
<td>AG</td>
<td>...</td>
<td>+</td>
</tr>
<tr>
<td>3</td>
<td>GG</td>
<td>CG</td>
<td>AG</td>
<td>CC</td>
<td>AA</td>
<td>CC</td>
<td>AG</td>
<td>...</td>
<td>+</td>
</tr>
<tr>
<td>4</td>
<td>AA</td>
<td>CC</td>
<td>GG</td>
<td>AG</td>
<td>AA</td>
<td>GG</td>
<td>CC</td>
<td>...</td>
<td>-</td>
</tr>
<tr>
<td>5</td>
<td>AG</td>
<td>GG</td>
<td>AA</td>
<td>AG</td>
<td>GG</td>
<td>AA</td>
<td>GG</td>
<td>...</td>
<td>-</td>
</tr>
</tbody>
</table>

DO first parses its genomic dataset and builds an index tree using their data. We will build the index tree using the technique introduced in our previous work. Let us denote by $T = (n_1, n_2, ..., n_m)$ the index tree with $m$ nodes, where $n_j \in T$ is the $j$-th node in the tree, and each node is encoded as $n_j = (\text{index}, \text{sid}, \text{val}, \text{count}, \text{list})$, where the index is a unique ID for the node assigned by the Breadth-first search (BFS) traversal order. In addition, sid is the level/depth of the node in the tree, which corresponds to the location of each SNP and can be used to indicate the exact position that this SNP occurs in a query. Furthermore, val indicates the actual SNP which is encoded by a 4-bit vector to cover all possible SNP pairs. The element count is the times of the corresponding SNP occurs in the current position (sid) in the dataset. Finally, the variable list records the index to all children nodes of the current nodes. An example of the index tree for Table 1 is shown in Figure 3. The tree starts with a root node, and all genomic sequences are derived from root node.
Figure 3. An illustrative example of an Index Tree for genomic data from 5 individuals as shown in Table 1. In practice, since not all individuals have a unique SNP combination, the count (encrypted by HME) cannot be predicted directly from the tree structure.

Once DO completed the tree building process, DO extracts the count for each node $n_j$ denoted as $c_j$, then encrypts $c_j$ with the received key $k_{pub}$. Here, let us denote by $c^E_i$ the encrypted count such that $C^E = \{c^E_i \mid i \in \{1, 2, 3, ..., m\}\}$ represents all the encrypted counts. Then, we remove all the counts from the generated index tree (as shown in Figure 3) and only keep the structure information. The structure tree is then encrypted by $k_{pub}$ which is denoted by $T^E$. Finally, two encrypted data $C^E$ and $T^E$ will be uploaded to CSP for both storage and computation outsourcing.

Step 3. Secure Query Execution. After DO completes the data provisioning process, the Researchers can request query execution on the CSP. Our framework supports SQL like search query, which means researchers can use the search query as follows:

```
SELECT count(*) FROM GENDATABASE WHERE
pos_1 = SNP_1 AND pos_2 = SNP_2 AND pos_3 = SNP_3 .... AND pos_i = SNP_i, AND cancer = T/F.
```

The SQL like query will be wrapped into a query vector $q$ followed by an encryption. In addition, the encryption process will only encrypt the SNP values in the query $q$ by using the public key $k_{pub}$, where we consider the position information is public. The vector query and its encryption format are denoted as follows:

$q = \langle <\text{pos}_1, \text{SNP}_1>>, <\text{pos}_2, \text{SNP}_2>>, <\text{pos}_3, \text{SNP}_3>>, ..., <\text{pos}_i, \text{SNP}_i>>, <\text{cancer}, \text{T/F}>\rangle$

$q^E = \langle <\text{pos}_1, \text{SNP}_1^E>>, <\text{pos}_2, \text{SNP}_2^E>>, <\text{pos}_3, \text{SNP}_3^E>>, ..., <\text{pos}_i, \text{SNP}_i^E>>, <\text{cancer}, (\text{T/F})^E>>\rangle$

After receiving an encrypted query $q^E$ from the researcher, the CSP will pass $q^E$ into enclave, in the meantime, CSP will also load $T^E$ into its enclave space. Within the enclave, both encrypted query $q^E$ and encrypted index tree structure $T^E$ are decrypted using the private key $k_{priv}$ to get the plaintext query $q$, $T$. Query search is then executed on the enclave. Algorithm 1 presents the details of query algorithm within the secure enclave. The inputs for this algorithm are the search query and the index tree structure. The algorithm starts with an empty list $l$, and then steps into the loop. For each round, the second element of each pair from $q$ is pop out and stored in a temporary container Pair($<p1, p2>$). Then, we traverse on the tree $T$ to find the node that located at the $p1$ level of $T$ and its node val is equal to $p2$. If a node is found, the index of the node will be returned. Otherwise, an empty value will be returned. The outputs of the algorithm will be a list $l = (i_1, i_2, i_3, ..., i_n)$, including all indices of these nodes that satisfy the search query. In the next step, CSP will perform HME based aggregation outside the enclave based on this list.

First, CSP looks up the list $l$, and use the elements in $l$ as the index to retrieve the encrypted count of the corresponding node. Due to the HME property of Paillier Cryptosystem, suppose we have two message $m1$ and $m2$, if we encrypt the two message with Paillier public key and then perform multiplication of the two ciphertexts we will get the sum of the corresponding plaintexts. According to such homomorphic property, CSP will perform HME aggregation of the lowest level nodes in the list $l$ to compute the encrypted overall counts $Sc^E$, which will be sent to
enclave for re-encryption before releasing the results to researcher.

To perform re-encryption, Researchers need to create a different pair of Paillier key \((pk, sk)\) and share the public key \(pk\) with CSP. CSP pass the public key into enclave and then perform \(Sc^E\) decryption followed by an encryption step of \(Sc\) by using the new key and send back the ciphertext \(pk(Sc)\) to researcher. Researchers then decrypt the ciphertext and get the query count result.

```
Algorithm 1. Algorithm for Executing Search Query

Input : Search Query \(q\), Index Tree Structure \(T\)
Output: List of inodes that satisfy the search query, \(I\)

\[ L = T.getNodes(depth = 1) \quad // \text{Append all tree nodes with depth 1} \]
\[ cur = 0; \quad // \text{cur indicate the current depth(level)} , \]
\[ \text{Initiate cur with 0 indicate start with root node.} \]

\textbf{while} \( cur < \text{max_depth}(T) \) \textbf{do:} \quad // Break loop when reach leaf nodes
  \textbf{if} \( cur \textbf{ not in } q.\text{keys}() \):
    Update \( L \) with the children of current nodes in \( L \)
  \textbf{else}
    Update \( L \) with the children of nodes where \( node.val == q[cur] \)
  \( cur++; \)
return \( L \);
```

**Results**

In our experiment, a dataset retrieved from the Personal Genome Project\textsuperscript{44} with 173 individuals and 60 to 2000 SNPs were used to perform the evaluation. The association test based on case/control group in GWAS could be a downstream application of our model. More specifically, for the chi-squared association test, \[ \sum \frac{(Obs - Exp)^2}{Exp} \], where \( Obs \) represents observed allele counts and \( Exp \) represents expected allele counts, one needs to query the genotype counts of case/control groups to compute the necessary expected frequencies and observed frequencies used in the analysis. For all experiments, we used 1024 bits Paillier HME (i.e., 80 bits security) for data encryption and decryption. All results were obtained on a Windows 10 PC with SGX-enabled i7-6820HK CPU and 48GB memory, on which we simulated all entities in the proposed framework.

Figure 4 shows the time cost for structure tree building, count encryption, and structure tree encryption. All these building time is assumed as one-time job and DO takes charge of building jobs. Table 2 presents the result of total size of different files generated by HME. We can see the encryption of node count is the most time consuming task. As shown in Table 2, we need to encrypt all count number of the tree and the encryption result can be large. But it is only encrypted once, and therefore the consumption is acceptable.

In Table 3, we show the overall query processing time. In the experiment, we test two different sizes of datasets with 500 and 1000 SNPs, respectively. We tested the query size with number of SNPs ranging from 5 to 200. The experimental results showed that HME computation was more time consuming than these of the node comparison within the enclave. As shown in Table 3, HME addition costs and size of queries are inversely proportional. It is because a longer query usually results in less matched nodes in the deep level of the tree, which reduced the number of HME addition operations.
Figure 4: Running time for tree building, count encryption, and tree structure encryption

Table 2: Performance in terms of the data size of the encrypted count and the encrypted structure tree

<table>
<thead>
<tr>
<th>Number of SNPs</th>
<th>60</th>
<th>120</th>
<th>180</th>
<th>300</th>
<th>500</th>
<th>1000</th>
<th>2000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Encrypted Structure Tree Size (KB)</td>
<td>40</td>
<td>58</td>
<td>86</td>
<td>87</td>
<td>252</td>
<td>508</td>
<td>1,020</td>
</tr>
<tr>
<td>Encrypted Count Size (KB)</td>
<td>3,130</td>
<td>4,606</td>
<td>6,664</td>
<td>7,604</td>
<td>21,107</td>
<td>42,819</td>
<td>86,241</td>
</tr>
</tbody>
</table>

Table 3: Performance in terms of overall query processing time (including both HME computation outside the enclave and the secure computation within the enclave) for two different datasets with 500 and 1000 SNPs

<table>
<thead>
<tr>
<th># of SNPs</th>
<th>500 SNPs</th>
<th>1000 SNPs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Query size</td>
<td>5</td>
<td>25</td>
</tr>
<tr>
<td>HME computing</td>
<td>1.085</td>
<td>0.059</td>
</tr>
<tr>
<td>Enclave computing</td>
<td>0.013</td>
<td>0.002</td>
</tr>
<tr>
<td>Overall computing</td>
<td>1.095</td>
<td>0.061</td>
</tr>
</tbody>
</table>

We compare the processing time required to execute a query in our proposed model with Hasan et al.\textsuperscript{55}. The experimental results showed that the size of SNPs has major impact on query processing time, thus, we used a larger number of SNPs instead of more individuals in our experiments. By using Hasan et al.\textsuperscript{55} model, a query consisting of 50 SNPs against a dataset of 300 SNPs each record requires around 6 seconds to process. In contrast, our proposed framework costs approximately 0.063 second to process the same query but against a larger dataset of 500 SNPs each record. Therefore, our framework is at least 95 times faster than Hasan et al.\textsuperscript{55} to process a secure query.

Discussion and limitations

Security Model: In this proposed framework, there are two private secret keys, i.e., the private genomic data from DO and the private search query from Researchers. As ASP only handles attestation and authentication, we consider ASP will not access any of the two secret keys. We consider the following potential leakage situations to show our model is secure and privacy-preserving.

1. **DO data leakage to CSP:** The tree is built on DO side and no other entities have the ability to access the raw data stored on DO. Thus, no one except for the DO can access data during tree building stage. In addition, the tree information will be split into two parts i.e., tree structure and counts of nodes. Both parts will be encrypted and remain protected by HME and SGX based secure hardware at the untrusted CSP side. Thus, CSP will not have the ability to access secrets of DO.
2. **DO data leakage to Researcher:** Researchers also do not have the ability to access data stored on DO, they are also not able to infer information in the tree building process.
3. **Researcher query leakage to CSP:** Researchers’ search queries are encoded and encrypted when sent to CSP. The query will remain sealed or encrypted on CSP outside enclave and can only become in plaintext when enters
a trusted enclave. Thus CSP do not have ability to access the query itself. The count of query operation is performed outside enclave, but the results are calculated through ciphertext via Paillier-based HME computation. Thus CSP also has no chance to infer or access the results of count queries.

4. **Researcher query leakage to DO:** Researchers do not have any interactions with DOs through the secure query execution stage. Thus there is no chance to leak query secrets to DO.

**Limitations:** The proposed SCOTCH framework has several limitations: First, With the increases of number of individual and SNPs, the branches of the structure tree will also increase significantly. Since we need to put the tree structure inside the enclave, the memory limitation of SGX may become an obstacle. When dataset scales up the computation complexity for our proposed increases accordingly. In this paper we only present the method for building tree structure based on position, no further optimization is included. However, based on our proposed method, one can apply bloom filter to optimize the algorithm when dataset scales up, which is an efficient method that can help us to speed up the the searching or re-organizing over tree structure based SNP frequency. Second, we still observe a significant storage overhead for homomorphically encrypted count information, as the count of each node needs to be encrypted independently. Moreover, the number of nodes will also increase, as the dataset increases. Third, even though the SGX-based framework can be constructed to defend malicious attacker, the proposed SCOTCH assume a semi-honest security model to cope with HME. Finally, we consider the position information of the querying SNPs to be non-private, which might not be always true. Advanced protection based on oblivious random access could be used to potentially mitigate such concerns. The above limitations warrant the further investigation along this line.

**Conclusion**

In this paper, we propose a hybrid secure framework based on both HME and SGX to conduct privacy-preserving genomic count queries for cloud computing. The proposed method provides an innovative data provisioning method on the combined hardware/software architecture, which encrypts private genomic datasets into two separate data structures, i.e., homomorphic encrypted tree structure and hardware sealed node counts. Such method reduces the storage overhead and improves the security and integrity of private data. A SGX-enhanced dynamic secure homomorphic encryption key pair generation method is also proposed to improve the security for key management. The results for our experimental studies demonstrated the efficiency of the proposed method. For example, it took less than 1 second to securely query 200 SNPs against a dataset with 1000 SNPs and 173 individuals. In comparison to the previous work 55, which required about 2 seconds to handle a query with 60 SNPs against a database with 240 SNPs, the proposed use only 0.061 seconds to handle a query with 75 SNPs against a database with 500 SNPs. It is worth mentioning that the complexity for tree search is mainly depends on the number of SNPs in the query and the database, while the number of individual in the database have very minor impact on the overall performance 55.

**Funding**

This work was funded in part by the Patient-Centered Outcomes Research Institute (PCORI) under contract ME-1310-07058, NHGRI R00HG08175, NIGMS R01GM114612, R01GM118574, R01GM118609, NLM R21LM012060, NIBIB U01EB023685, NSERC Discovery Grants (RGPI-2015-04147) and University of Manitoba Startup Grant.

**Authors’ contribution**

All authors approved the final manuscript. CW, YJ, XJ, NM and SW designed the method. CW, YJ and SW implemented the algorithms and devised the simulation experiments. CW, YJ, XJ, SW wrote a majority of the manuscript. FC, NM, MA and MS provided useful discussion, edits and critical suggestions.

**References**

32 FlexSC. https://github.com/yhuang912/FlexSC.
34 Constable SD, Tang Y, Wang S, Jiang X, Chapin S. Privacy-Preserving GWAS Analysis on Federated Genomic


Classifying Clinical Trial Eligibility Criteria to Facilitate Phased Cohort Identification Using Clinical Data Repositories

Amy Y. Wang, MD1, William J. Lancaster, MD2, Matthew C. Wyatt, MSHI2, Luke V. Rasmussen, MS1, Daniel G. Fort, PhD, MPH1,3, James J. Cimino, MD, FACMI4

1Department of Preventive Medicine, Northwestern University Feinberg School of Medicine, Chicago, IL;
2Department of Biomedical Informatics, Vanderbilt University, Nashville, TN;
3Ochsner Health System, New Orleans, LA
4Informatics Institute and Department of Medicine
University of Alabama at Birmingham School of Medicine, Birmingham, AL

Abstract

A major challenge in using electronic health record repositories for research is the difficulty matching subject eligibility criteria to query capabilities of the repositories. We propose categories for study criteria corresponding to the effort needed for querying those criteria: “easy” (supporting automated queries), mixed (initial automated querying with manual review), “hard” (fully manual record review), and “impossible” or “point of enrollment” (not typically in health repositories). We obtained a sample of 292 criteria from 20 studies from ClinicalTrials.gov. Six independent reviewers, three each from two academic research institutions, rated criteria according to our four types. We observed high interrater reliability both within and between institutions. The analysis demonstrated typical features of criteria that map with varying levels of difficulty to repositories. We propose using these features to improve enrollment workflow through more standardized study criteria, self-service repository queries, and analyst-mediated retrievals.

Introduction

There is a long history of using paper-based health records to identify potential research subjects. The advent of electronic health records (EHRs) has greatly facilitated researchers’ access to relevant patient data, especially when data are transferred to specialized data repositories or warehouses. Researchers may use such information for estimating the availability of eligible research subjects in some larger target population (cohort estimation), identifying potential patients for enrollment in a research study (cohort identification), or finding patients whose existing data (whether in summary or detailed form) can be used to explore research questions (data reuse). What most such endeavors have in common is a defined, study-specific set of patient characteristics, known as eligibility criteria. These criteria include those that must be present for a patient to be included as a study subject (inclusion criteria) or render the patient unsuitable for inclusion in the study (exclusion criteria). The use of EHR data for subject identification (for any purpose) will typically begin with querying a repository to find patients who meet inclusion criteria and then remove those patients with exclusion criteria.

Ideally, a researcher can access a repository in “self-service” mode to complete the search and retrieval processes independently. However, the required queries are often too complex for this approach and may require the assistance of a data analyst who is familiar with the data and tools and can mediate between the researcher and the repository. Even then, the process may be challenging. Eligibility criteria may be expressed in complex arrangements that are not amenable to automated searching, or they may require data that are either not accessible through query tools or are not recorded in the health record. Studies of matching enrollment criteria using EHR repositories have demonstrated success rates ranging from 23% to 44% due to requirements for temporal restrictions, calculated conditions, or inaccurate medical diagnostic coding.

For example, a study might seek patients with diabetes mellitus on escalating doses of insulin who arrive in the emergency room with elevated blood glucose levels but will be ineligible if they receive treatment prior to enrollment in the study. In this hypothetical case, the researcher or analyst can automatically identify patients with diabetes mellitus who have received insulin previously but can only determine whether they have received escalating doses through complex queries or manual analysis of records of patients who appear to be at least minimally eligible. Finding patients through the repository breaks down completely when a study requires patients who present to the emergency room with new and acute problems, since repository data are only retrospective.
Subject recruitment is further complicated because not only must potential subjects be identified immediately, but they must be enrolled before receiving treatment. It is unreasonable and unethical to expect clinicians to interrupt patient care and delay potentially lifesaving treatment in acute settings. Researchers with such complex, time-sensitive eligibility criteria must often resort to educating front-line health professionals (or posting paper reminders), hoping that clinicians will remember in time to enroll their patients as study subjects. Automated alerts tied to EHRs can be used for simple eligibility criteria but will be difficult to implement in more complex situations that are challenging, even for experienced data analysts.

One solution mimics the practice of using an initial screening protocol in which a large set of potentially eligible patients is identified, and each individual record is examined subsequently against more specific criteria. For the EHR repository version of this method, the researcher or analyst performs an initial query to identify patient records that meet some criteria and then reviews full records manually for the remaining criteria, perhaps flagging patients who might be eligible if they appear in care settings (such as the hypothetical patients with diabetes above) for further evaluation. This multi-step process can be facilitated by characterizing eligibility criteria in advance as either being amenable to retrieval with repository tools (“easy”), requiring some more elaborate mechanism (“hard”), or a combination of the two (“mixed”). If a researcher specifies criteria in this manner in a research protocol, repository users can then perform an initial retrieval step (especially when the user is an analyst rather than the researcher) to identify an initial set of patients. This cohort then passes on to the next stage in the process for application of additional study criteria through other means.

The categorization of study criteria by level of difficulty has been previously demonstrated. Three reviewers from a single research institution with combined expertise in clinical medicine, research, EHRs, and querying databases for study cohorts demonstrated high agreement when assigning levels of difficulty to study criteria. We have now expanded the original study to include a similar set of expert reviewers from a second research institution. The purpose of this follow-on study is to further characterize the use of these categories and determine if the prior findings are generalizable. Demonstrating inter-institutional reliability and examining differences in categorization between institutions are important steps in determining the overall feasibility of our proposed phased approach for identifying study cohorts.

![Figure 1](image-url). Sample ClinicalTrials.gov record, showing Eligibility Criteria in Tabular View. (Screen has been edited somewhat to allow inclusion of header and tabular information in the same view.)
Table 1. Research descriptions from ClinicalTrials.gov that were selected for this study. Studies were paired to include ten UAB studies and ten studies that followed them sequentially in the ClinicalTrials.gov database.

<table>
<thead>
<tr>
<th>NCT ID</th>
<th>Study Title</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>01098981</td>
<td>Phase 3, Randomized, Placebo-Controlled, Double-Blinded Trial of the Combined Lysis of Thrombus With Ultrasound and Systemic Tissue Plasminogen Activator (tPA) for Emergent Revascularization in Acute Ischemic Stroke (CLOTBUST-ER)</td>
<td><a href="https://clinicaltrials.gov/ct2/show/NCT01098981">https://clinicaltrials.gov/ct2/show/NCT01098981</a></td>
</tr>
<tr>
<td>01098994</td>
<td>Haptoglobin Phenotype, Vitamin E and High-density Lipoprotein (HDL) Function in Type 1 Diabetes (HAP-E)</td>
<td><a href="https://clinicaltrials.gov/ct2/show/NCT01098994">https://clinicaltrials.gov/ct2/show/NCT01098994</a></td>
</tr>
<tr>
<td>01382212</td>
<td>A Study to Evaluate the Safety of Paricalcitol Capsules in Pediatric Subjects Ages 10 to 16 With Stage 5 Chronic Kidney Disease Receiving Peritoneal Dialysis</td>
<td><a href="https://clinicaltrials.gov/ct2/show/NCT01382212">https://clinicaltrials.gov/ct2/show/NCT01382212</a></td>
</tr>
<tr>
<td>01382225</td>
<td>Sodium Hyaluronate Ophthalmic Solution, 0.18% for Treatment of Dry Eye Syndrome</td>
<td><a href="https://clinicaltrials.gov/ct2/show/NCT01382225">https://clinicaltrials.gov/ct2/show/NCT01382225</a></td>
</tr>
<tr>
<td>01797445</td>
<td>Study to Evaluate the Safety and Efficacy of E/C/F/TAF (Genvoya®) Versus E/C/F/TDF (Stribild®) in HIV-1 Positive, Antiretroviral Treatment-Taive Adults</td>
<td><a href="https://clinicaltrials.gov/ct2/show/NCT01797445">https://clinicaltrials.gov/ct2/show/NCT01797445</a></td>
</tr>
<tr>
<td>01797458</td>
<td>European Study on Three Different Approaches to Managing Class 2 Cavities in Primary Teeth</td>
<td><a href="https://clinicaltrials.gov/ct2/show/NCT01797458">https://clinicaltrials.gov/ct2/show/NCT01797458</a></td>
</tr>
<tr>
<td>02121795</td>
<td>Switch Study to Evaluate P/TAF in HIV-1 Positive Participants Who Are Virologically Suppressed on Regimens Containing ETC/TDF</td>
<td><a href="https://clinicaltrials.gov/ct2/show/NCT02121795">https://clinicaltrials.gov/ct2/show/NCT02121795</a></td>
</tr>
<tr>
<td>01720446</td>
<td>Trial to Evaluate Cardiovascular and Other Long-term Outcomes With Semaglutide in Subjects With Type 2 Diabetes (SUSTAIN™ 6)</td>
<td><a href="https://clinicaltrials.gov/ct2/show/NCT01720446">https://clinicaltrials.gov/ct2/show/NCT01720446</a></td>
</tr>
<tr>
<td>01720459</td>
<td>Effects of Micronized Trans-resveratrol Treatment on Polycystic Ovary Syndrome (PCOS) Patients</td>
<td><a href="https://clinicaltrials.gov/ct2/show/NCT01720459">https://clinicaltrials.gov/ct2/show/NCT01720459</a></td>
</tr>
<tr>
<td>01897233</td>
<td>Study of Lumacaftor in Combination With Ivacaftor in Subjects 6 Through 11 Years of Age With Cystic Fibrosis, Homozygous for the F508del-CFTR Mutation</td>
<td><a href="https://clinicaltrials.gov/ct2/show/NCT01897233">https://clinicaltrials.gov/ct2/show/NCT01897233</a></td>
</tr>
<tr>
<td>01567527</td>
<td>Efficacy, Safety, and Tolerability of an Intramuscular Formulation of Aripiprazole (OPC-14597) as Maintenance Treatment in Bipolar I Patients</td>
<td><a href="https://clinicaltrials.gov/ct2/show/NCT01567527">https://clinicaltrials.gov/ct2/show/NCT01567527</a></td>
</tr>
<tr>
<td>01567533</td>
<td>A Pilot Study Evaluating Safety of Sotaglifloptin Combined With Peg-IFN Alfa-2a + Ribavirin in Chronic Hepatitis C Patients</td>
<td><a href="https://clinicaltrials.gov/ct2/show/NCT01567533">https://clinicaltrials.gov/ct2/show/NCT01567533</a></td>
</tr>
<tr>
<td>01936688</td>
<td>A Study to Evaluate the Efficacy and Safety/Tolerability of Subcutaneous MK-3222 in Participants With Moderate-to-Severe Chronic Plaque Psoriasis (MK-3222-012)</td>
<td><a href="https://clinicaltrials.gov/ct2/show/NCT01936688">https://clinicaltrials.gov/ct2/show/NCT01936688</a></td>
</tr>
<tr>
<td>01936701</td>
<td>Comparison of the Effect of Hydrophobic Acrylic and Silicone 3-piece IOLs on Posterior Capsule Opacification</td>
<td><a href="https://clinicaltrials.gov/ct2/show/NCT01936701">https://clinicaltrials.gov/ct2/show/NCT01936701</a></td>
</tr>
<tr>
<td>01833533</td>
<td>A Study to Evaluate Chronic Hepatitis C Infection in Adults With Genotype 1a Infection (PEARL-IV)</td>
<td><a href="https://clinicaltrials.gov/ct2/show/NCT01833533">https://clinicaltrials.gov/ct2/show/NCT01833533</a></td>
</tr>
<tr>
<td>01833546</td>
<td>A Japanese Phase 1 Trial of TH-302 in Subjects With Solid Tumors and Pancreatic Cancer</td>
<td><a href="https://clinicaltrials.gov/ct2/show/NCT01833546">https://clinicaltrials.gov/ct2/show/NCT01833546</a></td>
</tr>
</tbody>
</table>

Methods

We obtained the ClinicalTrials.gov (NCT) identifiers for a convenience sample of ten studies conducted at the University of Alabama at Birmingham (UAB) that were chosen for a separate study of the use of EHR data for cohort prediction (unpublished data). We then incremented each NCT ID successively until the next ID in sequence matched another study in ClinicalTrials.gov. For example, the NCT ID NCT01098981 would be incremented to NCT01098994 to find the next additional study. These twenty studies provided the eligibility criteria that served as the data set for our study (Table 1). The NCT record for each study was examined to identify its eligibility criteria (Figure 1) which were manually copied from a Web browser into a spreadsheet. Unnecessary words were removed from each criterion to make it easier to identify those that were essentially redundant. Duplicate criteria were then removed, and the remaining criteria served as the set to be reviewed for the study.

The authors of this study served as experts for rating the eligibility criteria. We included three experts each from medical schools within two academic research institutions, University of Alabama School of Medicine (UASOM)
and Northwestern University Feinberg School of Medicine (NU), to improve the generalizability of our findings and explore factors that might be institution- or repository-specific.

- JJC is a physician who was the principle architect of the National Institutes of Health’s Biomedical Translational Research Information System (BTRIS) and has extensive experience using it to match patients to research criteria.
- WJL is a physician (formerly at UASOM) who is a research fellow with extensive experience with electronic health records and moderate experience with UASOM’s i2b2 (Informatics for Integrating Biology to the Bedside) repository.
- MCW is a systems analyst architect with responsibility for and experience using both i2b2 and PowerInsight (Cerner Corporation, Kansas City, MO), the clinical data warehouse attached to the EHR used by UASOM.
- LVR is a clinical research associate at NU who leads the implementation of the i2b2 instance within the Northwestern Medical Enterprise Data Warehouse (NMEDW) and trains faculty and staff on its proper use.
- AYW is a physician with extensive experience developing and implementing health information standards and moderate experience using EHRs and using the i2b2 repository within NMEDW.
- DGF is a research fellow at NU with considerable experience researching and using cohort selection queries for clinical trial recruitment.

Each author reviewed the resulting criteria and rated them with an ordinal scale from one to four (Table 2) that was assessed whether a typical clinical data repository, such as i2b2, BTRIS or PowerInsight, could identify study subjects matching each criterion based on information available in a typical EHR system. Given the complexity of some criteria, in which multiple clinical concepts or properties are represented together (e.g., classes of diagnoses, severity of disease in addition to presence of disease), ratings were assigned based on the highest level of difficulty. For example, if a criterion named a diagnosis and then specified its severity against a separate evaluation scale and a rater assigned 1 and 3 respectively for each sub-criterion, the overall rating for this item would be 3. Reviewers were instructed to leave the rating blank if there was not a clear answer. The ratings were compiled and summarized to determine the characteristics of criteria that could be readily used with EHR repositories versus those that could not.

Table 2. Rating Scale for Eligibility Criteria. Each criterion was assessed by each rater for its suitability for retrieval from an electronic health record data repository.

<table>
<thead>
<tr>
<th>Rating</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 – “Easy”</td>
<td>The criterion could be used easily to identify subjects in an EHR data repository, assuming the data were present, using the repository’s user interface</td>
</tr>
<tr>
<td>2 – “Mixed”</td>
<td>The criterion could be partly retrieved using the repository’s user interface but then would require further manual review</td>
</tr>
<tr>
<td>3 – “Hard”</td>
<td>Matching the criterion would require manual review of the record</td>
</tr>
<tr>
<td>4 – “Impossible”</td>
<td>“Time of enrollment” – The repository would not be likely to include sufficient information to match patients to the criterion, either because the source EHR would not be likely to contain the information or the repository would not include the information in a timely manner</td>
</tr>
</tbody>
</table>

Since this study involved a fully crossed design, in which each criterion was rated by multiple coders using an ordinal coding system, we calculated the intra-class correlation score (ICC). This analysis was performed using R v3.3.1 and the irr package v0.84 with results integrated using StatTag v3.0. Additionally, we analyzed the number of “near agreement” results across institutions. We defined “near agreement” within an institution as a criterion in which two of the three ratings matched exactly, and the discrepant score was only offset by one category. Likewise, “near agreement” across all raters was classified as at least four of the six raters having matching scores, and the discrepant scores were only offset by one category each.

Results

Criteria Data Set

The 20 clinical trials used in this study are listed in Table 1. The ClinicalTrials.gov records for these twenty studies were cleaned to remove blank spaces, formatting, bullets, numbering, and other redundant text which did not affect meaning. Leading clauses with age and gender criteria (e.g., “women” in “women who are pregnant”) were removed, as these criteria are indicated separately in ClinicalTrials.gov and are already straightforward to retrieve. Other text, such as “with” in “with diabetes mellitus” and “taking” in “taking aspirin,” was removed to yield distinct concepts (e.g., diseases or medications). This process yielded a set of 301 criteria, of which 292 were found to be conceptually unique and were reviewed by each rater (1,752 ratings total). Example ratings are shown in Table 3.
Table 2. * indicates criteria for which there was complete or near agreement, included in further analysis. Raters used “X” or “?” to indicate that the meaning of the criterion was unclear or that the categorization was unknown.

<table>
<thead>
<tr>
<th>UAB Raters</th>
<th>NU Raters</th>
<th>Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>A B C A B C</td>
<td>A B C</td>
<td>Abnormal liver function as defined in the protocol at Screening *</td>
</tr>
<tr>
<td>1 1 1</td>
<td>2</td>
<td>absolute neutrophil count ≥ 1,000/mm³ *</td>
</tr>
<tr>
<td>1 1 3</td>
<td>3 1</td>
<td>allergy to starch powder or iodine. *</td>
</tr>
<tr>
<td>1 1 1</td>
<td>1 1 1</td>
<td>Anti-cancer treatment prior to trial entry *</td>
</tr>
<tr>
<td>1 1 1</td>
<td>1 1 1</td>
<td>BMI 40 - 80 kg / m² *</td>
</tr>
<tr>
<td>1 1 2</td>
<td>1 1 1</td>
<td>digoxin within 6 months of starting treatment. *</td>
</tr>
<tr>
<td>1 1 3</td>
<td>1 1 1</td>
<td>immunosuppressants within 6 months of starting treatment</td>
</tr>
<tr>
<td>1 1 3</td>
<td>1 1 1</td>
<td>Anti-diabetic drug naive, or treated with one or two oral antidiabetic drug (OADs), or treated with human Neutral Protamin Hagedorn (NPH) insulin or long-acting insulin analogue or pre-mixed insulin, both types of insulin either alone or in combination with one or two OADs</td>
</tr>
<tr>
<td>1 2 2</td>
<td>2 2 2</td>
<td>currently being treated for secondary hyperparathyroidism *</td>
</tr>
<tr>
<td>2 3 1</td>
<td>2 2 4</td>
<td>Chronic heart failure New York Heart Association (NYHA) class IV</td>
</tr>
<tr>
<td>4</td>
<td>x</td>
<td>For women, effective contraception during the trial and a negative pregnancy test (urine) before enrollment</td>
</tr>
<tr>
<td>2 2 2</td>
<td>2 2 2</td>
<td>Active or untreated latent tuberculosis (TB) *</td>
</tr>
<tr>
<td>2 2 2</td>
<td>2 2 4</td>
<td>Pregnancy or lactation period *</td>
</tr>
<tr>
<td>2 3 4</td>
<td>2 2 1</td>
<td>Dermal disorder including infection at anticipated treatment sites in either axilla.</td>
</tr>
<tr>
<td>2 3 4</td>
<td>4 4 4</td>
<td>Evidence of lens opacity or cataract at the Screening</td>
</tr>
<tr>
<td>3 1</td>
<td>4 4 4</td>
<td>Presence of corneal and conjunctival staining.</td>
</tr>
<tr>
<td>3 3 3</td>
<td>1 1 1</td>
<td>HDSS score of 3 or 4. *</td>
</tr>
<tr>
<td>3 3 3</td>
<td>4 4 3</td>
<td>Prior treatment with any investigational drug within the preceding 4 weeks prior to study entry. *</td>
</tr>
<tr>
<td>3 3 3</td>
<td>4 4 4</td>
<td>Screening genotype report must show sensitivity to elvitegravir, emtricitabine, tenofovir DF *</td>
</tr>
<tr>
<td>3 3 4</td>
<td>4 4 4</td>
<td>person deprived of liberty by judicial or administrative decision *</td>
</tr>
<tr>
<td>3 4 4</td>
<td>3 3 4</td>
<td>Able to swallow tablets *</td>
</tr>
<tr>
<td>3 4 4</td>
<td>3 3 3</td>
<td>good health *</td>
</tr>
<tr>
<td>3 4 4</td>
<td>3 3 3</td>
<td>good overall physical constitution *</td>
</tr>
<tr>
<td>3 4 4</td>
<td>3 3 3</td>
<td>Legal incapacity or limited legal capacity *</td>
</tr>
<tr>
<td>3 4 4</td>
<td>4 4 4</td>
<td>Living conditions suggesting an inability to track all scheduled visits by the protocol *</td>
</tr>
<tr>
<td>3 4 4</td>
<td>4 4 4</td>
<td>Maintenance of a diet consisting of &lt;40 g of carbohydrate per day within 3 months of screening *</td>
</tr>
<tr>
<td>3 4 4</td>
<td>4 4 4</td>
<td>must have completed Part 2 of the base study</td>
</tr>
<tr>
<td>4 4 4</td>
<td>4 4 4</td>
<td>Ability to understand and sign a written informed consent form, which must be obtained prior to initiation of study procedures *</td>
</tr>
<tr>
<td>4 4 4</td>
<td>4 4 4</td>
<td>Facial hair *</td>
</tr>
<tr>
<td>4 4 4</td>
<td>4 4 4</td>
<td>Parents/children who refuse to participate in the study *</td>
</tr>
<tr>
<td>4 4 4</td>
<td>4 4 4</td>
<td>Unwillingness/inability to limit antioxidant supplement use to study-provided supplements *</td>
</tr>
<tr>
<td>4 4 4</td>
<td>4 4 4</td>
<td>Willing to be examined *</td>
</tr>
<tr>
<td>4 4 4</td>
<td>4 4 4</td>
<td>Current alcohol or substance use judged by the investigator to potentially interfere with study compliance</td>
</tr>
</tbody>
</table>

Aggregate Criteria Ratings

Aggregate ratings are summarized in Table 4. For UAB raters, the ICC was 0.91 (95% CI: 0.88, 0.93) for 286 criteria (excluding 6 criteria with blank responses). Of 286 criteria, 232 (81.12%) demonstrated complete or near-complete agreement (2 reviewers at a site agreed completely while the third disagreed by only one category). For NU raters, the ICC for agreement was 0.92 (95% CI: 0.90, 0.93) for 292 criteria. Of 292 criteria, 262 (89.73%) demonstrated complete or near-complete agreement.

For all raters from UAB and NU collectively, (including 286 criteria), the ICC for consistency was 0.92 (95% CI: 0.90, 0.93) and the ICC for agreement was 0.92 (95% CI: 0.90, 0.93). Of the 286 criteria where all raters provided a response, 181 (63.3%) indicated complete or near-complete agreement (at least 4 reviewers agreed completely while the others differed by only one category). All ICC results were statistically significant (p < 0.00001).
Table 4. Individual and combined aggregate ratings for University of Alabama at Birmingham (UAB) and Northwestern University (NU).

<table>
<thead>
<tr>
<th>Characteristics of Easy Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>These criteria were defined as those that can be used easily to identify subjects in an EHR data repository, assuming the data were present, using the repository’s user interface. In general, these were easily related to the categories of data that are provided by repositories like i2b2 and BTRIS, such as laboratory results, vital signs, diagnoses, procedures, and allergies. The raters generally rated as “easy” criteria that contained defined, single concepts, such as diseases like “glaucoma” or individual laboratory tests combined with discrete values, such as “hemoglobin ≥ 9 g/dL.” They also selected “easy” for general descriptions that could be addressed through class-based queries, such as “Abnormal laboratory tests,” “Anti-cancer treatment prior to trial entry,” and “History of solid organ or hematological transplantation.” There was general agreement that criteria involving complex phenotypes were considered “easy” if they could be addressed through an assemblage of individual easy criteria, such as “infection with human immunodeficiency virus (HIV), Hepatitis B, or Hepatitis C” or “Aspartate aminotransferase (AST) and alanine aminotransferase (ALT) ≤ 5 × the upper limit of the normal range (ULN).”</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Characteristics of Mixed Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mixed criteria were defined as having elements of easy and hard criteria, and these criteria can be partly retrieved using the repository’s user interface but then would require further manual review. These criteria generally included some mention of an “easy” criterion, coupled with some restriction or co-occurring state or procedure that either was not itself an easy criterion, or the relationship between the two components can only be carried out by a manual review. Most mixed criteria involved the presence of some medical condition with restrictions on the state, such as severity (e.g., “Severe cardiovascular disease (defined by NYHA ≥3”), chronicity (e.g., “current manic episode with a duration of ≥ 2 years”), or failure to respond to treatment (e.g., “Active or untreated latent tuberculosis (TB).”). Some criteria consisted of an “easy” condition paired with a “hard” criterion, such as “pregnant or breast feeding” and “Seizure disorders requiring anticonvulsant therapy”. There were also cases of findings that would be expected to appear only as narrative text in a diagnostic procedure report. In these text reports, the procedure could be readily identified, but detecting the findings or test results would require manual review, such as in “At least one primary molar tooth with caries into dentine involving two dental surfaces (diagnosed according to International Caries Detection and Assessment System - ICDAS, codes 3 to 5”), “homozygous for the F508del-CFTR mutation”, and “Normal electrocardiogram (ECG).” Finally, there were some criteria that combined “easy” laboratory findings with “hard” ones, such as “Estimated glomerular filtration rate (eGFR) ≥ 50 mL/min according to the Cockcroft-Gault formula for creatinine clearance.”</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Characteristics of Hard Criteria</th>
</tr>
</thead>
</table>
| Hard criteria were defined as those which would require manual review of the patient record. These criteria span the breadth of EHR findings that typically appear in clinical notes and procedure reports. Some hard criteria were encountered as parts of “mixed” criteria, such as “females who are breastfeeding”. Others were statements of patient condition that would not be reflected in a typical problem list, such as “prisoner”. Scale-based assessments, such as “Pre-morbid modified Rankin score of 0-1” and “Eastern Cooperative Oncology Group (ECOG) performance status of 0 or 1” were common. Finally, some criteria might be found in the Assessment portion of clinical notes or require the researcher to use the clinical record to make the assessment, such as “Candidate for phototherapy or systemic...
therapy” and “Any other clinical condition or prior therapy that, in the opinion of the Investigator, would make the individual unsuitable for the study or unable to comply with dosing requirements.”

Characteristics of Impossible (Point of Enrollment) Criteria

Criteria were “impossible” or “point of enrollment” if the repository would not be likely to include sufficient information to match patients to the criterion. While some of the “impossible” criteria related to the time factor of the data (such as “Parents/children who refuse to participate in the study” and “presenting within timeframe for intravenous tPA treatment approved by local regulatory authorities but no more than 4.5 hours from onset of symptoms”), other cases depended on the raters’ opinions about whether relevant data are reliably found in EHRs.

Researchers may be surprised that physical findings they would consider basic were considered by the raters to be generally absent, such as “Facial Hair”, and “good health.” Similarly, the raters agreed that information a researcher would typically obtain from a prospective subject or caregiver, such as “patients with impaired decision making ability”, and “Able to swallow tablets,” would likely not be documented in routine clinical care.

Differences in Ratings between Institutions

There were systemic differences in ratings between UAB and NU (Table 3). For example, for criteria specifying that a disorder, medication, or laboratory result must be present at screening, UAB reviewers unanimously rated these criteria as “easy” while NU reviewers rated them as “impossible.” There were additional criteria which the NU reviewers consistently deemed more challenging than the UAB reviewers. There were other criteria that UAB reviewers found more difficult than NU reviewers.

General Recommendations

The data presented above can be summarized by separating out “easy,” “mixed,” “hard,” and “impossible” criteria into explicit lists (see Table 5). This list, in turn, can be used by those developing eligibility criteria such that they can anticipate the workflow needed to identify potential research subjects using repository query tools, manual record review, and real-time interventions at recruitment sites. Potential uses for this list are explored further in the Discussion section, below.

Discussion

The use of EHR data for clinical research is a rapidly evolving phenomenon. The development of methods for “large pragmatic trials” is especially dependent on EHR data. However, the methods for defining eligibility criteria remain largely unchanged in practice, with little apparent attention paid to the fact that criteria elicited from patients completing questionnaires might differ fundamentally from those elicited from EHR.

---

* A search of PubMed on March 10, 2016 for articles with the phrase “electronic health record data for research” yields 722 citations from 1996 to 2005 and 10,303 citations from 2006 to the present.
Much informatics research has examined the semantics of clinical research eligibility criteria. Projects such as the Ontology of Clinical Research (OCRe)\textsuperscript{19} and the Agreement on Standardized Protocol Inclusion Requirements for Eligibility (ASPIRE)\textsuperscript{20} have defined broad classes of criteria (such as demographics, disease-specific features, functional status, etc.) with the intent of making them computable for comparison across multiple studies. The semantic complexity of these classes of criteria has been studied in depth by Ross and colleagues.\textsuperscript{21} Weng and colleagues provided a comprehensive review of this work\textsuperscript{22} and extended it with classification of criteria through use of the National Library of Medicine’s Unified Medical Language System (UMLS) semantic types,\textsuperscript{23} UMLS semantic network,\textsuperscript{24} and National Institute of Health’s (NIH) Common Data Elements.\textsuperscript{25} To our knowledge, however, there have been no systematic studies of how such criteria translate operationally into the potential capabilities of EHR repository query tools.\textsuperscript{2,4}

We did not set out to create a reusable criteria rating scale, but rather sought to find a specific set of criteria that could be used for further analysis. Nevertheless, we experienced high interrater correlation in the use of our scale within and between our two institutions. The sets of researchers have diverse backgrounds (medicine, informatics, and information systems) and draw on experience with three very different repositories (PowerInsight, i2b2 and BTRIS). This result suggests the raters were tracking a real and meaningful concept and that a standardized and generalizable scoring metric could be generated and applied at other institutions. Examples of criteria on which we agreed and disagreed are included in Table 3 for the reader to judge whether our rating method appears generalizable.

We identified characteristics that affect the difficulty of applying different criteria. Based on these findings, we can develop strategies to convert mixed, difficult, or impossible criteria into easier categories. For example, many criteria consist of combinations of simpler criteria or are bound by restrictions such as severity or time constraints. Separating combined criteria into individual components would likely make them easier to apply. Also, specifying clinical trial criteria using controlled terminologies (and increasing the use of structured data entry in EHRs) can promote improved automated patient matching. A shared library of shared queries for criteria can improve the querying process.

There were interesting systematic differences in ratings between the two institutions. Some criteria consistently or even unanimously received completely opposite ratings at the two institutions. In particular, criteria specifying a condition that must be present at the time of screening were unanimously rated as “easy” at UAB but “impossible” at NU. This difference illustrates that the wording of inclusion and exclusion criteria may be ambiguous, subject to individual interpretation, and a source of systemic inconsistency. UAB reviewers may have interpreted “screening” to indicate the time a researcher performs an initial query on the EHR repository, and the correct information would be searchable in the EHR. NU reviewers may have interpreted “screening” to mean the time that a patient presents for an initial screening encounter with a study coordinator. Because the relevant EHR data are retrospective, they may not be current at the time of a screening visit. This discrepancy does not invalidate the use of our categories for criteria. Instead, it highlights that criteria must be written so that their meanings are expressed clearly and unambiguously. It would be revealing to examine other systematic differences in ratings and differences in repository management and query processes at the two institutions to uncover other possible reasons for these differences.

Our examination of a randomly collected set of enrollment criteria suggests that a majority can be readily classified as having some components retrievable from EHR repositories (i.e. rated “easy” or “mixed”). While not necessarily representative, it does imply a large degree of inclusion and exclusion criteria are accessible via structured data. Our preliminary work in the adjacent field of cohort selection queries leads us to suggest that a significant amount of potential participant identification could be accomplished with intelligent use of these structured data.

The most interesting category may be those criteria widely rated as “mixed,” in which a “first pass” or approximate identification may be possible using structured data. This rating therefore represents an obvious break point in complexity. We suggest future research focus on such criteria, ideally looking to separate or better explicate what can be solved with structured data and EHR retrieval and what might must be retrieved with either full chart review or patient screening.

We do not make any claims, given our sample size and sampling method, about the actual ratios between easy, hard and impossible (time of enrollment) criteria across the spectrum of clinical trials. But, it seems logical to take advantage of the knowledge that such distinctions exist for developing the methods to be used for matching patients to criteria when EHR data repositories are being used for at least part of the process. A potential next step is to develop an interactive “wizard” with a graphical user interface that prompts researchers to enter criteria in structured ways that progress in order of increasing complexity. For example:
The wizard would ask first for easy elements that can be addressed readily with a repository search tool:

- What laboratory test result would you like, and with what range of values?
- What class of medications should the patient be taking (or not taking) and for what duration?
- What procedure should have already been or not been performed?
- Are diagnoses from billing data sufficient or are higher-quality sources (such as problem lists) required?

The wizard would then address more complex elements needing manual review of records found with initial criteria:

- What calculation would you like performed on the previously specified laboratory test result?
- What should the indication for the previously specified medication be?
- What findings should be present in, or absent from, the text report for the previously requested procedure?
- What complex phenotypic pattern should the patient have?

The application would then proceed to elements that can only be ascertained at the time of enrollment:

- What acute event will the otherwise eligible patient have (or not have) at the time of presentation?
- What will the patient have to agree to do (or not do) in order to be eligible for enrollment?

This wizard would then produce a coherent set of eligibility criteria that could specify a multi-staged enrollment process that starts with a set of data queries, proceeds to manual review of full records, and then specifies what additional information will be needed on specific potential subjects through direct contact, perhaps by flagging them in an EHR’s alerting system to notify the researcher when they appear in a patient care setting.7

The findings of our study are a building block toward the development of an approach to improve the specification of eligibility criteria. An understanding of the pragmatics of interfacing with actual clinical data in repositories can inform previous work on the semantics and interoperability of criteria. A set of questions, such as those listed above, guided by knowledge of criteria semantics, will help to establish a dialogue between the investigator and the data analyst (or provide an opportunity for introspection if the investigator is accessing a repository in self-service mode). Further study can lead to refinements in the categories and further strategies and tools for prioritizing criteria. Improving self-service queries will likely require an iterative process of extensions, corrections and refinements.

**Conclusion**

The findings of this study support the hypothesis that clinical research eligibility criteria fall into stereotypical categories with respect to ease of querying data available in EHR repositories. Describing these categories explicitly supports the potential development of a structured process using a data entry form or interactive “wizard.” Such an application may encourage researchers to express their criteria in ways that promote realistic expectations for the recruitment process and improve data retrieval by analysts and researchers.

**Acknowledgments**

This research was supported in part by the University of Alabama School of Medicine Informatics Institute. The authors also thank the University of Alabama at Birmingham Center for Clinical and Translational Science, funded by the National Institutes of Health, National Center for the Advancement of Translational Science (NCATS) (1TL1TR001418-01); the Northwestern University Clinical and Translational Sciences (NUCATS) Institute, funded by NCATS (TL1TR001423); and the National Institute of General Medical Sciences (R01GM105688).

**References**


What are they trying to do?: An analysis of Action Identities in using electronic documentation in an EHR
Charlene R. Weir PhD, Catherine Staes PhD, MPH, Stacey Slager MS, Teresa Taft PhD, Valiammai Chidambaram M.Sc, Heidi Kramer PhD, Bruce E. Bray MD, Seneca Perri Moore PhD, RN

Department of Biomedical Informatics, University of Utah, Salt Lake City, Utah

Abstract

Documentation processes have changed substantially with EHR adoption. User satisfaction studies have focused on usability or cognitive analysis perspectives. Few studies have provided useful information to developers to improve designs. The purpose of this study is to report a 3-pronged approach to deepen understanding of the documentation process, with the intent to provide useful information for future design. This study was conducted in two phases, beginning with cognitive task interviews and observations, followed by post-observation interviews. Twenty-five constructs were identified across the phases, and we observed several patterns of note writing. Participants provided useful information to potentially inform future design. Our study illustrates how electronic documentation serves many clinical processes and is at the core of the medical record. Providers need multiple kinds of notes and ways to display notes. In order to meet provider goals, we must completely re-think the way electronic documentation is composed and displayed.

Introduction

Adoption of electronic health records (EHRs) has resulted in large-scale use of computerized documentation. An emerging tension has resulted between provider’s traditional documentation practices as a component of clinical care, communication and record keeping versus the drive to use computerization to support structured data, in order to feed automated decision support, billing and general quality tracking.

This tension plays out in the complaints of clinicians regarding the usefulness and quality of current electronic notes. These complaints range from a widespread concern with increased time burden, inefficiencies as well as patient safety issues associated with copy and paste tools.(1-9) In addition, the notes themselves have become longer and perhaps less informative. (8-12) In fact, electronic documentation is one of the main areas of clinician’s complaints about the burden of computers and is often thought to be wholly inadequate. Yet, clinicians spend on average 60% of their time in notes.

To address these complaints, vendors and EHR designers have incorporated many time saving features in the documentation software, including templates that automatically insert text, labs, problem lists and other information in addition to copy and paste tools. Although these features have improved efficiency, they have resulted in increasing concerns for patient safety as well as for the overall quality of clinical progress note documentation. In one qualitative study using focus groups of experienced VA users, Weir, et al. identified emergent issues including unreadable notes, template-created errors, lower quality information, and inability to track the patient’s story.(13) Embi, et al’s work found similar results with house staff in another institution.(2) Roosenbloom found that although users noted increased accessibility and improved communication, they also complained that electronic documentation was less professional and informative.(6) The source of this dissatisfaction remains unclear, but likely goes beyond time saving documentation features. Note writing is not serving all of the goals in clinical documentation. Prior research may not be digging deep enough to understand user’s goals at the level that could really be used to improve design.(14-16)

In addition, the use of electronic documentation has changed the way clinicians do work, suggesting that new action identities or goals may be emerging out of a transformation process that has emerged with the adoption of the EMR. Such a transformation is highly likely, because of all of the clinical workflow processes, electronic documentation has been the most disruptive. When paper was used, notes were short, to the point and unreadable. With electronic documentation, notes can be as long as 5 pages, full of unnecessary information that is hard to read and integrate. Note writing can take hours or more.(17, 18) Given this transformation, we need to understand the current practice patterns and workflows and model the purpose of electronic documentation in the new world. The workflow of documentation has changed as electronic notes are used to support handovers, serve as event logs, and as tools to organize thinking.
The purpose of this paper is to report a qualitative study to clarify user’s action identities or goals associated with computerized clinical documentation. Hopefully, these results will contribute to the dialogue about how we can rethink the role of progress notes in an EHR.

**Action Identities or Task Goals**

Action representations are the complex mental representations or schemas that support the full action cycle. These mental representations are hierarchically organized and include the meaning of the task, the contextual cues that will trigger the action, the physical actions learned to support the task, and the overall purpose or goal that is trying to be achieved. When these complex schemas are activated, perception becomes more goal oriented, related memory becomes more available and many behavioral tendencies are automatic. Because many different actions can support the same high level goal, understanding goals and their associated action identities may be one of the most important steps in designing systems that will in the end change behavior.(19-21)

Measuring goals is complex because experts are often not aware of their actions (but may be aware of their goals) and action identification (how people label about what they are doing) varies as a function of expertise, difficulty and role. For example, novices often describe their behavior in terms of the actions involved and rarely refer to purpose or goal. Similarly, those in charge will describe their work in terms of meaning or big-picture goals, whereas subordinates are much more likely to describe their tasks in more concrete terms. Everybody will report their goals in terms of the context in which the question is asked.(20, 22) In this study, we approached the measurement of goals from several directions in order to more fully understand the mental representation of clinicians regarding computerized documentation. We focused this study mostly on inpatient physicians, but our future plans will extend to other disciplines and settings.

**Methods**

The study was conducted in two phases. In phase I we conducted cognitive task interviews in order to initially explore users’ perception of the documentation process. In phase II, we engaged in observations of clinicians across diverse clinical settings writing notes and completing documentation (with very little interruption) in order to capture the context of documentation and workflow patterns. We also conducted short post-observation interviews.

**Setting and Sample**

The University of Utah Institutional Review Board approved this study. Data collection was conducted at a large, >500 bed academic hospital located in Salt Lake City, Utah. Participants were physician interns, residents, attendings, hospitalists, and consultants, and were selected by convenience sampling methods. Sample size for each phase is listed in Table 1. Phase I was conducted in May 2014, phase II/observation interview participant data was collected from October 2014-February 2015. All phases included participants across a diversity of inpatient clinical domains (Table 1).

<table>
<thead>
<tr>
<th>Table 1. Participant Sample Across 2 Phases</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Phase 1: Cognitive Task Interviews (n=18)</strong></td>
</tr>
<tr>
<td><strong>Physician level</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td><strong>Clinical service/ dept.</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>
Procedures

Phase I. Phase I of this study consisted of cognitive task interviews conducted with participants during the week prior to the implementation of a new Epic EMR system replacing a previously used Cerner EMR. A list of potential participants were selected by first identifying the desired clinical domains, then creating a list of current residents in each domain as well as at least 1 attending and/or hospitalists. Potential participants were approached in their clinical units and asked to participate in a brief, recorded interview (average 15”). Questions were open-ended (and loosely based on the procedures for eliciting a Critical Incident(23): 1) Please recall a typical recent documentation episode in last 24 hours; 2) Describe the context (when, where, type of note, how did you start/finish, how do you fit it into workflow); 3) What was your main purpose in writing the note (e.g. what were you trying to achieve, definition of success, goals); and 4) What prevents you from reaching those goals/purposes. These questions were adaptations from the Action Identification methods of Vallacher and Wegner (1984),(20) whose approach to identifying action identities starts with an initial question “What are you doing”, followed by “You are doing X in order to do what?” followed by “You accomplish Y, but doing what?” (thereby mapping out the whole value to action mental representation). We used similar questions as sub-probes. Prior work suggested that note writing included cognitive activities of synthesizing and organizing their own thoughts, communicating to other team members in a way that they would understand, creating summaries of the hospital progress, and following rules for billing. Three investigators independently coded the transcribed data looking explicitly for statements related to goals, purpose, and action statements. Because goals and task descriptions are often imbedded, assumed, or nuanced, we took time to look for emergent concepts in vivo and through consensus in an iterative process. We used @Atlas.ti qualitative analysis software for coding, then organized into a Microsoft Excel spreadsheet for discussion.

Phase II. Phase II consisted of observations and very short interviews with participants in the clinical setting during targeted time periods of high rates of documentation activities. Data collection forms (using I pads) were iteratively pilot tested and refined using the goals and action tasks derived from the codes identified in phase 1. Six investigators approached potential participants in their clinical units and explained the purpose and procedures of the study. We identified important times, so that about 50% of the times sampled were in the morning, before and after rounds, afternoons in order to capture handovers, and early evening in order to capture admission from the Emergency Room. For every time period, we entered a physician group area and asked if someone would be interested in volunteering. Always somebody did, and anybody we directly asked (such as a hospitalist or outside Attending) never refused. The institution is a teaching organization, meaning that in every clinical setting there are house staff teams, attendings, and fellows. Our goal was to enroll a representative sampling of each role. Occasionally, an outside consultant would appear.

The procedures consisted of entering a setting with several physicians working on computers and ask if anyone would volunteer to be observed writing a note. Nobody refused when asked directly. Those agreeing to participate were observed during a documentation session wherein a participant created one of several types of notes in the Epic EMR (2016) including: rounds preparation, daily progress note, history and physical exam, consultation note, transfer note, or a discharge summary. We measured the time to complete the notes, where else they went in the chart in order to complete the note, their use of templates, copy and paste, and what determined when they could complete the note (turns out to be more complicated in that they would be waiting for more data, waiting to get input from their Attending, or waiting to get more lab results). If they offered their evaluations (such as this is a pain to do in EPIC), we would probe to ask why. The brief interview (< 3 minutes) simply asked their goals, barriers to achieving their goals, perceived efficacy, and any suggestions they had for the Epic system implementation.

Observations and interview responses were captured in written notes by each investigator and entered into a secure REDCap database (@Vanderbuilt, 2015). No audio recording was done. Interview responses were aggregated and extracted from the data set and uploaded into Atlas ti@. Three researchers independently reviewed and then iteratively discussed the selected data and associated codes until consensus was reached in order to identify mid-level action identities regarding the documentation process.

Results

Interviews

We identified 20 actions identities in the phase I cognitive task interviews (Table 2). The goal of supporting ‘billing’ was the most prominent goal identified in Phase 1 (n=17) and in additional 9 in Phae 2 (total = 26). ‘Demonstrating rationale for care provided’ followed as the next most frequently occurring goal for documentation (n=7), with
‘understanding rationale for care provided,’ ‘provider information exchange,’ and ‘educational tool’ following in frequency (n=5 each). The remaining constructs occurred <4 times.

There were a total of 22 action identities that emerged from the interviews in phase II (Table 2). The most frequently occurring was ‘creating and monitoring plans’ (n=15), followed by ‘provider information exchange’ (n=13), ‘summarize patient status’ (n=12), ‘information repository and resource’ as well as ‘to be accurate’ (each n=11), ‘demonstrating rationale for care provided’ (n=10), ‘billing’ and ‘to be comprehensive’ (each n=9), ‘rules & regulations compliance’ (n=8), ‘providing “gist”’ (n=7), and ‘providing clear and informative information’ (n=3). The remaining constructs all occurred with a frequency <4.

Twenty-five unique action identities were identified in total across both initial phases. Specific constructs, their descriptions, and example excerpts are listed below in Table 2.

Table 2. Phase I/II Interview Constructs

<table>
<thead>
<tr>
<th>ACTION Identity</th>
<th>#</th>
<th>Cognitive Task Interview Quotations</th>
<th>CATEGORY</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assist with workflow process</td>
<td>3</td>
<td>Observation</td>
<td>Efficiency and Time Pressure Management</td>
</tr>
<tr>
<td>Comply with billing and support for reimbursement</td>
<td>26</td>
<td>“I know the real purpose is for billing... like the hospital can’t bill for a day in the hospital without some physicians seeing them and documenting what they’ve done.”</td>
<td>Billing Compliance/ Follow Rules</td>
</tr>
<tr>
<td>Create/monitor plans for use across time and providers</td>
<td>15</td>
<td>“See what plan for the day is, especially changes that are going to be happening that day; people that you plan to contact and like specific goals.”</td>
<td>Care Coordination</td>
</tr>
<tr>
<td>Demonstrate rationale for care provided</td>
<td>17</td>
<td>“… the majority of the medical billing is based on critical care time and you need to provide documentation of life-threatening organ failures and you need to document what you’re doing to change those.”</td>
<td>Communicate Thinking for Current and Future Teams</td>
</tr>
<tr>
<td>Document care provided</td>
<td>4</td>
<td>“Like if you’re looking for a certain urine output or blood pressure being maintained, that’s documented...”</td>
<td>Thinking for Current and Future Teams</td>
</tr>
<tr>
<td>Document efficiently (in as little time as possible)</td>
<td>4</td>
<td>“I’ve tried to figure out ways to become more efficient and for a while there I was just kind of making a progress note in a Microsoft like notepad and then pasting it in...I’ve all but completely abandoned importing stuff from elsewhere in the medical records, which saves me a lot of time and I don’t know what belongs on a progress note anyway....”</td>
<td>Efficiency and Time Pressure Assist with thinking and organizing thoughts</td>
</tr>
<tr>
<td>Provide educational tool for residents</td>
<td>5</td>
<td>“My job is to get the residents to see the problem through and guide them if they’re not thinking the problem through in a way that I think is sensible.”</td>
<td>Education Assist with thinking and organizing thoughts</td>
</tr>
<tr>
<td>Organize Information exchange for handovers</td>
<td>2</td>
<td>“…if the night team needs to look at your note they can know what you were thinking for that day.”</td>
<td>Care Coordination</td>
</tr>
<tr>
<td>Create Information repository and resource</td>
<td>11</td>
<td>“…when I go see consults I review history and physical and prior progress notes and consult notes from other services in our current EMR...basically use that information to collect data before I see the patient.”</td>
<td>Assist with thinking and organizing thoughts</td>
</tr>
<tr>
<td>Validate Information</td>
<td>3</td>
<td>Observation</td>
<td>Assist with</td>
</tr>
<tr>
<td>Task Description</td>
<td>Score</td>
<td>Description</td>
<td>Category</td>
</tr>
<tr>
<td>-------------------------------------------------------</td>
<td>-------</td>
<td>------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td>-----------------------------------------------</td>
</tr>
<tr>
<td>Avoid Liability</td>
<td>1</td>
<td>“There’s medical/legal reasons like if a patient were ever to come back and sue us for something, then we’re able to look back and see what we did.”</td>
<td>Billing/ Follow Rules</td>
</tr>
<tr>
<td>Monitor pt. status progression</td>
<td>4</td>
<td>“…to reveal what part of the medical record can help me show what’s new versus old in this patient’s problem.”</td>
<td>Assist with thinking and organizing thoughts</td>
</tr>
<tr>
<td>Be efficient and fast</td>
<td>6</td>
<td>Observation (use of templates and copy and paste tools)</td>
<td>Assist with thinking and organizing thoughts</td>
</tr>
<tr>
<td>Provide &quot;gist&quot;</td>
<td>7</td>
<td>“… if you have to write down everything, you write down what’s important and what like the salient features of the days…”</td>
<td>Care Coordination</td>
</tr>
<tr>
<td>Provide clear and informative information</td>
<td>7</td>
<td>“…the same crappy story often gets perpetuated even if it’s wrong because we sort of look at other prior history…”</td>
<td>Communicate</td>
</tr>
<tr>
<td>Provider information exchange</td>
<td>13</td>
<td>“The notes are extremely helpful for me to remind myself and any other teams members…to just be up to date with the plan and why the patient has the treatments they do or orders in that they do.”</td>
<td>Communicate</td>
</tr>
<tr>
<td>Provide an information resource for patient</td>
<td>3</td>
<td>Observation</td>
<td>Communicate to Patients</td>
</tr>
<tr>
<td>Follow rules/regulations compliance and QI</td>
<td>8</td>
<td>“I think the template model that we use here seems to be targeted towards…documentation requirements rather than necessarily communication…”</td>
<td>Billing/ Follow rules</td>
</tr>
<tr>
<td>Summarize patient status</td>
<td>12</td>
<td>“I would simply put -- try to capture the events of the last day and how the patient’s doing and their physical exam for the day and any significant findings and what the plan is going forward.”</td>
<td>Care Coordination</td>
</tr>
<tr>
<td>Support cognitive processing</td>
<td>4</td>
<td>“…we formulate a plan of what we’re going to do and then discuss it and then compare/contrast our plan to what’s been done already and then proceed with whatever our plan is and then the results of our plan come back. We regroup, reformulate, rethink, reread and redo.”</td>
<td>Assist with thinking and organizing thoughts</td>
</tr>
<tr>
<td>Tell the patient’s story</td>
<td>6</td>
<td>“… I don’t just write, ‘I saw the patient and I agree with the note.’… I try to tell a story more in my note… I try to really just do mine as a paragraph… summation of sort… of why they’re”</td>
<td>Thinking for Current and Future Teams</td>
</tr>
</tbody>
</table>
here and what we’re doing about these things versus sort of a garbled sort of list of things.”

<table>
<thead>
<tr>
<th>To be accurate</th>
<th>11</th>
<th>“It’s reviewing the health staff’s note, making sure that what it says is what they actually said on rounds, making sure that the information is accurate…” Communication</th>
</tr>
</thead>
<tbody>
<tr>
<td>To be comprehensive</td>
<td>9</td>
<td>“…we do an assessment note that usually goes in-depth with the patient’s problems and the active plan for the day…it’s important and I try to get those on the chart the same day because if I don’t actually summarize what the health staff don’t say, then the consult team and stuff won’t know what’s going on.” Thinking for Current and Future Teams</td>
</tr>
<tr>
<td>Understand rationale for care provided</td>
<td>5</td>
<td>“Mostly to see the thinking pattern of the doctors who have seen the patient already to see if I understand where the plan has gone so far.” Care Coordination</td>
</tr>
<tr>
<td>Update patient's current condition</td>
<td>4</td>
<td>Observation Care Coordination</td>
</tr>
</tbody>
</table>

**Observations**

During our observations, we observed 3 patterns of note writing: 1) preparing notes (sometimes began and left to be completed, or used for rounds); 2) writing daily progress notes (sometimes attached to rounds) containing a detailed log; and 3) brief, 24-hour summaries of what happened in the day. Each clinical setting varied significantly in note length and content.

**Suggestions to Improve Design**

In phase II, participants were asked if they had any suggestions to improve the documentation process. The participants provided useful information regarding data display, customizability, searching for information, and training. Suggestions for data display improvements were highly concentrated around “note bloat,” with the great deal of excessive information in the display being the greatest source of frustration, inefficiency, and potential risk to patient safety. Many features of the EHR could be greatly improved by limiting the amount of information displayed and eliminating redundancies. A common documentation goal was to write concise, informative notes that could communicate easily the “gist” of the patient’s situation. (Communicate to Team or Assist with Thinking). For example, several residents talked about the need for a better discharge summary process so that they could easily see the course of the hospitalization. Because each team has unique clinical information needs and diverse role responsibilities, more customizability is needed across specialties and providers. Apparently, communication to nurses and other allied health is subsumed in the documentation process as well as part of workflow, but not supported in current functionality (Care Coordination).

Participants also suggested that searching and tracking of information could be improved. All residents had to go to other parts of the chart while writing notes. (Assist with Thinking) Sophisticated search engines within the note writing environment would be useful and are now widely used across information technology industries. Finally, the easiest way to improve existing design was the simple request to provide additional training. Many participants knew that some of the features they wanted already exist within the system; however, the early training they received did not adequately prepare them for the optimal level of mastery. Several participants asked for brief “YouTube” type videos that addressed certain topics that they could search for as needed, or having periodic, on-unit re-trainings at 6-month intervals.

**Discussion**

Identification of clinician’s documentation goals provided insight into current usage patterns of electronic documentation. As experience with EMRs has increased, users have impacted the design of the technology and the technology has impacted the users. This co-evolution is continuing and demonstrates the transformative nature of health information technology. Assessing this change is challenging and this work helps put forward an approach that could be used over time to clarify the change.

**Measuring Satisfaction**
This work can support a conceptual model of documentation satisfaction that can be used across systems and time. Because the constructs were derived empirically from clinicians who have been working and using two EMRs for many years, the results capture the current transformational changes in workflow created by electronic documentation. Developing a model that can be generalizable across systems as well as across time will allow for future development impact to be measured effectively.

**Clinical Implications**

Our study illustrates how electronic documentation serves many clinical processes and is at the core of the medical record. Providers must be able to successfully write notes and use documentation as a communication tool as they provide patient care and fulfill institutional responsibilities. It is the area where clinicians spend most of their time. EHR interfaces that are designed with complicated, confusing, or superfluous data display and/or entry mechanisms may interfere with workflow by at best frustrating and dissatisfying providers, or at worst by hindering patient care.(24) Specific areas where documentation processes were impactful were handoffs, rounds and transitions. The tension between structured data and narrative data is most visible in regards to billing requirements.(18) Clinicians were very aware of the need to support billing requirements, but often expressed an attitude that it was more an institutional requirement than a clinical one. They also requested more decision support in meeting those billing requirements. This has additional implications for physician workflow, as the link between clinically relevant data representation and billing requirements is confusing in many EHRs.(25). To support the goals of communication and planning, EHRs should provide explicit, standardized terminology and data elements in order to optimize and support clinical workflow.(26)

As billing was the most prominent issue for providers in phase I of the study, the data suggest that streamlining the documentation processes in a way that makes billing easier while concurrently making documentation of care more seamless would be beneficial. A space for rationale of care and/or communication space outside of an actual note might be beneficial too. More intuitive clinical decision support (CDS) with best practice alerts (BPAs) may also prove effective in guiding documentation.

**Design Implications**

The health care domain provides one of the best use cases for user-centered design (UCD). While billing, administration, and secondary use of patient information are important, electronic documentation in this high demand/high risk sphere must be responsive to user needs so they can spend more time providing care and less time in front of a computer. These goals might be mutually exclusive and it might be time to rethink the medical record design. Information display literature focuses largely on what content is displayed rather than how it appears, however UCD insists on the opposite. Improving data display by reducing “note bloat,” adding customizability features, organizing information into clinical patterns and improve information searching would workflow. A streaming feed of most recent orders could appear alongside a patient's chart, which includes labs and vitals automatically uploaded from instruments that would serve as communication and information and have hyperlinks to longer reports in the patient chart.

In addition, providers need to have more than one kind of note and more than one way to display notes. There is a wide range of preferences for how providers prefer to do their documentation, also varying by service, which must be considered. As not all services are the same, providers cannot and should not be expected to document the same way. For example, in psychiatry, interactions with patients are more conversation and observation-based, which lends itself more to text narratives than drop-down lists or smart phrases. Providing a note writing area for thoughts, being able to tag information, or run summary reports with exactly the information a user wants (and nothing they don’t) from that contains the whole story of the patient, or just the “gist,” or choose tagged info with plans or assessments, etc. would greatly improve the safety, efficiency, and satisfaction of physicians. The variation across clinical settings suggests that documentation is evolving differently in different domains. To support this diversity, system designers must create more documentation options.

**Limitations**

This study was done at one site only, with one system only, and using volunteer and available participants. In addition, the focus was narrow, with observations of only physicians in the inpatient setting with an emphasis on internal medicine areas. The methods were qualitative, not experimental. However, every effort was made to get as close as possible to the clinical work process. Collecting Phase 1 interviews during a transition may be a limitation.
Conclusion

In this study, we evaluated workflow and documentation goals across a spectrum of clinical services and provider levels. The current formulation and mechanisms for documentation may not be able to meet all those goals. A transformation in the way progress notes are composed and displayed might be essential. Separating out functionality to support expressivity in documentation (telling the patient’s story), complying with billing regulations, communicating to other providers, and support for higher-level thinking may be necessary.

Acknowledgements

This study was funded by the University of Utah Health Science Center to Charlene R. Weir (PI)

References

Interactive Visualization and Exploration of Patient Progression in a Hospital Setting

Wathsala Widanagamaachchi\textsuperscript{1,2}, Yarden Livnat\textsuperscript{1}, Peer-Timo Bremer\textsuperscript{1,3}, Scott Duvall\textsuperscript{2}, Valerio Pascucci\textsuperscript{1}

\textsuperscript{1}SCI Institute, University of Utah, Salt Lake City, UT, USA; \textsuperscript{2}Internal Medicine, University of Utah, Salt Lake City, UT, USA; \textsuperscript{3}Lawrence Livermore National Laboratory, Livermore, CA, USA.

As medical organizations increasingly adopt the use of electronic health records (EHRs), large volumes of clinical data are being captured on a daily basis. These data provide comprehensive information about patients and have the potential to improve a wide range of application domains in healthcare. Physicians and clinical researchers are interested in finding effective ways to understand this abundance of data. Use of visual analytics to explore healthcare data is one such research direction. Here, we present a visualization and analysis environment to understand patient progression over time. Through the use of optimized data structures and progressive visualization techniques, we allow users to interactively explore how patients and their progression change over time. Compared to existing techniques, our work provides additional flexibility in analyzing patient data and has the potential to be used in a real-time hospital setting. Finally, we demonstrate the utility of our approach using a publicly available intensive care unit (ICU) database.

1 Introduction

The US healthcare system is producing hundreds of thousands of patient records detailing a wide range of information from admission times and dates, to symptoms and outcomes. Until recently, this data has been difficult to access, especially in bulk, often lacked a useful organization, and thus has been generally underutilized for clinical research. With the increasing use of EHRs, this paradigm changes, allowing researchers easy access to a large collection of information. If used effectively, this data may lead to better predictions of patient outcomes, personalized medication, and more targeted interventions. However, to realize this potential requires the ability to understand the clinical data in detail. Given the massive amounts of available data, for example, ICUs may collect real-time data streams of all patients\textsuperscript{1}, which implies automatic or semiautomatic techniques to identify and explore interesting patterns and underlying trends. In this context, visualizing and exploring patient progression over time can provide valuable insights and facilitate the decision-making of physicians and clinical researchers.

Several factors need to be taken into consideration when analyzing this type of data: First, given the large number of patients, an individual, per-patient analysis is time-consuming and does not lend itself to finding commonalities and trends. Instead, patients should be grouped according to various criteria, such as symptoms, outcomes, etc. Second, to compare groups of patients who arrive at different times, their records must be aligned, for example, by their time of admission, time of major procedures or other common factors. Third, patient progression over time needs to be presented in a concise manner to allow simultaneous exploration of large numbers of patients. Finally, to utilize such a system in a hospital setting, the analysis must be interactive, allowing users to quickly explore different hypotheses.

The ideal system described above presents a number of practical challenges, especially for the large databases of interest. First, there exist a number of potentially interesting metrics by which to group patients and thus any analysis must be flexible and efficient enough to change the metric on-the-fly. Furthermore, whereas some metrics are easy to apply and absolute (e.g., splits by gender), others depend on specifying a similarity threshold that determines when two patients are considered to be in the same group. However, in practice this threshold is typically not known a priori, and in fact understanding how patient distributions and progression change with different thresholds may provide important insights. Most existing approaches focus on a single metric and a preselected threshold\textsuperscript{2,3}; we present a system that allows users to freely explore patient grouping metrics and thresholds in an interactive setting.

Another challenge is the size and complexity of the data. Given a large number of patients and high temporal resolution, it is often difficult to grasp the progression of certain groups, let alone identify salient ones. Therefore, presenting data in a concise manner and providing support for various parameter selections and simplifications is crucial to provide the necessary insights.
From an analysis perspective, providing an effective exploration of patient progression requires three abilities: first, grouping patients within a time step at different similarity thresholds; second, correlating patient groups over time; and third, interactively visualizing and exploring patient progression to understand how different similarities affect their behavior. In this paper, we extract patient groups across multiple patient similarities and explore their progression with the aid of tracking graphs, where a concise representation of feature evolution is captured as a collection of feature tracks, see Figure 1. We provide clinical researchers with a visualization and analysis environment that is developed based on our earlier research in the scientific domain\textsuperscript{4,5}. This prior system couples feature grouping and correlation components with visualization techniques to explore the temporal evolution of features in combustion data sets.

Our system process data in several steps. First, we use the patient similarity metric introduced by Lee et al.\textsuperscript{6} to group patients across multiple similarity thresholds. Here, it is important to note that any other patient similarity metric can be used within this system as well. Second, patient groups are correlated over time by tracking the individual patients within a group. In order to allow interactive extraction of data, our system uses optimized data structures to store these patient group and correlation details. Within the system, tracking graphs are used to present a global concise view of patient progression, and progressive visualization techniques are employed to enable interactive exploration of data. Finally, in collaboration with clinical researchers, we apply our visualization and analysis environment to a publicly available ICU database, the clinical database of Multiparameter Intelligent Monitoring in Intensive Care (MIMIC II) databases\textsuperscript{7}, and explore the temporal progression of patients for varying similarity thresholds.

2 Related Work

A subset of the relevant related work is presented here to provide context and background for our research work. Analyzing time-varying data sets usually involves feature extraction and tracking steps. For healthcare data, tracking the progression of patient groups, i.e., the features-of-interest, is relevant to clinical researchers. Among the many feature definitions and their computation techniques found in the literature, techniques that extract feature information for all or a large range of values in a single pass are particularly useful. These techniques often result in hierarchical representations. For instance, hierarchical clustering\textsuperscript{8,9} and various other topological techniques\textsuperscript{10–12} have been used to effectively capture flexible feature hierarchies.

Hierarchical clustering is considered to be one of the most popular methods for creating a feature hierarchy. It partitions data into homogeneous groups based on a measure of similarity through the use of clustering. Depending on the similarity measure used, the results can lead to very different hierarchies. Moreover, many sequential and parallel algorithms for hierarchical clustering are available in the literature\textsuperscript{8,9}. This type of clustering imposes a hierarchical structure on the underlying data irrespective of whether such a structure is appropriate. However, due to its simplicity, many applications have used this method to explore the clustering hierarchy of features. In this work, we also make use of hierarchical clustering to group patients within a time step at different similarity thresholds. In topological analysis, techniques exist that are able to efficiently extract and encode entire feature families in a single analysis pass. Reeb graphs\textsuperscript{13}, contour trees\textsuperscript{10}, merge trees\textsuperscript{12}, and Morse-Smale complexes\textsuperscript{11} are several such techniques. Among them, Reeb graph, contour tree, and merge tree are contour-based and the Morse-Smale complex is gradient-based. As a result, the Morse-Smale complex captures very different structural information.

Visualizing the temporal evolution of features has long been a problem of interest within the visualization community. Depending on the subject area, many different techniques have been developed to address this problem. Traditionally, abstraction, illustration, morphing or animation-based techniques have been used to visualize temporal evolution of features\textsuperscript{14,15}. Tracking graphs that show the feature evolution as a collection of feature tracks that split or merge over time are considered to be an effective representation for visualizing feature evolution\textsuperscript{4}. These graphs provide concise...
global views of feature evolution and are more amenable to filtering and simplifications. As clinical researchers are particularly interested in concise representations, we make use of tracking graphs to visualize patient progression over time.

Rind et al.\(^{16}\) present a comprehensive survey of information visualization systems used to visualize, explore, and query EHRs. These approaches related to EHRs can be broadly categorized into two categories: those that focus on a single patient record\(^ {17}\) and those concerned with a collection of patient records\(^ {2}\). Approaches in the first category focus on providing comprehensive information about a single patient (e.g., patient history, significant events, medication, and treatment), and the second category aims at presenting an overview from multiple patients. The latter provides less detail on each individual patient and focuses more on recognizing patterns and outliers within patient groups. Among these approaches that fall in the second category, LifeFlow\(^ {3}\) and OutFlow\(^ {2}\) are particularly interesting as they visualize event sequences in EHRs. LifeFlow uses color for a compact view and OutFlow uses a graph-based representation. In contrast, we do not visualize the progression of patient groups as an event. At a particular time step, the current event of a patient is one of the parameters considered within the patient similarity metric used. Also, within our system any similarity metric can be used to define patient similarities, providing more flexibility.

3 System Components

An interactive visualization and analysis environment is essential to gain an in-depth understanding of patient progression. In this paper, we refine a prior system that relies on dynamically constructed tracking graphs to enable feature extraction, tracking, and simplification\(^ {4,5}\). This system is designed to study general time-varying features. However, so far it has only been applied to analyze features in scientific simulations. Here, we extend its functionality to effectively visualize patient progression in healthcare data. This section describes our system partitioned into several subsections dealing with: patient grouping, patient correlation, visualization, exploration, and implementation.

3.1 Grouping Patients Within a Time Step

The first step towards understanding clinical data is defining its features and a time step size based on which subsequent analysis is to be conducted. For our intended research, the feature-of-interest is a patient group (i.e., similar set of patients), and a day is considered to be the appropriate time step size. Next, for each time step in the data set, these patient groups need to be extracted and aligned. In this work, to ensure all patients’ hospital stays start at the same time, we align data based on a patient’s admission time.

Once features are extracted and aligned, they should be grouped based on an appropriate grouping method. By maintaining a notion of scale, this feature grouping naturally approximates a meaningful hierarchy. The naive approach of creating this hierarchy is by exhaustively precomputing all possible features at all possible scales. Many popular grouping algorithms also produce nested sets of features for varying scale, which in turn create feature hierarchies (e.g., hierarchical clustering techniques progressively merge elements\(^ {8}\) and threshold-based segmentation creates increasingly larger regions\(^ {11}\)). In this case, we use hierarchical clustering. Patient groups in each time step are clustered based on their similarity to generate a hierarchical representation in the form of a tree. During clustering, we use the similarity metric by Lee et al.\(^ {6}\) to define patient similarities but any other similarity metric could be used as well.

Figure 2(a) shows an example where such a hierarchy is constructed by progressively merging individual patients, with the most similar ones clustered first. Each leaf in the hierarchy represents a patient and each branch a patient group. Along with the hierarchy, various patient group-based attributes such as patient count, mean age and mean heart rate are computed and stored on a per-branch basis. For a given data set, an offline preprocessing step is used to compute these patient hierarchies, and the results are stored in a look-up structure to allow interactive exploration of patient groups. Within this look-up structure, for each patient group, its parent details and patient group-based attributes are stored. To determine correspondences across patients later, each patient is marked with a unique ID. This hierarchy is computed for each time step in the data, and is stored in a separate file to allow interactive exploration of patient groups. Once the hierarchy is computed, patient groups and their attributes can be quickly and easily extracted for any similarity threshold within its range, see Figure 2(b). Given a similarity threshold \(s\) within the full range of \(r\), the corresponding patient groups can be extracted by “cutting” the hierarchy at \(s\). This creates a forest of subtrees, where each subtree represents a patient group existing at \(s\).
3.2 Correlating Groups of Patients Over Time

Once patient groups are identified, the next step is to correlate them over time by tracking individual patients within a group. Two patient groups in consecutive time steps are considered to be correlated if they share at least one patient. All such correlations are extracted for each time step. To efficiently store and interactively extract these patient group correlations, we utilize the meta-graph structure of Widanagamaachchi et al.4. Similar to the patient hierarchy, this meta-graph structure is able to encode patient group correlations and their attributes for a range of similarity thresholds.

The meta-graph is generated in two steps. First, per-patient correlations are computed using the patient IDs computed above. For example, two patients in consecutive time steps are considered to be correlated if they have the same ID. As individual patients are represented by leaf branches in the patient hierarchy, this step results in correlations across leaf branches in consecutive time steps. If a correlation exists, we assign an edge with the weight of 1 across the two corresponding leaf branches, \((s_j', s_j'^{t+1}, 1)\). Second, these per-patient correlations are accumulated along the patient hierarchy to compute the per-patient group correlations. At the accumulation time, if a correspondence already exists, we accumulate only the edge weights.

Just as in the patient hierarchy, various correlation-based attributes such as the amount of patient overlap are computed and stored within the meta-graph structure. Again, once the meta-graph is computed, patient group correlations and their attributes can be quickly extracted for any similarity threshold within the full parameter range. For a selected similarity threshold \(s\), first, patient groups existing at \(s\) for each time step in the data set are obtained using the precomputed patient hierarchies. Then, correlations that exist across those extracted patient groups are obtained from the meta-graph structure. Together, these extracted patient groups and their correlations form the tracking graph at \(f\), see Figure 3. This meta-graph structure is also created in an offline preprocessing step and the resulting structure is stored in multiple files (i.e., one file per time step), each containing a set of edges representing its correlations to patient groups in the next time step.

3.3 Visualizing and Exploring Patient Progression

Our system for exploring patient progression over time contains three views: patient grouping view, patient progression view, and patient view, see Figure 4. Within each view, various progressive visualization techniques are employed to achieve interactivity. For instance, data is always presented with respect to a focus time step that is processed first.
Data for the neighboring time steps is then extracted and presented in order of increasing distance. All views designed for only a single time step (i.e., patient grouping view and patient view), use the focus to determine their time step. The parameters such as hierarchy parameters and other filter parameters are coordinated across all views to provide a fully linked analysis environment.

3.3.1 Patient Grouping View

To enable researchers to gain a quick visual understanding of how patients group together for varying similarity thresholds, the patient hierarchy of the focus time step is visualized within this view. As the similarity threshold is changed, active patient groups within the hierarchy are also highlighted. In Figure 4(b), the selected similarity threshold within the hierarchy is displayed in a brown vertical line, and the active patient groups are highlighted in prominent colors.

3.3.2 Patient Progression View

This view visualizes the temporal progression of patients using tracking graphs. Starting from the user-defined focus time step, nodes and edges are iteratively added both forward and backward in time up to the user-defined time window to create the tracking graph, see Figure 4(c). Each node in the graph represents a patient group. A set of nodes in the same x coordinate indicates groups in one time step and edges across them indicate their correlations. For visual clarity, nodes in the focus time step are always displayed in prominent colors. Progressive techniques as in, specifically, a fast initial graph layout and a slower greedy one, are used to visualize these tracking graphs.

3.3.3 Patient View

Several visualization techniques are combined here to present a specialized view of patients. Specifically, we integrate word cloud, textual, geometric embeddings, and geospatial visualizations, see Figure 4(a).

The word cloud visualization is dedicated to providing a quick overview of textual information regarding patients. For a selected patient group, a word cloud is constructed from the patient group-based attributes stored within the patient hierarchy, see Figure 4(d). Here, to obtain more intuitive overviews, the numerical attributes are converted into ranges. This visualization displays high-frequency words using bigger fonts and brighter colors, and others in faded and smaller fonts. As the name suggests, textual visualization presents textual details of patients in their native domain (i.e., as text), see Figure 4(e). For a selected patient group, this component displays its attributes such as hospital admission ID, patient ID, care unit and age.

Regardless of the data type, visualizing geometric embedding reveals interesting details and trends about data. The geometric embedding visualization displays the geometric embedding of patients in either 2D or 3D, see Figure 4(f). As geometric embedding details are not very obvious for the clinical data, for each time step, the GraphViz's ‘neato’
layout algorithm together with patient similarity details is used to compute the 2D embedding of patients. In the geospatial visualization, when relevant information is available, we allow data exploration to be augmented with geospatial visualizations, see Figure 4(g). For instance, if a patient has his physical location details available for each moment in time (both during and/or prior to his hospital stay), this information will be visualized within this view. In addition to visualizing patient geospatial locations, their trajectories can also be displayed to easily identify data trends related to geographic locations.

3.3.4 Interactive Exploration

As tracking graphs can easily become complex and difficult to understand, various simplifications have to be performed on them to successfully understand their underlying trends. Specifically, we enable several simplification options. Through the linked-view interface, researchers are allowed to explore data sets by changing the focus time step and time window. They can select a particular day within a patient’s hospital stay, expand and contract its neighboring days to view progression both forward and backward in time. Within our system, the similarity threshold within the patient hierarchy, correlation amount within the meta-graph and other attribute values available (i.e., patient group-based and correlation-based), can all be explored. We also allow tracking graphs to be filtered by the length of stay of a patient group, which enables small patient stays to be eliminated from the analysis. Valence two and zero nodes of a tracking graph can be hidden to prevent visual clutter, nodes can be scaled based on their size, and progressions of certain patient groups can be highlighted. To help researchers maintain context across systems’ views, we also make use of correlated color maps and allow nodes to be colored using various patient group-based attributes. All these options combined enable researchers to interactively simplify tracking graphs, isolate interesting patient progressions and explore their parameter space.

3.4 Dataflow

Our system is implemented using the ViSUS framework\textsuperscript{19}, which provides the basic building blocks for designing a streaming, asynchronous dataflow. Figure 5 shows the dataflow utilized within our system. The Data Reader module is dedicated to reading data into the system. It checks whether all data required for the current tracking graph has been loaded. If needed, it loads the required data and passes it to the Filter module. This module filters the received patient group and correlation details for the current parameter and attributes values. This filtered information is then simultaneously sent to layout and view modules.

Each of the two layout modules computes a graph layout and sends those layout details to the relevant view modules for rendering. The Hierarchy Graph Layout module computes the initial layout for the tracking graph and sends this information to the Patient Grouping View and Patient Progression View modules. This hierarchy graph layout is computed only once for each time step as the data is read for the first time. The second layout module, Greedy Graph Layout, computes a greedy layout for the tracking graph each time its parameters change and passes them to the Patient Progression View module. This greedy layout minimizes the edge crossings within the tracking graph.

Our dataflow contains three view modules. The first view module, Patient Grouping View, visualizes the patient hierarchy of the focus time step. Once the module receives
the necessary node and hierarchy details from Filter module and the layout details from Hierarchy Graph Layout module, it renders the patient hierarchy. The Patient Progression View module initially renders the tracking graph using the hierarchy graph layout. Then, as the greedy layout becomes available, it is integrated with the current graph. The third view module, Patient View, provides more specific views of patients (geometric embedding, geospatial, word cloud, and textual visualizations). Once this module receives the required data, depending on which visualization mode is selected, the corresponding computations and renderings are triggered. Each time parameters and/or selections are changed, the current processing within the dataflow is interrupted and restarted. However, rendering within the views maintains the current state for visual continuity.

4 Results

We enable clinical researchers to study the progression of patients via interactive exploration of dynamically constructed tracking graphs. The effectiveness of our framework is demonstrated with the use of a publicly available ICU database. The clinical database of Multiparameter Intelligent Monitoring in Intensive Care (MIMIC II) databases\(^7\) contains comprehensive EHR data collected from hospital medical information systems (both patient bedside workstations and hospital archives). This data is obtained from a set of ICUs including medical, surgical, coronary care, and neonatal in a single tertiary teaching hospital in the 2001 to 2008 time period. It includes patient information that falls into various categories such as general, physiological, medications, fluid balance, notes, and reports, see Table 1. The entire database totals to about $\approx 27\,\text{GB}$ and contains information about tens of thousands of ICU patients.

In order to visualize MIMIC II clinical data within our framework, the relevant patient hierarchies and meta-graph structures need to be computed and stored. This is done in an offline preprocessing step. First, for each day in a patient’s hospital stay, patient details available in the database (e.g., admission ID, age, gender, race, ICD-9 code, drug code, hospital stay length, mean heart rate, mean temperature, and max urine output) are extracted, which results in 38291 patient admissions from 32536 patients. These details are then aligned to make sure all admissions fall on the first time step of the resultant data set. The resulting data set after aligning contains 174 time steps (i.e., 174 days).

Patients in each time step are then clustered together using the metric by Lee et al.\(^6\). This patient similarity metric was previously applied to the same MIMIC II clinical database to identify patient similarities within the first day of the ICU stay. In order to apply the metric to our research, the required clinical, administrative, and categorical variables are extracted from the database for each day within a patient’s hospital stay. Next, correlations across patient groups are computed by tracking individual patients within the groups. Once the patient groups and their correlation details are stored in our data format, the total data size is reduced to $\approx 680\,\text{MB}$. By precomputing the patient hierarchies and meta-graph structures and storing them using optimized data structures, we allow interactive exploration of patient progression over time for several gigabytes of data.

Researchers are provided with the flexibility to vary the patient similarity thresholds and explore the entire parameter space interactively. Such interaction provides an understanding of how patients group together for varying similarity thresholds within a particular day in their hospital stay. In a hospital setting, such interaction provides the flexibility to explore patient progression details to better predict patient outcomes for a specific patient. For example, as new patients come in, for each day in their hospital stay, a patient’s outcome for the next day can be predicted based on

<table>
<thead>
<tr>
<th>General</th>
<th>Patient demographics, hospital admissions, discharge dates, room tracking, death dates (in or out of the hospital), ICD-9 codes, unique code for healthcare provider, and type (RN, MD, RT, etc).</th>
</tr>
</thead>
<tbody>
<tr>
<td>Physiological</td>
<td>Hourly vital sign metrics, SAPS, SOFA, ventilator settings, etc.</td>
</tr>
<tr>
<td>Medications</td>
<td>IV meds, provider order entry data, etc.</td>
</tr>
<tr>
<td>Lab Tests</td>
<td>Chemistry, hematology, ABGs, imaging, etc.</td>
</tr>
<tr>
<td>Fluid Balance</td>
<td>Intake (solutions, blood, etc), output (urine, estimated blood loss, etc).</td>
</tr>
<tr>
<td>Notes &amp; Reports</td>
<td>Discharge summary, nursing progress notes, etc; cardiac catheterization, ECG, radiology, and echo reports.</td>
</tr>
</tbody>
</table>

Table 1: An overview of the data categories within MIMIC II clinical database
their similarity to the existing patient groups. Figure 6 shows several examples of patient groups and their progression for 30, 34, 35 and 38 similarity thresholds. As the similarity thresholds decrease, more patients are grouped together, reducing the complexity of the tracking graph. For a specific similarity metric, exploring the full range of similarities enables researchers to gain insights on that metric’s range of values. In our case, upon exploration we realized that for this particular patient similarity metric, the appropriate similarity threshold range is 30-38. Any similarity threshold below or above that range either grouped all patients into one group or divided each patient to be in a separate group.

Figure 6: Effects of varying the similarity threshold to explore the temporal progression of patients. Here, patient groups and a portion of their corresponding tracking graphs are shown at 30, 34, 35, and 38 similarity thresholds. The focus time step of the tracking graphs is indicated with a black arrow, and the nodes are scaled based on the patient group’s size. In each graph, patient progression for 10 time steps both forward and backward in time from the focus time step is displayed.

Our system presents a global concise view of patient progression over time using tracking graphs. The full tracking graph showing the patient progression over time at 36 similarity threshold is displayed in Figure 7. By observing these tracking graphs, specifically feature track length indicating the hospital length of stay of patients, it is clear that although many of the patient stays are less than 90 days (i.e., 3 months), our data set also contains several longer patient stays. Of 32536 patients, we found 6 patients with hospital lengths of stay greater than 90 days.

Figure 7: The entire tracking graph showing the complete patient progression for the MIMIC II clinical database. The graph contains 1110 nodes and 1288 edges for a total of 174 time steps. Here, 36 similarity threshold is used.

More importantly, various simplification options available in our system allow researchers to further simplify the tracking graphs. For example, filtering the tracking graph by correlation amount allows removal of the least frequent patient progression paths from the tracking graph, making frequent patterns more prominent. If an analysis is to be conducted only on longer hospital stays of patients, filtering options available within our system, specifically filtering tracking graphs by the length of a feature track, are useful. Furthermore, various patient group based attributes stored along with the patient hierarchy such as mean age and mean heart rate can also be used to filter patient groups. Figure 8 illustrates several such simplification results.

Additionally, the patient view of our system is useful for obtaining an overview of patient groups. As the user selects a certain patient group, this view displays the details of its patients. The word cloud visualization provides a quick visual overview of the information within a selected patient group, see Figure 4(d). The numerical attributes such as age and hospital length of stay are converted to ranges to obtain more intuitive results. The exact patient details are also presented in the textual visualization within the system, see Figure 4(e).

Due to the generality of its design, this proposed framework provides new avenues for exploring healthcare data. Depending on the domain question we wish to answer, different options can be used within the visual analysis process. Instead of the patient similarity metric used, any other patient similarity based metric (e.g. age based, symptom based) can be used to group patients within time steps. Different patient attributes can be stored along with the patient
hierarchies. For example, say we wish to study whether patients with diabetes stay longer in the ICU than other patients. We can use a patient similarity metric based on diabetes-specific symptoms to group patients for each time step. It then allows the users to explore the length of stay of patient groups where the patients are grouped based on their diabetes-specific symptoms.

5 Conclusion and Future Work

In this work, we present a visualization and analysis environment for understanding patient progression over time. The system’s interactive abilities to explore patient progression for different similarity metrics and for varying similarities are a distinct advantage over existing techniques used in healthcare. Using our system, researchers are able to explore how patients group together and progress over time, identify frequent progression paths, and also refer back to the native space of data for a visual understanding. By combining optimized data structures and progressive visualization techniques, we enable interactive exploration of terabytes size data, which provides the platform to use this type of analysis in a hospital setting. Within this work, an existing patient similarity metric is utilized for defining patient similarities. At each moment in time, patient similarities are computed by looking at a patient’s current clinical, administrative, and categorical information. A better similarity metric would be one that considers both the current information of the patient and the entire history starting from the hospital admission time. In order to obtain better results, we hope to utilize such a similarity metric in the future. In this work, we demonstrate the applicability of our approach using a publicly available ICU database. We are looking into obtaining additional healthcare databases to use within our system, specifically, databases with geospatial information for which the patient view within our system would prove to be more beneficial. Finally, we aspire to use our visualization and analysis environment in a real-time setting to assist the decision-making process of our collaborating physicians and clinical researchers.

References


Residence, Living Situation, and Living Conditions Information Documentation in Clinical Practice

Tamara J. Winden, PhD1,3, Elizabeth S. Chen, PhD4, Yan Wang, PhD1, Elizabeth Lindemann2, Genevieve B. Melton, MD, PhD1,2

1Institute for Health Informatics, 2Dept of Surgery, Univ. of Minnesota, Minneapolis, MN; 3Research Administration, Allina Health, Minneapolis, MN; 4Center for Biomedical Informatics, Brown University, Providence, RI

Abstract

Social determinants of health (SDOH) have an important role in diagnosis, prevention, health outcomes, and quality of life. Currently, SDOH information in electronic health record (EHR) systems is often contained in unstructured text. The objective of this study is to examine an important subset of SDOH documentation for Residence, Living Situation and Living Conditions in an enterprise EHR informed by previous model representations. In addition to two publicly available clinical note sources, notes created by Social Work, Physical Therapy, and Occupational Therapy, along with free text Social Documentation entries were reviewed. Sentences were classified, annotated, and evaluated once mapped to element entities and attributes. Overall, 2,491 total notes yielded 616, 813, and 30 sentences related to Residence, Living Situation, and Living Conditions. This study demonstrated the need for additional elements in the model representation, more representative values and content culminating in a more comprehensive model representation for these key SDOH.

Introduction and Background

Social and individual behavioral factors play an important role in diagnosis, prevention, health outcomes, and quality of life.1, 2 As defined by the World Health Organization, “social determinants of health are the conditions in which people are born, grow, live, work and age”.3 Social determinants of health (SDOH) can cause illness, exacerbate or contribute to chronic illness, and conversely can also improve health.

Previous studies have demonstrated the deleterious effect of behaviors such as alcohol and tobacco use on health outcomes.4-7 Housing has also been relatively well studied, especially the impact of homelessness on various conditions.8-14 Other housing conditions have been correlated with health outcomes. Residential status, specifically housing instability, has been studied in relation to outcomes in certain disease or treatment groups, and found to be a risk for poor outcomes.8-10 It overall appears that there is a complex interconnectedness between poor housing and poor health.15 Costa-Font found that owning a home, or housing equity overrides the effect of income as a determinant of health and (absences) of disability in old age.12 In contrast, permanent supportive housing can addresses homelessness and health disparities.16 This is also some evidence that improving housing can contribute to improved health.17 For example, Jacobs et al. found evidence that specific housing interventions can improve certain health outcomes.18

However, other aspects of social determinants related to Residence, Living Situation, and Living Conditions have not been investigated as thoroughly. For example, the impact of housing type (Residence) such as single family home, assisted living, group living situations, has not been investigated. Studies have also shown that certain housing models can be beneficial to specific patient groups.14, 19 In addition, significant exposures risks have been associated with indoor environments.20, 21 For example, with multi-unit dwellings there is risk of exposure to hazards such as second-hand smoke as smoke can seep into neighboring units resulting in involuntary exposure.22 Knowledge of a patient’s physical living space, type of dwelling, stairs, safety mechanisms, etc. could be of benefit to the clinician or therapist in providing care to that patient and obtaining the proper support in further promoting their wellness.

With whom a patient lives (Living Situation) as well as the conditions (Living Conditions) under which they live also have heath implications. While living with others creates a support network for the patient, housing density increases exposure to communicable diseases, causes stress in adults and poor long-term health in both children and adults.23

The increase in the use of EHRs provides unprecedented opportunity to collect and analyze SDOH information in conjunction with clinical data in secondary use for research and process improvement. SDOH information can be
used in a myriad of ways from health outcomes evaluations to predictive modeling for prevention. The National Academy of Medicine has completed a consensus study on social determinants in the EHR.\textsuperscript{24, 25} The Committee on Recommended Social and Behavioral Domains and Measures for Electronic Health Records has identified domains and measures that capture the social determinants of health to inform the development of recommendations for Stage 3 meaningful use of electronic health records (EHRs). The NAM final report recommended documentation of race and ethnicity, education, financial resource strain, stress, depression, physical activity, tobacco use and exposure, alcohol use, social connections and social isolation, exposure to violence (intimate partner violence), and neighborhood and community compositional characteristics. The final recommendations unfortunately did not include \textit{Residence, Living Situation, or Living Conditions}.

Currently, comprehensive documentation standards for many SDOH do not exist.\textsuperscript{26} As a result, EHR systems and their associated user interfaces are not optimized for the consistent collection of discrete social history information leaving this important information often buried in free text notes or as unstructured text fields in the social history sections of the EHR. Natural language processing techniques are being developed that allow us to extract social history information from the notes and into discrete datasets but primarily only around substance use information which has more developed and robust information models providing a “target” for discrete representation.\textsuperscript{27, 28}

The overall goal of this study is to evaluate documentation of three SDOH topic areas (\textit{Residence, Living Situation, and Living Conditions}) using publically available note sources and Enterprise EHR free text documentation. This work also builds from work of Chen et al. and Melton et al. to model social history from progress notes and public health surveys, including living situation, residence, social support, and occupation.\textsuperscript{29} Other previous work to harmonize interface terminologies, standards, specifications, coding terminologies, vocabularies, documentation guidelines, measures, and surveys provides a model representation of our three topic areas.\textsuperscript{26} This study also further refines the model representation of \textit{Residence, Living Situation, and Living Conditions} informed by additional interdisciplinary EHR system content.

\textbf{Methods}

This study is focused on the three topic areas \textit{Residence, Living Situation, and Living Conditions}. Clinical notes from three sources were evaluated. Individual sentences related to the topic areas were identified and were classified, into one or more of the target topic areas. Lastly, statement-level annotation was performed, with a 10\% secondary review for Kappa, and elements were mapped to generate a model representation (Figure 1).

\textit{Topic areas}: Using definitions developed from previous work, \textit{Residence} describes dwelling types, physical residence, and geographic location and includes safety considerations such as railings or number of floors and steps.\textsuperscript{26} \textit{Living Situation} describes with whom the patient lives such as roommates, family members, multi-resident dwelling as well as how many others they live with. Lastly, \textit{Living Conditions} describes environmental cleanliness and precautions against infection and disease and includes such things as animals, and presence of mold or an unclean living space.

\begin{figure}[h]
\centering
\includegraphics[width=\linewidth]{methods.png}
\caption{Overview of Methods.}
\end{figure}
Data Sources: The data sources utilized for this work were: 1) MTSamples.com (MTS), a publicly accessible clinical note data source; 2) University of Pittsburgh Medical Center (UPMC) NLP Repository with de-identified clinical notes (following execution of a data use agreement for research); and 3) multiple sources of data from the University of Minnesota-affiliated Fairview Health Services (FHS) electronic health record system. Overall, there were 491 history and physical and consult notes from MTS and 200 history and physical and consult notes from the UPMC included. The majority of clinical data for the study otherwise originated from the FHS EHR system through the University of Minnesota Academic Health Center Information Exchange (AHC-IE) data repository. We included only patients who had consented for their medical records to be used in research with inpatient encounters in 2013. For the purposes of this study to also obtain a broader understanding of documentation of this information by non-provider clinicians including several inter-disciplinary fields, we randomly selected 200 random progress notes authored by social workers, 200 progress notes authored by physical therapists, and 200 progress notes authored by occupational therapists, as well as 1,200 social documentation notes. Social Documentation is a portion of the Social History section in the EHR composed of a single free text field that can be documented on by any EHR clinical user.

Annotation Guidelines Development: Guidelines for sentence-level annotation were developed through literature review and included 28 classifications covering most social determinants of health. Related to this work, of those 28, 7 classifications were further analyzed for this study which included: Residence, Residence Exposure, Residence Other, Living Condition, Living Situation, Living Situation Exposure, and Living Situation Other.

Guidelines for statement-level annotations were developed through previous work reviewing existing standards and terminologies. Separate schemas were developed for each of the three topic areas (Tables 1, 2 and 3).

Table 1. Residence Annotation Guidelines Entities, Attributes and Relationships.

<table>
<thead>
<tr>
<th>Entities (27)</th>
<th>Attributes (16)</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Status</td>
<td>• Specificity</td>
</tr>
<tr>
<td>• Subject</td>
<td>o Exact</td>
</tr>
<tr>
<td>o Family member</td>
<td>o InexactQuantitative</td>
</tr>
<tr>
<td>o Side of family</td>
<td>o InexactQualitative</td>
</tr>
<tr>
<td>o Other</td>
<td>o Other</td>
</tr>
<tr>
<td>• Negation</td>
<td>• Location</td>
</tr>
<tr>
<td>• Certainty</td>
<td>o City</td>
</tr>
<tr>
<td>• Temporal</td>
<td>o State</td>
</tr>
<tr>
<td>o Start Date</td>
<td>o County</td>
</tr>
<tr>
<td>o End Date</td>
<td>o Country</td>
</tr>
<tr>
<td>o Start age</td>
<td>o Other</td>
</tr>
<tr>
<td>o End age</td>
<td>• Certainty</td>
</tr>
<tr>
<td>o Duration</td>
<td>o Unknown</td>
</tr>
<tr>
<td>o Duration Since Time Point</td>
<td>o Uncertain</td>
</tr>
<tr>
<td>o Time point</td>
<td>o Certain</td>
</tr>
<tr>
<td>o Time frame</td>
<td>o Other</td>
</tr>
<tr>
<td>o Residence Age</td>
<td>• Side of family</td>
</tr>
<tr>
<td>o Residence Build Time Point</td>
<td>o Paternal</td>
</tr>
<tr>
<td>• Residence Type</td>
<td>o Maternal</td>
</tr>
<tr>
<td>o Residence Subtype (Type Details)</td>
<td>o Both paternal and maternal</td>
</tr>
<tr>
<td>• Residence Name</td>
<td>o Unknown</td>
</tr>
<tr>
<td>• Geographic Location</td>
<td>o Other</td>
</tr>
<tr>
<td>o Location Detail</td>
<td>• Other</td>
</tr>
<tr>
<td>• Residence Detail</td>
<td>• Type Relationships</td>
</tr>
<tr>
<td>o Detail Subtype (Type details)</td>
<td>• Status Relationships</td>
</tr>
<tr>
<td>• Quantity</td>
<td>• Amount Relationships</td>
</tr>
<tr>
<td>• Other</td>
<td>• Negation, Certainty, and Context Relationships</td>
</tr>
</tbody>
</table>
Table 2. Living Situation Annotation Guidelines Entities, Attributes and Relationships.

<table>
<thead>
<tr>
<th>Entities (21)</th>
<th>Attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Status</td>
<td>Specificity</td>
</tr>
<tr>
<td>Subject</td>
<td>o Exact</td>
</tr>
<tr>
<td></td>
<td>o InexactQuantitative</td>
</tr>
<tr>
<td></td>
<td>o InexactQualitative</td>
</tr>
<tr>
<td></td>
<td>o Other</td>
</tr>
<tr>
<td>Negation</td>
<td>o Unknown</td>
</tr>
<tr>
<td>Certainty</td>
<td>o Uncertain</td>
</tr>
<tr>
<td>Temporal</td>
<td>o Certain</td>
</tr>
<tr>
<td></td>
<td>o Other</td>
</tr>
<tr>
<td>Subject</td>
<td>o Paternal</td>
</tr>
<tr>
<td></td>
<td>o Maternal</td>
</tr>
<tr>
<td></td>
<td>o Both paternal and maternal</td>
</tr>
<tr>
<td></td>
<td>o Unknown</td>
</tr>
<tr>
<td></td>
<td>o Other</td>
</tr>
<tr>
<td>Quantity</td>
<td>o Exact</td>
</tr>
<tr>
<td></td>
<td>o InexactQuantitative</td>
</tr>
<tr>
<td></td>
<td>o InexactQualitative</td>
</tr>
<tr>
<td></td>
<td>o Other</td>
</tr>
<tr>
<td>Detail Subtype</td>
<td>o Unknown</td>
</tr>
<tr>
<td></td>
<td>o Certainty</td>
</tr>
<tr>
<td></td>
<td>o Certain</td>
</tr>
<tr>
<td></td>
<td>o Other</td>
</tr>
<tr>
<td>Subject</td>
<td>o Paternal</td>
</tr>
<tr>
<td></td>
<td>o Maternal</td>
</tr>
<tr>
<td></td>
<td>o Both paternal and maternal</td>
</tr>
<tr>
<td></td>
<td>o Unknown</td>
</tr>
<tr>
<td></td>
<td>o Other</td>
</tr>
</tbody>
</table>

Table 3. Living Condition Annotation Guidelines Entities, Attributes and Relationships.

<table>
<thead>
<tr>
<th>Entities (21)</th>
<th>Attributes (13)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Status</td>
<td>Specificity</td>
</tr>
<tr>
<td>Subject</td>
<td>o Exact</td>
</tr>
<tr>
<td></td>
<td>o InexactQuantitative</td>
</tr>
<tr>
<td></td>
<td>o InexactQualitative</td>
</tr>
<tr>
<td></td>
<td>o Other</td>
</tr>
<tr>
<td>Negation</td>
<td>o Unknown</td>
</tr>
<tr>
<td>Certainty</td>
<td>o Uncertain</td>
</tr>
<tr>
<td>Temporal</td>
<td>o Certain</td>
</tr>
<tr>
<td></td>
<td>o Other</td>
</tr>
<tr>
<td>Subject</td>
<td>o Paternal</td>
</tr>
<tr>
<td></td>
<td>o Maternal</td>
</tr>
<tr>
<td></td>
<td>o Both paternal and maternal</td>
</tr>
<tr>
<td></td>
<td>o Unknown</td>
</tr>
<tr>
<td></td>
<td>o Other</td>
</tr>
<tr>
<td>Quantity</td>
<td>o Exact</td>
</tr>
<tr>
<td>Living Conditions Type</td>
<td>o InexactQuantitative</td>
</tr>
<tr>
<td></td>
<td>o InexactQualitative</td>
</tr>
<tr>
<td></td>
<td>o Other</td>
</tr>
<tr>
<td>Side of Family</td>
<td>o Unknown</td>
</tr>
<tr>
<td></td>
<td>o Certain</td>
</tr>
<tr>
<td></td>
<td>o Other</td>
</tr>
<tr>
<td>Side of Family</td>
<td>o Paternal</td>
</tr>
<tr>
<td></td>
<td>o Maternal</td>
</tr>
<tr>
<td></td>
<td>o Both paternal and maternal</td>
</tr>
<tr>
<td></td>
<td>o Unknown</td>
</tr>
<tr>
<td></td>
<td>o Other</td>
</tr>
<tr>
<td>Relationship</td>
<td>o Type Relationships</td>
</tr>
<tr>
<td></td>
<td>o Status Relationships</td>
</tr>
<tr>
<td></td>
<td>o Amount Relationships</td>
</tr>
<tr>
<td></td>
<td>o Negation, Certainty, and Context Relationships to all entities</td>
</tr>
</tbody>
</table>

Sentence-level Annotation

All notes were initially reviewed and sentence-level annotation was performed by a single reviewer using General Architecture for Text Engineering (GATE) to identify and classify sentences related to the three topic areas of Residence, Living Situation, and Living Conditions. Sentences containing information related to more than one of the three topic areas were classified into each appropriate topic for statement-level annotation. For example, the sentence “He lives in <city name> with his mother and step-father” was classified as both Residence (“lives in <city name>”) and Living Situation (“lives...with his mother and step-father). Table 4 shows example statements for each topic area.

Table 4. Example sentences classified into each topic area.

<table>
<thead>
<tr>
<th>Topic Area</th>
<th>Example Sentences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Residence</td>
<td>• Lives in 4 level house</td>
</tr>
<tr>
<td></td>
<td>• Lives in &lt;city, state name&gt;</td>
</tr>
<tr>
<td></td>
<td>• The patient has been residing at &lt;name of facility&gt;</td>
</tr>
</tbody>
</table>
• Home age 10-25 years

<table>
<thead>
<tr>
<th>Living Situation</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Lives at home with her husband and two daughters</td>
</tr>
<tr>
<td></td>
<td>Lives at home with husband and 5 children.</td>
</tr>
<tr>
<td></td>
<td>Lives with his parents and 11 year old sister</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Living Conditions</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>They have city water</td>
</tr>
<tr>
<td></td>
<td>Childs home has well water</td>
</tr>
<tr>
<td></td>
<td>…had some mold behind the stove and refrigerator</td>
</tr>
</tbody>
</table>

Statement-level Annotations

Statement-level annotation was performed on the classified sentences using the brat rapid annotation tool (brat) to identify elements, attributes and relationships. The schema was modified iteratively to accommodate newly found elements and attributes for three sustentative iterations to include subject and temporal entities not previously encountered to ensure a stabilized schema. The original Residence annotation schema was amended to include “Subject”, “Time Point”, and “Time Frame”. The Living Situation schema was amended to include one new element “Current Age” that refers to the age of the persons with whom the patient lives. And the Living Conditions schema was amended to include the element “Living Conditions Detail Subtype” which refers to the subcategory of type.

The statement-level brat annotations was performed by a single reviewer and a subset of 10% of sentences were annotated by a second reviewer to ensure internal consistency and to assess inter-rater reliability. Values sets were compiled from the annotations for each entity found in the data sources as were schema amendments. The model representations from previous work were then amended with additional elements from this analysis to create and enhanced model representation of Residence, Living Situation, and Living Conditions (Tables 6A, B, C).

Results

In total, 2,491 notes were reviewed by two reviewers, resulting in 1,459 sentences classified into the three topic areas of Residence, Living Situation, and Living Conditions. The initial classification analysis resulted in 616 sentences categorized as Residence, 813 sentences categorized as Living Situation, and 30 sentences categorized as Living Conditions (Table 5). MTSamples, FHS Physical Therapy, and FHS Occupational Therapy notes did not have any sentences that could be classified under the Living Conditions topic.

Table 5. Number of notes reviewed and number of sentences classified. (*Total number of sentences are not mutually exclusive)

<table>
<thead>
<tr>
<th>Data Source</th>
<th>Total Notes</th>
<th>Sentences Classified</th>
<th>Sentences Classified</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Residence</td>
<td>Living Situation</td>
</tr>
<tr>
<td>MTSamples</td>
<td>491</td>
<td>36</td>
<td>88</td>
</tr>
<tr>
<td>UPMC</td>
<td>200</td>
<td>42</td>
<td>54</td>
</tr>
<tr>
<td>FHS</td>
<td>1200</td>
<td>296</td>
<td>453</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>88</td>
<td>64</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>98</td>
<td>86</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>56</td>
<td>68</td>
</tr>
<tr>
<td>Total sentences reviewed</td>
<td>2491</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total number of sentences classified*</td>
<td>616</td>
<td>813</td>
<td>30</td>
</tr>
</tbody>
</table>

The statement-level annotation yielded an inter-rater reliability of $K = 0.84\%$ and proportion agreement of 0.98%. Overall the FHS set of notes were the most comprehensive and this had the highest contribution to this work. In totality, the 616 sentences or statements for Residence yielded significant contributions to the overall model. Of these sentences, Status was documented in 60.1% (416) of sentences, Residence Type 51.8% (359), and Geographic Location Detail (i.e., specific city, state country locations) in 38.4% (302) (Table 6A). Temporal elements were present but to a much lower degree.

For Living Situation, a total of 813 sentences were analyzed and, as with Residence, Status was highly prevalent being present 823 times and in total there were 1303 references to Subject other than patient or family member (Table 6B).
For *Living Conditions* in the 30 sentences were annotated, Living Condition Type was the most prevalent entity found with 39 instances with many sentences referencing more than one Living Condition Type per sentence, followed by subject being found 16 times (Table 6C).

**Tables 6A, B, C. Elements, counts, and example values/patterns Residence, Living Situation, and Living Conditions.** For each source n= the total number of unique sentences that we eventually annotated for distinct elements. Percent of unique sentences that contained the element (Total number of instances of that element). Example values and patterns in bold are newly added to the existing model as a result of this work. Bolded Example Values represent items added to the existing model through this study.

<table>
<thead>
<tr>
<th>Table 6A. Residence</th>
<th>Elements</th>
<th>MTS (n=36)</th>
<th>UPMC (n=42)</th>
<th>FHS (n=504)</th>
<th>Total (n=581)</th>
<th>Example Values and Patterns</th>
</tr>
</thead>
<tbody>
<tr>
<td>Status</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>lives in, resides in, homeless</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>&lt;ownership status&gt; lives, live, living, moved, resides, residing, lived, staying, built, buying, came</td>
</tr>
<tr>
<td>Subject</td>
<td>2.9% (1)</td>
<td>-</td>
<td>0.8% (5)</td>
<td>0.9% (6)</td>
<td></td>
<td>mother's, in-laws, friends, daughter and son-in-law, &lt;family member&gt;</td>
</tr>
<tr>
<td>Negation</td>
<td>-</td>
<td>-</td>
<td>0.2% (1)</td>
<td>0.2% (1)</td>
<td></td>
<td>no &lt;residence detail&gt;, don't</td>
</tr>
<tr>
<td>Certainty</td>
<td>2.9% (1)</td>
<td>-</td>
<td>0.4% (2)</td>
<td>0.5% (3)</td>
<td></td>
<td>yes/present, no/absent, unknown, didn't know, apparently</td>
</tr>
<tr>
<td>Quantity</td>
<td>-</td>
<td>-</td>
<td>0.2% (1)</td>
<td>0.2% (1)</td>
<td></td>
<td>&lt;#&gt; steps, &lt;#&gt; floors/levels &lt;#&gt; residence detail, several, &lt;#&gt;</td>
</tr>
<tr>
<td>Temporal</td>
<td>14.3% (6)</td>
<td>7.1% (3)</td>
<td>6.9% (40)</td>
<td>7.4% (49)</td>
<td></td>
<td>currently, prior to hosp, recently, now</td>
</tr>
<tr>
<td>Duration</td>
<td>8.6% (3)</td>
<td>-</td>
<td>1.6% (9)</td>
<td>1.9% (12)</td>
<td></td>
<td>&lt;#&gt; years, &lt;#&gt; months, &lt;#&gt; days, few weeks/years</td>
</tr>
<tr>
<td>Duration Since Time Point</td>
<td>2.9% (1)</td>
<td>-</td>
<td>1.2% (6)</td>
<td>1.2% (7)</td>
<td></td>
<td>Since &lt;year&gt;, end of &lt;month&gt;, after &lt;medical incident&gt;</td>
</tr>
<tr>
<td>End Date</td>
<td>-</td>
<td>-</td>
<td>0.2% (1)</td>
<td>0.2% (1)</td>
<td></td>
<td>&lt;date&gt;</td>
</tr>
<tr>
<td>Residence Age</td>
<td>-</td>
<td>-</td>
<td>1.2% (7)</td>
<td>1.0% (7)</td>
<td></td>
<td>New, newer, 10-25 years, built before 1950</td>
</tr>
<tr>
<td>Residence Build Time Point</td>
<td>-</td>
<td>-</td>
<td>0.6% (3)</td>
<td>0.5% (3)</td>
<td></td>
<td>&lt;date&gt;</td>
</tr>
<tr>
<td>Start Age</td>
<td>-</td>
<td>-</td>
<td>0.2% (1)</td>
<td>0.2% (1)</td>
<td></td>
<td>&lt;age&gt;</td>
</tr>
<tr>
<td>Start Date</td>
<td>-</td>
<td>-</td>
<td>1.2% (6)</td>
<td>1.0% (6)</td>
<td></td>
<td>Summer, &lt;year&gt;, &lt;MM/YYYY&gt;, &lt;DD/MM/YYYY&gt;</td>
</tr>
<tr>
<td>Residence Type</td>
<td>45.7% (17)</td>
<td>76.2% (36)</td>
<td>50.2% (306)</td>
<td>51.8% (359)</td>
<td></td>
<td>house, apartment, nursing home, mobile home, &lt;dwelling type&gt;, home, assisted living, house, townhome, group home, condominium, senior housing</td>
</tr>
<tr>
<td>Residence Subtype</td>
<td>5.7% (2)</td>
<td>9.5% (4)</td>
<td>4.2% (21)</td>
<td>4.6% (27)</td>
<td></td>
<td>Multi-level, level, story, bedroom, floor, split-level</td>
</tr>
<tr>
<td>Residence Detail</td>
<td>5.7% (2)</td>
<td>2.4% (1)</td>
<td>2.2% (12)</td>
<td>2.4% (15)</td>
<td></td>
<td>own/rent safety devices, stairs, appliances, carpeted, independent, living, rear entry, basement</td>
</tr>
<tr>
<td>Residence Name</td>
<td>MTS (n=96)</td>
<td>UPMC (n=53)</td>
<td>FHS (n=665)</td>
<td>TOTAL (n=814)</td>
<td>Example Values and Patterns</td>
<td></td>
</tr>
<tr>
<td>----------------</td>
<td>------------</td>
<td>-------------</td>
<td>-------------</td>
<td>---------------</td>
<td>-----------------------------</td>
<td></td>
</tr>
<tr>
<td>Geographic Location</td>
<td>8.6% (3)</td>
<td>-</td>
<td>2.6% (14)</td>
<td>2.8% (17)</td>
<td>&lt;facility name&gt;, &lt;general geographic location&gt;, campus, locally, nearby, community, up here, there, here</td>
<td></td>
</tr>
<tr>
<td>Location Detail</td>
<td>42.9% (18)</td>
<td>14.3% (7)</td>
<td>40.1% (277)</td>
<td>38.4% (302)</td>
<td>Specific geographic &lt;country&gt;, &lt;state&gt;, &lt;neighborhood&gt;, &lt;zip&gt;, &lt;street address&gt;</td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Elements</th>
<th>MTS (n=96)</th>
<th>UPMC (n=53)</th>
<th>FHS (n=665)</th>
<th>TOTAL (n=814)</th>
<th>Example Values and Patterns</th>
</tr>
</thead>
<tbody>
<tr>
<td>Status</td>
<td>90.6% (100)</td>
<td>96.2% (53)</td>
<td>86.9% (670)</td>
<td>88.0% (823)</td>
<td>lives, lives with, live, living, resides, residing, visiting, lived, moving in, moved, stay, staying</td>
</tr>
<tr>
<td>Subject</td>
<td>26.0% (27)</td>
<td>22.6% (12)</td>
<td>37.3% (293)</td>
<td>35.0% (332)</td>
<td>spouse, parents, mother, father, child, roommate, family, &lt;Family member&gt; alone, child, &lt;name&gt;, boyfriend, significant other, roommate</td>
</tr>
<tr>
<td>Negation</td>
<td>1.0% (1)</td>
<td>-</td>
<td>0.5% (3)</td>
<td>0.5% (4)</td>
<td>no &lt;subject&gt; &lt;living situation detail&gt;, do not, don’t, not, no</td>
</tr>
<tr>
<td>Certainty</td>
<td>2.1% (2)</td>
<td>-</td>
<td>-</td>
<td>0.2% (2)</td>
<td>yes/present, no/absent, unknown, apparently, either</td>
</tr>
<tr>
<td>Quantity</td>
<td>6.3% (7)</td>
<td>3.8% (3)</td>
<td>3.9% (32)</td>
<td>4.2% (42)</td>
<td>&lt;#&gt; subjects &lt;#&gt; in household</td>
</tr>
<tr>
<td>Temporal</td>
<td>-</td>
<td>-</td>
<td>0.2% (1)</td>
<td>0.1% (1)</td>
<td>Every other week</td>
</tr>
<tr>
<td>Duration</td>
<td>1.0% (1)</td>
<td>-</td>
<td>0.3% (2)</td>
<td>0.4% (3)</td>
<td>Few weeks, &lt;###&gt; years</td>
</tr>
<tr>
<td>Duration Since Time Point</td>
<td>-</td>
<td>-</td>
<td>0.3% (2)</td>
<td>0.2% (2)</td>
<td>Approximately, past several years</td>
</tr>
<tr>
<td>End Date</td>
<td>-</td>
<td>-</td>
<td>0.2% (1)</td>
<td>0.1% (1)</td>
<td>&lt;MM/YY&gt;</td>
</tr>
<tr>
<td>Timeframe</td>
<td>7.3% (8)</td>
<td>5.7% (3)</td>
<td>3.5% (28)</td>
<td>4.1% (39)</td>
<td>Currently, previously, prior to hospitalization, recently</td>
</tr>
<tr>
<td>Time Point</td>
<td>2.1% (2)</td>
<td>-</td>
<td>1.1% (8)</td>
<td>1.1% (10)</td>
<td>&lt;MM/DD/YY&gt;, &lt;YYYY&gt;, this week, now,</td>
</tr>
<tr>
<td>Living Situation Detail</td>
<td>4.2% (4)</td>
<td>5.7% (4)</td>
<td>2.6% (20)</td>
<td>2.9% (28)</td>
<td>inadequate, crowded, alone, privacy, together, independently, foster care</td>
</tr>
<tr>
<td>Family Member</td>
<td>65.6% (104)</td>
<td>75.5% (50)</td>
<td>68.7% (817)</td>
<td>68.8% (971)</td>
<td>Wife, brother, Child(ren), dad, daughter, father, husband, mom, mother</td>
</tr>
<tr>
<td>Side of Family</td>
<td>-</td>
<td>-</td>
<td>0.2% (1)</td>
<td>0.1% (1)</td>
<td>maternal</td>
</tr>
<tr>
<td>Current Age</td>
<td>8.3% (13)</td>
<td>-</td>
<td>3.0% (26)</td>
<td>3.4% (39)</td>
<td>&lt;###&gt; year old, age &lt;###&gt; years,</td>
</tr>
<tr>
<td>Other</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

1772
Table 6C. Living Conditions

<table>
<thead>
<tr>
<th>Element</th>
<th>MTS (n=0)</th>
<th>UPMC (n=5)</th>
<th>FHS (n=33)</th>
<th>Total (n=38)</th>
<th>Example Values and Patterns</th>
</tr>
</thead>
<tbody>
<tr>
<td>Status</td>
<td>-</td>
<td>40.0% (2)</td>
<td>12.1% (4)</td>
<td>15.8% (6)</td>
<td>housing contains, lives, no, live, living</td>
</tr>
<tr>
<td>Subject</td>
<td>-</td>
<td>-</td>
<td>24.2% (16)</td>
<td>21.1% (16)</td>
<td>patient, &lt;family member,&gt;</td>
</tr>
<tr>
<td>Negation</td>
<td>-</td>
<td>40.0% (2)</td>
<td>24.2% (10)</td>
<td>26.3% (12)</td>
<td>no &lt;living condition detail&gt;, without</td>
</tr>
<tr>
<td>Certainty</td>
<td>-</td>
<td>20.0% (1)</td>
<td>3.0% (2)</td>
<td>5.3% (3)</td>
<td>yes/present, no/absent, unknown, apparently</td>
</tr>
<tr>
<td>Quantity</td>
<td>-</td>
<td>20.0% (1)</td>
<td>15.2% (8)</td>
<td>15.8% (9)</td>
<td>Excessive animals &lt;#&gt; Living conditions type, good deal</td>
</tr>
<tr>
<td>Temporal</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Living Conditions Detail</td>
<td>-</td>
<td>20.0% (1)</td>
<td>12.1% (6)</td>
<td>13.2% (7)</td>
<td>water damage, home smelled of urine</td>
</tr>
<tr>
<td>Living Conditions Type</td>
<td>-</td>
<td>40.0% (9)</td>
<td>48.5% (30)</td>
<td>47.4% (39)</td>
<td>mold, insects, rodents, animals, water, heat, well water, filtered water, city water, condemned, electricity, excrement, light, water fluorinated</td>
</tr>
<tr>
<td>Living Conditions Type Subtype</td>
<td>-</td>
<td>20.0% (1)</td>
<td>3.0% (1)</td>
<td>5.3% (2)</td>
<td>Modifier of Living Conditions Type, unhygienic, presence of laundry facilities</td>
</tr>
<tr>
<td>Other</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

Discussion

While social determinants of health (SDOH) play an important role in the provision of patient care, they also play an important role in secondary use of data for research and quality improvement. Unfortunately, SDOH information is not well documented as discrete data in the EHR. By leveraging a diverse collection of notes including general social history documentation notes and notes authored by different types of clinical authors our evaluation helps information about the three topic areas of Living Situation, Residence, and Living Conditions. By far, Living Situation was the most common topic found of the three topic areas followed by Residence with much less information around Living Conditions in our dataset. The prevalence of documentation related to Living Situation appears to be an effort by clinicians to indicate the relative amount of potential social support for patients.

Many sentences included elements that crossed topic areas. For example, “She lives in <CITY NAME> with her parents and 2 older sisters.”, which includes Residence and Living Situation. Another example: “Lives at home with mom and two pets.”, crosses all three topic areas. Not surprisingly, physical therapist and social worker authored notes had a higher proportion of sentences related to Residence. The Social Documentation notes, which could have been authored by any clinician type, had the highest number of sentences related to Living Situation and Living Conditions.

Further analysis and mapping of the Residence sentences to the element axes showed persistent use of status, residence type and geographic location. Temporal entities and attributes were less prevalent with most documentation describing current state with some references to past situations. For Living Situation there was again significant presence of status as we as subject specifically family member. Living conditions was much less represented in these data sources and most references were related to type of water available.

The enhanced model representations (Tables 6A, B, C), built upon the foundation of previous standards evaluation work, now represent the analysis of 27 data sources including existing standards, terminologies, guidelines, and measures and Surveys as well as the analysis of 2491 notes. The analysis of the notes added more elements. For
Residence and Living Situation the temporality elements “Duration Since Time Point”, “Time Point”, and “Time Frame” were added as was the element “Residence Type Subtype” and “Living Situation Detail Subtype”. For Living Conditions the temporal elements “Duration Since Time Point” and “Time Point” were added. Overall, the EHR unstructured text significantly contributed to enhance and strengthen the model representation. Value lists are much more complete for each of the elements.

Although we obtained an inter-rater reliability of $K = 0.84$ and proportion agreement of 0.98, we observed some inconsistency between reviewers regarding the difference between Geographic Location as opposed to Location Detail. This inconsistency was resolved and the annotation schema was amended accordingly. Another challenge area was around statements regarding safety concepts. In some cases safety items such as railings and stairs were annotated as Residence Detail and in other they were annotated as Living Conditions Detail. More work will be needed to sort out where these concepts logically fit the best. Lastly, while it is expected that not every sentence would have every element, in examination of these results it was noted that in the Residence annotations the number of sentences that had a “status” documented was lower than expected. After further manual review the issue was traced back to several of the question answer type “Living Arrangements” could be considered a section header and, in past work, section headers have been identified separately from text. For this work these sentences were left as is and will be considered in future work and iterations of the annotation schema.

In summary, this work has demonstrated that the SDOH topic areas of Residence, Living Situation, and Living Conditions are being documented in the EHR within unstructured text, specifically general progress notes, Social Documentation notes, and notes authored by Social Workers, Physical Therapists, and Occupational Therapists. This analysis contributes to overall representation models for these three topic areas. Next steps will include an evaluation of flow sheet documentation related to these three topic areas and further enhancement of the model representation that can be used to extract information from EHR text, to design discrete data collection tools for the EHR, and to contribute to the development of ontology for the social history topics of Residence, Living Situation, and Living Conditions.

Acknowledgements

This work was supported by the National Library of Medicine of the National Institutes of Health (R01LM011364) and University of Minnesota Clinical and Translational Science Award (8UL1TR000114-02). The content is solely the responsibility of the authors and does not represent the official views of the National Institutes of Health.

References

Towards precision informatics of pharmacovigilance: OAE-CTCAE mapping and OAE-based representation and analysis of adverse events in patients treated with cancer drugs

Mei U Wong, BS1, Rebecca Racz, PharmD, ORISE Fellow1,2, Edison Ong, Bioinformatics PhD student1, Yongqun He, DVM, PhD1

1University of Michigan, Ann Arbor, MI, USA; 2Division of Applied Regulatory Science, Office of Clinical Pharmacology, Food and Drug Administration, Silver Spring, MD, USA

Abstract
A critical issue in the usage of cancer drugs is its association with various adverse events (AEs) in some, but not all, patients. The National Cancer Institute (NCI) Common Terminology Criteria for Adverse Events (CTCAE) is a controlled terminology for AE classification and analysis in cancer clinical trials. The Ontology of Adverse Events (OAE) is a community-based ontology in the domain of AEs. In this study, OAE was first updated by including AE severity grading and OAE-CTCAE mapping. An OAE subset containing CTCAE-related terms and their associated OAE terms was generated to facilitate term usage. A use case study based on a published cancer drug clinical trial demonstrates that OAE provides better hierarchical representation, includes semantic relations, and supports automated reasoning. Demonstrated with a single patient analysis, the OAE framework supports precision informatics for representing AEs and related genetic and clinical conditions in individual patients treated with cancer drugs.

Introduction
While cancer drugs play a critical role in treating cancer patients, the adverse events (AEs) associated with cancer drugs can be very detrimental. As a result, the surveillance and analysis of cancer drug AEs are important to improve public health. To support such surveillance and analysis, it is critical to maintain proper terminology and documentation. A foundation of AE reporting and analysis is the controlled terminology of various AEs. MedDRA [1], WHO-ART [2], and the Common Terminology Criteria for Adverse Events (CTCAE) [3] are commonly used AE controlled terminology systems. These controlled terminologies tend to have drawbacks, such as missing term definitions, poorly defined hierarchies, and lack of semantic relations among terms across hierarchies [4]. Ontologies have been proposed and developed to address these drawbacks [4-6].

As a product of the US National Cancer Institute (NCI), CTCAE is a specialized controlled terminology that provides the standardized classification of AEs of drugs used in cancer therapy. CTCAE has substantially evolved since its inception in 1983. Initially published in 2009, CTCAE v4.0 lists 790 AE terms, and each of these AE terms may be associated with a range of severity grades [7]. Most US and UK cancer drug trials encode their observations based on CTCAE.

The Ontology of Adverse Events (OAE) is a community-based open source ontology that logically represents various AEs [6]. OAE represents various AEs based on patient anatomic regions and clinical outcomes, including symptoms, signs, and abnormal processes. In OAE, an AE is defined as a pathological bodily process that occurs after a medical intervention (e.g., drug treatment). By logically linking the medical intervention, patient, patient conditions, and AE outcomes, OAE provides a robust framework that supports systematic analysis of AE case classification and analysis. With over 5,000 terms, OAE has been used in many scenarios [6, 8-10]. Figure 1 shows the general OAE model of an adverse event of a human patient treated with a drug. Specifically, a human patient is administered a drug in order to treat a specific disease (e.g., cancer). The drug has its active chemical ingredient, dose, and it is administered through a specific route (e.g., oral). After an incubation time, the patient will experience a specific adverse event(s) that occur in a specific anatomic location. The patient conditions (e.g., age, gender, and disease) may affect the drug treatment effectiveness as well as AE outcomes. Even for the same type of AE, there exist different severity grades (as detailed later in this paper).

To support ontology interoperability, in addition to its own terms, OAE imports many terms from existing ontologies such as Basic Formal Ontology (BFO) [11], Ontology for Biomedical Investigations (OBI) [12], and Ontology for General Medical Science (OGMS) [13] (Figure 1). For example, OAE imports many terms from the
upper level BFO [11] and aligns with BFO [6]. Both administration and drug AE are classified as BFO: ‘process’. On the other hand, the drug administration is subclass type of OBI: ‘planned process’ (i.e., a process intentionally planned by human), and an AE is a subclass of OGMS: ‘pathological bodily process’ (Figure 1). The disease, drug, quality, and anatomical entity can be represented by Disease Ontology (DOID) [14], Drug Ontology (DrON) [15], Phenotypic Quality Ontology (PATO) [16], UBERON [17], and NCBITaxon [18], respectively (Figure 1). Terms from these ontologies are semantically linked to provide a comprehensive picture of a drug AE and its associated drug, drug administration, anatomical location where the AE occurs, and detailed patient conditions (e.g., gender, age, and health condition).

Figure 1. OAE modeling of drug AE and related terms in a systematic framework. The terms inside boxes are ontology classes, and terms in the middle of arrows are relations (i.e., object properties). The boxes with blue text indicate patient information, and the boxes with red text represent AE-related terms. See more explanation in the text.

In the Figure 1 modeling, the relation ‘preceded by’ is used to link drug administration and drug AE. This relation is simply a temporal relation that indicates the drug AE occurs after the drug administration. The usage of this relation is aligned with the FDA definition that a drug AE does not have to be causal. It means that the drug administration does not have to be the cause of the AE. An AE may be just part of a disease process that started with the patient treated with the drug. For example, a patient who has gastric cancer and is treated with a drug has gastrointestinal bleeding. The bleeding AE may be caused by the gastric cancer or the drug usage. If we do not know the cause, we can use the relation ‘preceded by’. To represent a defined causal relation, OAE uses the relation ‘induced by’, which is a special subclass of the ‘preceded by’ relation [6]. If we know a causal relation between the drug administration and the bleeding, the relation ‘induced by’ can be used.

In addition, OAE and CTCAE may assign term hierarchies differently. For example, In CTCAE, ‘Rash pustular’ (CTCAE code E11545) is asserted as a subclass of ‘Infections and infestations’ in CTCAE (Figure 2A). This assertion is questionable since not every rash pustular is associated with an infection. CTCAE includes three other types of rash where ‘Rash acniform’ and ‘Rash maculo-papular’ are under ‘Skin and subcutaneous tissue disorders’, and ‘Papulopustular rash’ is under ‘Infections and infestations’. However, CTCAE does not have a class called ‘rash’ (Figure 2A). In contrast, OAE has ‘rash AE’ (OAE_0000528) and over ten specific subclasses of ‘rash AE’ (Figure 2B). Such comparisons suggest possible issues with CTCAE classifications and broader coverage of OAE.

Since CTCAE is commonly used in cancer drug AE reporting, we hypothesized that the CTCAE-OAE mapping followed by OAE-based analysis would improve cancer drug AE classification and analysis and support personalized pharmacovigilance investigation. Not all drug-treated patients will experience AEs. For patients under different conditions, the AE patterns may vary greatly, and should be investigated by considering different variables including personal conditions and drug usage. To address this hypothesis, we updated OAE with a systematic CTCAE-OAE mapping strategy and applied the updated OAE to two different use cases to demonstrate how OAE supports better cancer drug AE classifications and personalized drug-AE case representation.
Figure 2. CTCAE and OAE term hierarchy comparison. This comparison uses the Grade 2 pustular rash AE as the example. (A) CTCAE hierarchy. Rash pustular is under ‘Infection and infestation’. (B) OAE hierarchy. OAE includes the term ‘rash AE’ while CTCAE does not have a matching term.

Methods

**OAE ontology visualization and editing.** The Protégé OWL editor ([http://protege.stanford.edu/](http://protege.stanford.edu/)) was used for ontology visualization and editing.

**OAE modeling of AE severity and semi-automatic OAE-CTCAE mapping.** We first generated an OAE design pattern that represents AE severity. Our OAE severity representation strategy is similar to the CTCAE severity grading in terms of severity definitions, but also differs in terms of their classification and usage. The CTCAE source studied in this project primarily comes from the CTCAE 4.0.3, which is the version available in NCBO Biportal ([http://biportal.bioontology.org/ontologies/CTCAE](http://biportal.bioontology.org/ontologies/CTCAE)). To ensure high quality of the OAE-CTCAE mapping, manual curation was applied to annotate individual CTCAE records and map the records to OAE. The annotations, such as, definition, definition source, MedDRA ID, etc., were manually obtained from either CTCAE or online resources, and the mapped results along with the annotated information were initially represented in a pre-defined Excel template. After second review, the Excel data were then automatically converted to an OWL file using the Ontotator software program [19], and merged to OAE. A Java-based program was developed to automatically assign severity axioms to OAE terms to indicate possible grades for specific OAE AE terms.

An important feature of CTCAE is its classification of AE severity. Specifically, the CTCAE v4.0 includes Grades 0 through 5 with unique clinical descriptions of severity for each AE based on the general guideline: Grade 0 – signs and symptoms within normal limits; Grade 1 - Mild AE; Grade 2 - Moderate AE; Grade 3 - Severe AE; Grade 4: life-threatening or disabling AE; and Grade 5 - Death related to AE [7].

Corresponding to the six grades of severity in CTCAE, OAE also has 6 grades of ‘AE severity’. However, OAE and CTCAE implement the severity grades differently. The OAE ‘AE severity’ is a subclass under the term ‘process quality’ (PATO_0001236). The term ‘process quality’, imported from the PATO, represents a quality which inheres in a process. For example, ‘skin ulceration’ includes five grades (Figure 3A).

To represent a specific AE severity, OAE uses two approaches. First, for an AE with a defined severity range, OAE uses an axiom with the relation term ‘has participant quality’ to link the AE and a severity grade of the participant undergoing the AE process. For example, OAE defines an axiom of ‘skin ulceration AE’:

’ve participant quality’ some (‘AE severity G1’ or ‘AE severity G2’ or ‘AE severity G3’ or ‘AE severity G4’ or ‘AE severity G5’)

1778
The second approach in OAE for severity representation is to generate specific subclasses with severity specifically defined. For example, OAE represents 4 different subclasses of ‘skin ulceration AE’ corresponding to the Grade 1-4 of this AE (Figure 3B).

Figure 3. Comparison of AE severity definitions in CTCAE and OAE. (A) CTCAE term ‘Skin ulceration’ and its five grades. (B) OAE term ‘Skin ulceration’, all its child terms, and related axioms. Ontofox was used to extract all children under ‘skin ulceration AE’ and all their related terms and axioms from OAE. The Ontofox output was displayed in Protege OWL editor, and the screenshot was taken from the display. The OAE annotation includes its mapping to CTCAE ontology term ID. The linkage of the OAE term to the AE severity grade is logically defined using the new relation ‘has participant quality’. The ulcer diameter is included in another axiom to specify the Grade 2 severity. OAE does not include ‘Grade 5 skin ulceration AE’ as explained in the text.

Note that we did not generate a class ‘Grade 5 skin ulceration AE’ because Grade 5 always means death. Since our ‘skin ulceration AE’ axiom has shown Grade 5 is possible, we do not need to generate a specific term for Grade 5. However, OAE represents Grade 1-4 individually (Figure 3B) since each of the grades has its specific criterion other than the general grade definitions. For example, according to CTCAE definition, Grade 2 skin ulceration AE shows a combined area of ulcers <1 cm and nonblanchable erythema of intact skin with associated warmth or edema. In align with the CTCAE definition, we represent such a case in OAE with a specific class “Grade 2 skin ulceration AE” (OAE_0002618) (Figure 3B). In addition, OAE includes an ontological axiom to logically represent the ulcer size criterion of “Grade 2 skin ulceration AE”:

’reresults in formation of’ some (ulcer and ‘has diameter in cm’ some decimal[>0, <1])

With the above axiom, when the ulcer diameter is measured, an OAE-based grading system will be able to automatically classify the ulceration with a Grade 2 severity. The CTCAE does not include such an axiom definition, and its usage requires manual interpretation.

Generation, deposition, and query of CTCAE-related OAE subset view (CTCAE-OAEview). Ontofox [20] was used to generate an OAE subset view that contained all CTCAE-matched OAE terms and their associated OAE terms. The CTCAE-OAEview was then deposited to the Ontobee linked ontology server [21] at http://www.ontobee.org/ontology/CTCAE-OAEview. Ontobee SPARQL (http://www.ontobee.org/sparql) was used to query CTCAE-OAEview for addressing various questions.

OAE application use case studies. Two use cases were analyzed. First, a randomized, multicenter, placebo-controlled phase 2 study of a cancer drug clinical trial was used for testing the application of the newly updated OAE in cancer clinical trial AE analysis [22]. This case study demonstrated how OAE could be used to support AE and AE grading hierarchical classification from a cancer clinical trial. The second use case reported a single patient treated with selumetinib [23] and illustrated the application of the OAE framework for a single patient, personalized pharmacovigilance study.
Results

**OAE-CTCAE mapping and annotation results**

Using our semi-automatic annotation methods, we have established in OAE the OAE-CTCAE mapping for all the CTCAE terms. Many CTCAE terms like “Blood and lymphatic system disorders, Other, specify” are not included in OAE because this term is not specific and does not meet the requirement of generating a new ontology class term. According to our OAE modeling design of AE severity as described earlier, only those grade-specific terms with details more than the general grade definitions were mapped and included in OAE as OAE classes. Those grade-specific CTCAE terms that do not contain additional information than the common Grade definitions (e.g., those terms with Grade 5) were not included in OAE. Overall, we added 1,140 grade level terms to OAE.

We also generated a CTCAE-related OAE subset view (CTCAE-OAEview) that includes all CTCAE-related OAE terms and the OAE terms associated with these CTCAE-related terms. The generation of such OAE subset (or “view”) supports easy usage of the CTCAE-mapped OAE system for cancer drug clinical trials and cancer drug AE case reporting and analysis. In addition, we can better study the CTCAE-related OAE subset to clarify the coverage of the CTCAE system. The CTCAE-OAEview has been deposited on Ontobee ([http://www.onotbee.org/ontology/CTCAE-OAEview](http://www.onotbee.org/ontology/CTCAE-OAEview)). CTCAE-OAEview includes 3,267 CTCAE-associated OAE terms. Among these terms, there are 2,169 terms with “OAE_” prefix, 854 terms with “UBERON_” prefix (for anatomic location), and other terms from different ontologies (Figure 4A).

*Figure 4.* CTCAE-OAEview information and its SPARQL query demonstration. (A) CTCAE-OAEview statistics page information on its Ontobee website ([http://www.onotbee.org/ontology/CTCAE-OAEview](http://www.onotbee.org/ontology/CTCAE-OAEview)). (B) SPARQL query of the CTCAE-OAEview using Ontobee SPARQL website ([http://www.onotbee.org/sparql](http://www.onotbee.org/sparql)). This query identified all the terms directly under ‘rash AE’ in CTCAE-OAEview.

**Use case demonstrating the advantages of OAE-based hierarchical classification**

We applied our OAE-CTCAE mapping method to study the AEs reported in a paper published in *The Lancet Oncology* [22]. Since there was no targeted therapy for KRAS-mutant non-small-cell lung cancer (NSCLC), this clinical study tested whether selumetinib plus docetaxel had a good effect on KRAS-mutant NSCLC. KRAS, encoding for the protein K-Ras that regulates cell division, is the most frequently mutated oncogene in NSCLC. Both selumetinib and docetaxel are antineoplastic drugs. Selumetinib is a potent, selective, orally available, ATP-independent inhibitor of mitogen-activated protein kinase kinase 1 and 2 (MEK1/MEK2), two essential mediators in the activation of the RAS/RAF/MEK/ERK pathway [24]. Inhibition of both MEK1 and 2 by selumetinib prevents cellular proliferation in various cancers. Docetaxel binds to and stabilizes tubulin, thereby inhibiting microtubule disassembly and resulting in cell death. This study found that selumetinib plus docetaxel had promising efficacy but was associated with a higher number of AEs compared to docetaxel alone. Overall, 26 AEs associated with two groups of cancer treatments: selumetinib plus docetaxel, and placebo + docetaxel [22].
**Figure 5.** OAE-based classification of adverse events shown in a cancer drug clinical trial. The 26 CTCAE-defined AEs came from Table 8 of the paper [22] were classified using CTCAE or OAE. (A) CTCAE hierarchy of 20 AEs reported in the paper. Note that 6 out of 26 terms in the paper could not be found from CTCAE. (B) Asserted hierarchy of the 26 terms and their associated terms in OAE. (C) Inferred hierarchy using a reasoner in the Protégé OWL editor. See the text for detailed explanation. Note that for (B) and (C), the Ontofox tool (http://ontofox.hgroup.org/) was then used to generate the hierarchical structure of the subset of OAE that contains all the 26 AE terms and the other OAE terms related to these 26 AE terms.

After the OAE-CTCAE mapping, we could use the CTCAE and OAE hierarchical structures to classify the 26 AEs and their top-level AEs (Figure 5). Interestingly, CTCAE4 only includes 20 of the 26 terms reported in the paper (Figure 5A) [22], while OAE includes all these 26 AE terms. In addition, our OAE-based AE analysis system demonstrated that OAE at least has the following advantages compared to the CTCAE alone:

First, OAE provides additional intermediate layers in the AE hierarchy that supports enhanced AE classification. For example, OAE includes an additional layer ‘abnormal defecation AE’ that covers two subclasses - constipation and diarrhea AEs. The combination of these two subclasses in the treatment group makes up to (32% + 73% =) 105% of AE occurrence rates (Figure 5B), suggesting the high susceptibility of this treatment to the abnormal defecation AE. The 105% AE occurrence rate was likely because some patients might have these two AEs simultaneously or at different points in time during the trial. To handle the temporal occurrence of AEs, OAE provides a mechanism to record the time and time period during which an AE occurs (Figure 1) [6].

Second, unlike CTCAE, OAE supports inferred hierarchy generation based on semantic axiom definitions. For example, although OAE does not assert the term ‘hematology investigation result abnormal AE’ as a subclass of ‘hematopoietic system AE’ (Figure 5A), such a subclass relation is inferred using a semantic reasoner (Figure 4C). Note that this term ‘hematology investigation result abnormal AE’ is not an AE term reported in Table 8 of the paper. However, its usage allows the clear classification of this term in OAE.

Third, unlike CTCAE, OAE includes many semantic relations that logically link different AE terms. For example, OAE provides the following axiom: ‘neutrophil count decreased AE’: ‘is evidence of’ some ‘neutropenia AE’. This relation closely links the laboratory count to the neutropenia AE.

In addition, the updated OAE ontology represents different grades of severity. While the contents of defining different grades of severity are the same between OAE and CTCAE, OAE uses its own strategy in representing different grades of severity as detailed earlier in this article. The use case paper summarizes the numbers of cases in Grade 3-4 of 26 different AEs [22]. Such grading results could have been done easily using the OAE grading strategy. As exemplified earlier, one advantage of the OAE grading system is that by measuring some parameter values (e.g., ulcer diameter), OAE will be able to automatically identify the grading level of an AE (e.g., ulceration).
Personal informatics use case demonstrating OAE-based modeling of patient AE after cancer drug treatment

As shown in Figure 1, OAE provides a logical framework that semantically links different variables (e.g., drug features, drug treatment detail, and patient conditions) with the AE outcomes. Each of these variables may affect the final AE outcome. For example, the drug ingredient, the usage of drug, the human quality (e.g., age, gender, concurrent disease, and disease severity) may all affect the outcome of the adverse event.

To show how the OAE-based framework supports practical use case, we used OAE to model a use case of selumetinib treatment on a patient with recurrent low-grade serous ovarian carcinoma (LGSOC) with KRAS mutation [23]. After treatment with selumetinib for more than 7 years, this patient’s tumor has not progressed and the patient has maintained a good general condition without severe toxicities. A next-generation sequencing study showed that her tumor included a G12V mutation in KRAS [23]. Overall, the patient’s LGSOC carcinoma had not progressed, and the patient experienced several mild AEs over the many years of therapy. The most notable AEs were intermittent rashes (grade 2 at the maximum) at anatomically disparate areas including the trunk, chest, and eyelids. She also intermittently experienced grade 1 stomatitis. Figure 6 demonstrates the OAE modeling of the AE process of this LGSOC patient treated with selumetinib.

Figure 6. OAE-based instance level modeling of a drug AE from a selumetinib-treated patient. The boxes with red text represent AE-related terms.

In this use case, the patient’s drug AEs were linked with the patient’s background, health condition, drug treatment details, and treatment efficiency (Figure 6). This comprehensive view under the OAE framework provides a systematic understanding of the patient treatment outcomes and possible variable associations. Note that in this model, the “intermittent” in the Grade 2 intermittent rash AE case is not modelled in detail here. This paper does not provide such detail either. For better representation of “intermittent” rash AE, we will need to specify when the rash occurred and then stopped over the course of drug exposure and over the course of the disease process.

Discussion

The major contributions of this study are multiple. First, OAE was updated to include terms representing different grades of AE severity. While such severity definition references the CTCAE definitions, the two definition strategies and their usage still differ in many aspects. Second, we achieved the task of OAE-CTCAE mapping by primarily using manual curation to obtain high accuracy and also using software tools such as Ontotator to facilitate ontology editing process. Note that it is possible to use a natural language processing (NLP) method to achieve part of the manual curation work. However, given the relatively small number of CTCAE terms, we did not perform such a strategy. Third, an OAE view of the CTCAE-related terms (CTCAE-OAEview) was generated, which provides a simplified OAE version to support CTCAE-related cancer drug AE data analysis. Finally, two use cases were applied to demonstrate how the updated OAE system can support AE hierarchical classification and personalized pharmacovigilance research. Future directions of the OAE-CTCAE mapping project include the completion of an OAE-based automatic AE grading classification system.
Since a large number of cancer drug AEs are represented using CTCAE, the CTCAE-OAE mapping allows easy conversion of the CTCAE-reported AE terms to OAE terms, which can then be utilized to support AE hierarchical classification (Figure 5). Based on the first use case study, we demonstrate the advantages of applying OAE to study CTCAE-annotated cancer clinical trial results. Our case study shows that after OAE-CTCAE mapping, OAE can be used to better support AE classification and analysis for cancer clinical trials. Furthermore, since OAE covers more terms than CTCAE, it is possible to apply OAE to leverage the coverage of AEs in cancer clinical trials. In the future, we will investigate more use cases, and provide solutions on best usage of both CTCAE and OAE to support cancer clinical trials.

A major difference between CTCAE and OAE is that CTCAE is a controlled terminology and OAE is an AE-specific ontology. As a controlled terminology, CTCAE includes only simple and loose ‘is a’ hierarchies. In comparison, as an ontology, OAE not only includes ‘is a’ hierarchies, but it also allows the generation of robust axioms that link different terms under different hierarchies [4]. Therefore, unlike CTCAE, OAE provides a platform to support automatic reasoning. As exemplified by the size-based grading ulcerations, Grade 1 ulceration means that the formed ulcer diameter is greater than 0 cm but less than 1 cm. The inclusion of this knowledge as a logical axiom in OAE makes it possible for computers to utilize this and other axioms to automatically reason the Grade 1 ulceration. A reasoning engine can also be developed based on such an ontological system. In the future, we will generate more axioms and based on them to develop an automatic AE grading classification system and semantic reasoning engine.

Personalized informatics, which includes systematic individual data analysis, is critical to personalized medicine. The second use case provided in this report demonstrates that the OAE modeling offers a logic linkage of different variables of this cancer patient. Such a modeling provides the ability to integrate various types of data, such as general patient information (sex, gender), medical condition (e.g., concurrent disease, and disease severity), drug name, drug dose, treatment route, time of drug treatment, or AE symptom occurrence, into a single comprehensive picture. Such personalized information illustration is different from the traditional method of summarizing results from pooled patient records. When many patients’ records are available, it is possible to conduct such personalized record organization for each patient, and develop new statistical methods to better analyze the relations among variables in different patients. Such a strategy would promote personalized medicine and public health.

Conclusion

The OAE ontology was updated to ontologically represent different AE severity grades and mapped to CTCAE. An OAE view of the CTCAE-related terms (CTCAE-OAEview) was generated to support better usage of the mapped information in OAE. Two use cases were applied to demonstrate that the OAE ontology can support cancer drug AE hierarchical classification and provide a semantic framework for personalized AE information representation and analysis.

Acknowledgements

We appreciate Dr. Darrell Abernethy’s insightful edits and comments on this manuscript. We also appreciate Dr. Sirarat Sarnivijai’s discussion and comments.

Disclaimer

This article reflects the views of the authors and should not be construed to represent FDA’s views or policies.

References

3 http://evs.nci.nih.gov/ftp1/CTCAE/About.html, access on October 27, 2016


Xie, J., Codd, C., Mo, K., and He, Y.: ‘Differential adverse event profiles associated with BCG as a preventive tuberculosis vaccine or therapeutic bladder cancer vaccine identified by comparative ontology-based VAERS and literature meta-analysis’, PloS one, 2016, 11, (10), pp. e0164792

Xie, J., Zhao, L., Zhou, S., and He, Y.: ‘Statistical and ontological analysis of adverse events associated with monovalent and combination vaccines against hepatitis A and B diseases’, Scientific reports, 2016, 6, pp. 34318


https://github.com/OGMS


https://github.com/pato-ontology/pato/


http://obofoundry.org/ontology/nebitaxon.html


1784
Implementation of a Medication Reconciliation Assistive Technology: 
A Qualitative Analysis

Theodore B. Wright, MD 1, 2, Kathleen Adams, MPH 1, Victoria L. Church, RN MS CNS 1, Mimi Ferraro, RN 1, Scott Ragland, BS 1, Anthony Sayers, MS, FAC-PM II 1, Stephanie Tallett, BA 1, Travis Lovejoy, PhD 1, Joan Ash, PhD MLS MS BA 2, Patricia J. Holahan, PhD 3, Blake J. Lesselroth, MD MB 1, 2; 1Veterans Affairs Portland Healthcare System, Portland, OR; 2Oregon Health and Sciences University, Portland OR; 3Stevens Institute of Technology, Hoboken NJ

Abstract

Objective: To aid the implementation of a medication reconciliation process within a hybrid primary-specialty care setting by using qualitative techniques to describe the climate of implementation and provide guidance for future projects. Methods: Guided by McMullen et al’s Rapid Assessment Process1, we performed semi-structured interviews prior to and iteratively throughout the implementation. Interviews were coded and analyzed using grounded theory 2 and cross-examined for validity. Results: We identified five barriers and five facilitators that impacted the implementation. Facilitators identified were process alignment with user values, and motivation and clinical champions fostered by the implementation team rather than the administration. Barriers included a perceived limited capacity for change, diverging priorities, and inconsistencies in process standards and role definitions. Discussion: A more complete, qualitative understanding of existing barriers and facilitators helps to guide critical decisions on the design and implementation of a successful medication reconciliation process.

Introduction

Medication errors are ubiquitous at interfaces in care delivery – including the ambulatory encounter2-4. An estimated one-quarter to one-half of errors are attributed to gaps in information or incomplete medication histories5-11. If undetected, these errors can result in adverse drug events (ADEs) and significant patient harm12, 13. Medication reconciliation (MR) – the process of collecting a medication history, comparing available documentation, and resolving discrepancies12, 13-16. Unfortunately, durable solutions have been notoriously difficult to implement, owing to barriers associated with information technology, organizational culture, and patient health literacy12, 20-26.

A number of promising health information technology (HIT) solutions have been developed to support MR – many of which have extended the capabilities of the Veterans' Affairs (VA) existing information systems12, 23, 24, 26-35. The Automated Patient History Intake Device (APHID) is one such HIT intervention20, 29, 36. Developed at VA Portland Healthcare System (VAPORHCS), APHID provides a standardized way for clinicians to collect and interpret medication information to identify and reconcile discrepancies. In addition to automatically assembling a list of prescriptions from across the enterprise, APHID displays the corresponding medication images in a questionnaire format for the patient and staff to review at the point-of-care. The developers used this approach with the intent to improve patient engagement and medication recall33, 34, 37-40.

It is important to note that APHID – like any decision support system – cannot be a standalone solution. Because it is subsumed under a larger, more complex activity system, implementation of software such as APHID requires a firm understanding of clinic environment, values, and workflow20, 29, 36. The effect of an HIT intervention – particularly one aimed towards improving MR – is heavily dependent upon sociotechnical fit; local culture, business processes, and implementation strategies all mediating successful adoption. Accordingly, accounts of “real-world” implementations are needed to help informaticians and quality improvement specialists identify, label, and manage the significant sociotechnical determinants of change.

To address this gap, we conducted a qualitative study of a MR quality improvement project. Over the span of eight months in an outpatient clinic, an informatics team rolled out a new MR workflow that incorporated the APHID
software. Concurrent to the rollout, we conducted semi-structured interviews with key stakeholders to guide the implementation. By characterizing the sociotechnical context of the clinic and the implementation strategies used by the team, we sought to identify the barriers and enablers to MR software adoption43.

**Background**

Facility leadership charged the informatics department with developing a comprehensive MR solution for ambulatory specialty clinics. This solution had to provide a standardized method for completing and documenting MR at each encounter. Furthermore, the solution needed to generate an updated medication list that could be furnished to the patient upon departure. In response, the informatics team developed and rolled out an “implementation bundle” that included 1) the APHID software, 2) structured note templates (based upon user-furnished requirements), 3) an “after-visit” summary report with medication list, 4) a redesigned workflow, 5) printed educational materials, and 6) at-the-elbow support provided by members of the informatics team.

The informatics team iteratively designed and usability tested software during a series of rapid design “sprints” with representative end-users. The team then introduced the implementation bundle in a phased roll-out, beginning with the nurses, proceeding with the physicians, and concluding with the medical support assistants. The nurses were instructed how to use the APHID software to collect and chart a medication history. Physicians were instructed how to use specially designed templates to review the history and document reconciliation efforts. They were also taught to generate an after-visit summary report. Medical support assistants were expected to print the after-visit summary along with follow-up appointment information at the time of patient checkout.

Despite initial implementation hurdles, the informatics team was able to successfully incorporate the tools and workflow into the clinic. At the time of manuscript preparation, the clinic had been using the MR bundle for four months following implementation. The full details of this implementation along with quantitative quality improvement statistics are described in a separate manuscript (in preparation).

**Theoretical Framework**

While an exhaustive review of implementation theory is beyond the scope of this manuscript, it is useful to briefly describe some of the contemporary theoretical frameworks that informed our approach to this study. Davis’ Technology Acceptance Model (TAM) is an extension of Ajzen and Fishbein’s Theory of Reasoned Action and has been widely used to describe users’ acceptance and adoption of technology in a wide array of disciplines, including healthcare44-46. The model states that an individual’s predilection to use a technology has two determinants: perceived ease of use and perceived usefulness.

Although TAM has been studied extensively and has been said to predict 30-40% of variance in end-user behavior, it is limited in several important ways47. First, TAM was developed outside the healthcare context and it may not suitably capture the complex-adaptive nature of healthcare systems48. Second, TAM only addresses individual difference factors and neglects the social processes associated with implementation such as the voluntariness of HIT use, team-based care interactions, and healthcare governance49, 5047. Finally, the model emphasizes the intent to use or frequency of use, as opposed to the quality or effectiveness of technology use.

To address these limitations, so-called “added variable” models have emerged to describe and explain the interaction between user and setting. Venkatesh developed the Unified Theory of Acceptance and Use of Technology (UTAUT) based on the TAM. The UTAUT is a widely cited model that adds facilitating conditions and has been shown to explain up to 70% of behavioral intention and 50% of actual use51. Our research group developed the Effective Technology Use (ETU) model49, which extends TAM and UTAUT in two important ways. First, ETU includes a parsimonious set of “belief-based” variables that are amenable to manipulation: 1) an internal locus of control (i.e., “Compatibility with Workplace Values”), 2) and two external contextual mediators (i.e., “Compatibility with Work Processes” and “Implementation Climate”). Second, ETU conceptualizes use of technology along a continuum ranging from avoidance to skillful and effective use. We used the ETU to develop semi-structured interview scripts described in the following Methods section. Our research team also used the framework throughout the analysis cycle to organize findings.
Methods

Setting
This study was conducted at a level 1a Veterans’ Affairs facility located in the Pacific Northwest United States from April to October 2016. Nearly 70,000 patients from Alaska, Washington, Idaho, and Oregon receive primary or specialty care from our center; more than 85% of the patients are 45 years of age or older and 91% are male. The informatics department selected the ID clinic as a “demonstration lab” for several reasons. First, the clinic is located in a single physical location, thereby avoiding the need to divide implementation personnel across multiple deployments. Second, ID staff provides primary care and specialty consultation to a medically complex patient population, allowing the implementation team to observe an array of reconciliation challenges. Finally, the ID clinic hosts a small number of physicians – typically, three ID physicians and 1-2 residents or fellows – hence, limiting the training burden shouldered by the implementation team.

Participants
Although we extended interview invitations to all staff members, several were unavailable to participate due to scheduling conflicts or clinical reassignment. Three physicians, two nurse, and five medical support assistants participated in a total of thirteen pre- and mid-implementation interviews.

Interviews
The research team used the ETU model to guide the development of semi-structured interviews scripts. We included questions and probes to address each construct in the model. For example, to gather information about user values we included the question, “how important is it to collect and reconcile a medication history in comparison to other tasks performed during a clinical encounter?” Similarly, to gather information about the climate for implementation we included the question, “how did leadership communicate to you about this implementation?” Questions were open-ended to encourage respondent-directed narratives and to capture emergent concepts.

Interviews were conducted in two discrete phases of the implementation: 1) pre-implementation interviews were completed before technology rollout and 2) mid-implementation interviews were conducted in a rolling fashion as end-users were oriented to the technology. Two research team members were present for each interview. One researcher conducted the majority of the scripted interview while the other acted as scribe, managed the recording equipment, and asked follow-up or clarifying questions.

Each interview guide was informed by the ETU model and began with broad open-ended questions, followed by a series of more specific probes. Specific scripts were developed for each clinical role (i.e., nurse, physician, or clerk). The pre-implementation interviews aimed to gather baseline information about the target users and workplace culture, while mid-implementation interviews focused on gathering information about tools, workflows, and support. The audio recordings were transcribed verbatim. Names were replaced with identification codes to protect individual identities.

Analysis
A team member uninvolved with the implementation completed the initial analysis of transcripts using open coding methods. Codes were applied by labeling each word, line, or passage. Initial codes were then pile-sorted into preliminary categories using a Grounded Theory technique to identify emergent themes. To establish validity, the data were presented to the implementation team for critique and feedback. The investigators then compared sample passages and codes to identify new codes and larger themes. This process was repeated until no new themes could be identified and consensus was reached among team members. Finally, themes were sorted into facilitators and barriers to successful technology use.

Results
After reviewing 132 pages of transcripts, the research team identified a total of five facilitating conditions and five barriers germane to the MR implantation. We outline herewith each facilitator or barrier along with exemplar quotes as well the ETU concept which most closely maps to each.
Facilitators

**Facilitator 1:** The innovation aligned with users’ clinical tasks, values, and priorities; prioritization varied as a function of professional role. (ETU Theme: Compatibility with Workplace Values)

While respondents in each professional role emphasized a different set of core patient care values, each respondent was able to spontaneously draw associations between a core value and one or more facets of the implementation. Medical support assistants emphasized the primacy of fostering patient-centered care:

> “I think that the veteran has had a very positive experience, and they’re very happy to see it, and I think it’s great… I think that it gives them more information, and they have something in their hand” - (MSA1).

Nurses emphasized the importance of delivering accurate and useful information to providers:

> “We want to make sure that we’re not only documenting properly but we’re also informing the provider so that they’re getting that information as soon as possible to address it in the, within the appointment.” - (LPN5)

Physicians emphasized the criticality of prescribing safety and patient communication:

> “I think when we start doing better medication reconciliation the profile will be more accurate… My hope would be that that will make for better patient care.” - (MD4)

> “One thing is that you know the patient has, each time, the right phone numbers and all of that stuff from our clinic, and so they constantly have the right information. I like that.” - (MD5)

**Facilitator 2:** Staff found the software interface easy to use; products were easily incorporated into existing workflows without reducing efficiency. (ETU Theme: Perceived Ease of Use)

Most users believed the software was easy to use and could be adapted to existing workflows:

> “Really, it’s just one more little step than what we usually do. So I think that it’s just very easy. I mean you just click on it, print it out, and hand it to them along with their appointment list. It’s very smooth. No issues” - (MSA1).

Many seemed undeterred by the time required to complete MR:

[Is it worth the time?] – “Yeah, I think so, and honestly, I’m not sure that it takes more time. It took some time to learn the tools, and well, it still is, but I think in the end, I’m not sure that it takes too much more time from us” - (MD5).

**Facilitator 3:** Despite the absence of executive sponsorship, the implementation team successfully used a practice facilitation strategy to establish clinician champions and foster clinic staff engagement. (ETU Theme: Implementation Climate)

Typologies for overarching implementation strategies help broadly categorize and disambiguate between approaches. In general, the implementation team tended to use a “practice facilitation” approach – encouraging clinician uptake by providing intensive support and preparing local champions. This is in contrast to a primarily top-down approach where existing organizational structures such as executive boards or section leaders promote and enforce change directives.

> “I thought the staff [Implementation team] was very helpful. They asked any questions I had or answered all of my questions, and they were very supportive. They asked me if there was anything that I needed or more additional information, and yeah, they were great. They were supportive” - (MSA1).

The implementation team encouraged clinical champions to adopt project ownership…

> “But having the emphasis that medication reconciliation and medication adherence are very, very important for patient outcomes, that’s a very important thing, and then saying we’re providing you with tools to facilitate. You’re doing a good job with medication reconciliation. That’s the other important thing.” – (MD4)
“you have to go with, you have this great opportunity to help the patient, and we’re providing so many tools to make
this so much easier for you and so much more of a streamlined process and as our head infection control nurse always
used to say, the reason bad things happen in medicine is because the holes in Swiss cheese will line up sometimes, and
having two people looking at things from medications from two different perspectives, that prevents these bad things
from happening. So I have to go with the angle of rather than mandating things in a certain way of trying to have tools
to help you help your patients. I think that that is the key” - (MD4).

…despite the absence of executive sponsorship.

“Thinking back to leadership support and stuff, I think it comes back to supporting staffing and supporting you guys to
do it, and supporting the staff to be there to be able to implement it. We just don’t have that. So I find that very
frustrating… Yeah, and I have no idea if. I mean I would assume that the Chief of Medicine would be supporting this. I
don’t actually know because I just haven’t been involved in discussions past MD1” - (MD5).

**Facilitator 4: At-the-elbow support was essential, and more valuable than printed educational materials. (ETU
Theme: Implementation Climate)**

Staff believed at-the-elbow support was a crucial training tool:

“I mean it would have been much more frustrating without having one of you guys there. I think that it was key to
have at least one of you guys sitting in the clinic. Probably, I mean I’m not sure if we needed it for every individual
provider, potentially to save some manpower, but one of you to be in clinic in real time is key” - (MD5).

Notably, users felt supporting staff did not necessarily need to have a clinical background:

[Was there much of a difference with us supporting you with a doctor versus someone else?] – “Not a huge difference;
Yeah, not a huge difference… Yeah, but no, for the technical part of it, it maybe was slightly. I can say if all things
being equal, potentially it would be nice to have a doctor there in kind of he understands that. I mean he really
understands the physician part of the workflow that you’re dealing with, so that’s potentially slightly more helpful. But
it’s almost as helpful to have non-[Physician] there” - (MD5).

Although customized paper educational materials were furnished to each professional role during training, they were
rarely referenced in clinic:

[What educational materials did you use to learn the new tools and processes?] – “It was hands on. You guys just kind
of came up and because I didn’t get to go to the initial walkthrough orientation. So I know that you for sure came out
and kind of explained to me the process…” [Do you remember getting any educational materials from us?] – “I do not
remember getting any educational material from you” – (MSA1).

**Facilitator 5: Providers valued having flexible tools and processes available to accommodate the emergent
properties of complex adaptive systems. (ETU Theme: Compatibility with Work Processes)**

Providers emphasized that the clinical situations encountered – and practice strategies used – are heterogeneous and
often improvised. Therefore, it is critical that tools can be adapted to the circumstances:

“The new note I haven't used at all because I do my own intake notes, and I have a way that I like them... Everybody is
different... But these patients are so complex, and some of them we’ve followed for so long that I have my way that I
like to have things in a specific place and follow them. So it helps me... Sometimes I’ll actually do that note and then
cut and paste it into my other notes (laughter), some of the components of it. So I liked having it there. It’s more. All
of these things to me are tools, but it’s nice to have different tools to be able to use in different circumstances for
different patients. So overall, I think that it’s definitely been helpful” - (MD4).

**Barriers**

**Barrier 1: With severely limited clinical time at baseline, users feared that using new technology would be time
consuming and interfere with care delivery. (ETU Theme: Compatibility with Work Processes)**

Clinicians and staff felt that they had insufficient time to complete even baseline clinical tasks:

“Because the doctors are just as busy if not sometimes more busy than nursing staff. They’re not going to always have
time to get to everything they need to get to in an encounter. In all fairness they’re busy, too” - (LPN2).
“The way the clinic works, there is so much delay I will have them skip vitals and stuff, and have them put them in a room. Or else you’ll inevitably get way behind” - (MD5).

Respondents representing each professional role expressed concerns over the time required to complete MR tasks:

“I think it’s going to be a learning curve a bit. I think it’s going to be time consuming at first. Other than that, I think if all three of us work on it together, the MSA’s, I think it can be successful. My concern is that it’s time consuming” - (MSA1).

Barrier 2: There is no accepted standard for prioritizing clinical tasks; staff tends to prioritize according to their individually-held values. (ETU Theme: Compatibility with Workplace Values)

Our VA facility is a matrix organization; control structures are siloed and clinical personnel tend to report to different managers. As a result, leadership messaging can be unclear or even discordant. We found that in resource-constrained settings, staff used personal and ad-hoc heuristics to organize and prioritize work:

[Regarding clinical tasks] “I don’t know that there really is a prioritization per se, unless they are saying things to trigger the need to look at, ‘I’m depressed or I’m thinking to hurt myself’… at least if we get to a point where we can get something set, where it relieves the guessing game, did we do it for this one, did we not do it for this one? That can slow things down” - (LPN2).

Barrier 3: Staff believed their capacity for change was limited by fixed resource constraints (staffing, environmental space, patient social issues, and medical complexity). (ETU Theme: Implementation Climate)

Staff identified fixed resource constraints such as physical configuration of the clinic (e.g., the number of exam rooms, or distance between exam rooms and waiting area) or medical complexity of patients (e.g., multiple active problems, high medication burden, co-management of patients) as important barriers to change:

“I think it is confusing for them [patients] as well, because they come through those double doors and they don’t know whether to go left, right, straight… I’ve seen a lot of CDU patients come out of the double-doors, and say, ‘where are the elevators, I need to check out.’ They don’t know where to check out” - (MSA3).

Barrier 4: Gaps in codified process standards and unclear role definitions lead to disruptions in workflow. (ETU Theme: Compatibility with Work Processes)

Stakeholders role expectations – including those of the patients – are used to negotiate shared mental models, rules of engagement, and production of deliverables. When these roles are not clear to everyone, workflows are disrupted:

“But the problem we realized at the time of that meeting was that everybody in our group does the process very differently. Has very different goals, and very different… and I was like, ‘this is not going to be easy’” - (MD4).

“I know with (sigh) the medication reconciliation part of it I feel that with patients cause they’re not used to the nurses going through that part … because some of the patients aren’t, either are not comfortable, or it take, just takes a little bit longer to do that, the medication reconciliation with patients that haven’t gone through it with the nurse cause they’re used to doing it with the, a provider” - (LPN5).

Barrier 5: Structured data forms were difficult to use and lacked sufficient expressivity for routine clinical encounters. (ETU Theme: Perceived Ease of Use)

The structured clinical notes requested by the ID staff and developed by the implementation team were found to be over-engineered, not intuitive, and inflexible at the point-of-service:

“I kind of feel bad that we made this whole [clinic] template and created it, and that in practice we’ve found that it really isn’t all that helpful. So I guess I’ve learned or I guess we’ve probably learned that there is some balance of over-scripting and under-scripting these things… I guess the [clinic] template was a little bit surprising. I thought that it would work great, but it didn’t seem to work in the end… I think that it’s over-customized. Yeah” - (MD5).
Discussion

Key Findings

This article is one of the first to use qualitative methods to describe and understand the sociotechnical determinants of MR adoption and, to our knowledge, only the second that explores ambulatory care culture. We identified several important facilitators and barriers to MR adoption that may inform the development and implementation of future MR processes.

Although articulated prioritization heuristics varied as a function of professional role, all users associated new tasks with a deeply held personal value. Medical support assistants typically orchestrated patient navigation activities (e.g., check-in, check-out, scheduling) and, therefore, more often extolled the virtues of our patient-centered documentation. Nurses and physicians, by contrast, valued the completeness of the medication adherence data. When asked why they used the tools, they cited the ability to improve prescribing safety and avoid medication misadventures. Interestingly, nurses also emphasized the significance of team-based care and quality information hand-offs. When end-users believed the innovation purpose aligned with their values, they were more likely to adjust workflows to accommodate the innovation. To summarize, it appears that when process improvement professionals select “practice facilitation” strategies, it is critical to understand the unique values associated with the different professional roles. This information can then be leveraged when recruiting clinical champions and developing information campaigns.

Our findings also suggest that lack of executive sponsorship – while undesirable – does not predict certain failure of an implementation. Although facility executives (i.e., the Chief of Staff Office) sponsored the project, there had been no direct communication with front-line staff regarding participation or expectations. In response, the implementation team adopted a set of “practice facilitation” techniques including 1) identifying and preparing champions, 2) ensuring ongoing training and facilitation, 3) providing audits and feedback, and 4) reminding clinicians of key tasks. In this regard, the implementation team acted as a liaison or proxy for the executives. Respondent reports suggest the implementation team effectively communicated project goals and expectations directly to key stakeholders. Stakeholders, in turn, were encouraged to reach out to the implementation team for questions or issues. Overall, this data supports the contention that an internal locus for change can be cultivated using a number of complementary and synergistic implementation strategies.

Finally, the data suggest that at-the-elbow support was an essential component of implementation success. Printed materials were quickly discarded and all but forgotten, whereas direct hands-on support helped reinforce tasks, navigate trouble spots, and build confidence in the process. This finding is unsurprising; staff training has long been recognized as a relatively weak intervention in process improvement campaigns. What may be more significant is that respondents valued the assistance of non-clinicians as well as clinicians, suggesting that the value of hands-on support is not necessarily limited to clinical colleagues. Future implementations may be able to use more non-clinician staff during training or to otherwise extend training beyond the capabilities of clinician “super-users”.

We identified many implementation barriers – some of which were at least partially addressed with targeted interventions. Fixed external constraints plagued the implementation and caused timeline slippage. Pre-implementation interviews suggested a climate of frustration with systematic barriers facing the clinic, such as clinician time, exam room space, and competing organizational expectations. Although optimism increased by mid-implementation, the research team identified a number of new barriers including unclear role definitions associated with the new workflow and poor note template usability. Fortunately, the structured approach to gathering feedback enabled the team to quickly identify problems and course correct. For example, a 15-minute schedule shift was enacted to provide more nurse-patient contact time. Staff roles were codified and reinforced with at-the-elbow support. Finally, the implementation team invested additional time role-modeling hand-offs to reinforce shared mental models and improve situational awareness.

In general, users indicated the software tools were easy to use, integrating well into existing workflows. The only notable exception appeared to be the initial structured note template. Providers found the input fields to be inefficient and difficult to navigate. Accordingly, the development team quickly abandoned this design for a more flexible MR input dialog that could be called by the user at any point in the workflow. What is exceptional about this finding is that the template was designed according to user-centered principles with requirements furnished and tested by the ID clinicians.
Human factors experts often find that end-users will describe—and build products—according to work-as-imagined rather than work-as-done\textsuperscript{60}. Fortunately, involvement of end-users and domain subject matter experts helped to build trust between implementation team and stakeholders. Furthermore, the collection of qualitative data at mid-implementation helped surface design challenges early and craft an agile response. This finding stands as a reminder that electronic health record documentation tools should balance the need to script workflow and collect computable data with the flexibility required to handle the emergent properties of the clinical environment.

**Implications for Further Work**

In the healthcare setting, technology is becoming nearly ubiquitous. However, advances in technology as well as biomedical knowledge\textsuperscript{61} will continually lead to the development of newer technologies and methods that will have an even greater potential to improve the way physicians care for patients. As such, it becomes ever more important to continue to evaluate and improve our understanding of technology implementation methods. There are many avenues for future research. As patients play a central role in the process, further exploration of the impact that patient values and expectations have on technology acceptance, even if the patient is not the primary user of the technology, would be of benefit for future MR implementations. More research is also needed regarding structured data elements in clinical documentation—How can structured documentation better be implemented without hindering the end user? Finally, in the near future as technologies with alternative input methods such as voice recognition using artificial intelligence as the primary interface mechanism become more available, research analyzing the intricacies of this novelty will be required.

**Limitations**

Limitations of our study include that it was conducted at a single site, limiting the generalizability of findings. Nurses and medical support assistants rotated through clinics at different time points during the implementation. Thus, it was not always possible to recruit the same participants from pre to mid-implementation. Finally, the total number of interviews is relatively low. Hence, this study should be considered exploratory in nature. Larger respondent samples from multiple clinics—or, preferably, multiple organizations—would help to enrich and validate the list of facilitators and barriers.

**Conclusions**

As we increasingly utilize new technologies to improve healthcare processes, more research is needed in how best to implement and integrate these technologies into clinical workflows. The findings in this qualitative analysis of an MR assistive technology implementation identify key sociotechnical aspects of technology adoption. While our findings generally agree with current literature on healthcare technology implementations, some aspects may be more specific to MR processes and these findings may be used to guide future efforts to implement MR processes.

**References**

2. Bradley EH, Curry LA, Devers KJ. Qualitative data analysis for health services research: developing taxonomy, themes, and theory. Health services research. 2007;42(4):1758-72.
Clinical Named Entity Recognition Using Deep Learning Models

Yonghui Wu, PhD, Min Jiang, MS, Jun Xu, PhD, Degui Zhi, PhD, Hua Xu, PhD
School of Biomedical Informatics, the University of Texas Health Science Center at
Houston, Houston, TX, USA

Abstract

Clinical Named Entity Recognition (NER) is a critical natural language processing (NLP) task to extract important concepts (named entities) from clinical narratives. Researchers have extensively investigated machine learning models for clinical NER. Recently, there have been increasing efforts to apply deep learning models to improve the performance of current clinical NER systems. This study examined two popular deep learning architectures, the Convolutional Neural Network (CNN) and the Recurrent Neural Network (RNN), to extract concepts from clinical texts. We compared the two deep neural network architectures with three baseline Conditional Random Fields (CRFs) models and two state-of-the-art clinical NER systems using the i2b2 2010 clinical concept extraction corpus. The evaluation results showed that the RNN model trained with the word embeddings achieved a new state-of-the-art performance (a strict F1 score of 85.94%) for the defined clinical NER task, outperforming the best-reported system that used both manually defined and unsupervised learning features. This study demonstrates the advantage of using deep neural network architectures for clinical concept extraction, including distributed feature representation, automatic feature learning, and long-term dependencies capture. This is one of the first studies to compare the two widely used deep learning models and demonstrate the superior performance of the RNN model for clinical NER.

Introduction

Clinical studies often require detailed patients’ information documented in clinical narratives. Named Entity Recognition (NER) is a fundamental Natural Language Processing (NLP) task to extract entities of interest (e.g., disease names, medication names and lab tests) from clinical narratives, thus to support clinical and translational research. Researchers have developed computational models and applied them in general clinical NLP systems. Most of the general clinical NLP systems such as MetaMap, MedLEE, and KnowledgeMap, applied rule-based methods that rely on existing medical vocabularies for NER. The clinical NLP community has organized several challenges to examine the performances of state-of-the-art methods. Most of the top-performed systems are primarily based on supervised machine learning models with manually defined features. To further improve the performance, researchers have explored various strategies within the current infrastructure of conventional machine learning models, including ensemble models that combine multiple machine learning methods, hybrid systems that combine machine learning with high-confidence rules, unsupervised features generated using clustering algorithms (e.g., Brown clustering), and domain adaptation to leverage labeled corpora from other domains.

Machine learning methods formulate the clinical NER task as a sequence labeling problem that aims to find the best label sequence (e.g., BIO format labels) for a given input sequence (individual words from clinical text). Researchers have applied many machine learning models, including Conditional Random Fields (CRFs), Maximum Entropy (ME), and Structured Support Vector Machines (SSVMs). Many top-ranked NER systems applied the CRFs model, which is the most popular solution among conventional machine learning algorithms. A typical state-of-the-art clinical NER system usually utilizes features from different linguistic levels, including orthographic information (e.g., capitalization of letters, prefix and suffix), syntactic information (e.g. POS tags), word n-grams, and semantic information (e.g., the UMLS concept unique identifier). Some hybrid models further leverage the concepts and semantic types from the existing clinical NLP systems such as MetaMap, cTAKES. To further improve the performance, researchers have also utilized ensemble methods to combine different machine learning models, such as re-ranking. More recently, researchers also start to examine the unsupervised features derived from large volumes of unlabeled corpora, such as the word clusters generated using Brown clustering and random indexing. The continuous and intensive hard work from the clinical NLP community have boosted the performance of clinical NER, while also identified several bottlenecks that impede further improvement, including:

1) Fragile feature representation. Initially, the dominant feature representation in clinical NER is the bag-of-word model, which is a simplified representation of a piece of text based on the presence/absence of its words...
irrespective of the orders between words, grammatical relation, and semantic information. The bag-of-word model is fragile for clinical NER due to the sparsity problem. For example, the following two similar clinical entities: “mildly dilated right atrium” and “somewhat enlarged left ventricle” have nothing in common using the bag-of-word feature representation. However, they are two related concepts. There is a need for more robust feature representations.

2) Task-specific and time-consuming human feature engineering.27,28 The typical machine learning based clinical NER is composed of two steps: feature extraction and parameter optimization, where feature extraction is the most critical but time-consuming step. In the conventional machine learning solutions, the feature extraction heavily depends on humans while the machine can only handle the parameter optimization supervised by the gold-standard annotations. Researchers manually screen the positive/negative samples to identify possible features (e.g., tokens containing capitalized letters are more likely to have special meanings) and design feature combinations (e.g., body location followed by a disorder mention), commonly referred to as “human feature engineering”. The human feature engineering has several problems. First, the features extracted by human are either incomplete – human cannot enumerate all possible features, or over specified – the same information is repeated in many complex features and feature combinations. Second, researchers must engineer the features again for a different task or different data source. There is an increasing need for automatic feature learning algorithms to release researchers from the time-consuming manual feature engineering.

3) Lack of long-term dependencies.28-30 The typical CRFs based NER systems usually require the set up of a word window for the input sequence of tokens. Many studies that examined the system prediction errors have reported false negatives caused by the lack of long-term dependencies. However, simply increasing the window size cannot solve this problem as it may dilute the signal with more noise into the feature space and greatly increase the training time. The clinical NER need a better architecture to capture long-term dependencies from clinical texts. Recently, there have been increasing efforts to explore a new emerging technology, deep learning31 (or deep neural networks), to improve the current clinical NLP systems. Deep learning is a sub-domain of machine learning that uses deep architectures to learn high-level feature representations. Currently, deep neural networks are commonly used as the unique deep architecture for high-level feature learning. Deep learning models introduced word embedding28,32 as a critical technique to train densely-valued vector representation of words to replace the fragile bag-of-word representation. Each row of the matrix is associated with a word in the vocabulary and each column of the matrix represents a latent feature. The input word sequence can be transformed into a vector by concatenating the corresponding word vectors from the embedding matrix. Deep neural network architectures can learn high-level features automatically to release researchers from time-consuming human feature engineering. To capture long-term dependencies in a word sequence, researchers designed two popular deep architectures, including the Convolutional Neural Networks (CNN)28,33 and the Recurrent Neural Networks (RNN)34. Recent research from the general NLP domain reported that the CNN and RNN developed using only the word embeddings achieved comparable performance as the state-of-the-art CRFs with human engineered features and knowledge from dictionaries.28,34 In the clinical NLP domain, in one of our previous works35, we examined the word embedding trained using a neural network and applied the embeddings to an SVMs model for word sense disambiguation (WSD)36 and a CRFs model for clinical NER. The evaluation results showed that the word embeddings could be used as useful features to improve the state-of-the-art performances of conventional machine learning models for WSD and NER. Later, we developed a CNN based NER method37 and applied it to the Chinese clinical notes, which outperformed a state-of-the-art CRFs model. Jagannatha et al.38 applied RNN for medical event detection from clinical notes and later compared several RNN models with different loss functions. The experimental results using annotated cancer patient notes showed that the RNN outperformed a baseline CRFs model with context features.

However, the previous study by Jagannatha et al.38 compared RNN with a baseline CRFs with only context word features and the evaluation corpus is not openly accessible. Our previous CNN study37 was evaluated using Chinese clinical corpus. It is still not clear which deep architecture is better and whether they could outperform the state-of-the-art clinical NER systems for English clinical corpora. Therefore, it is necessary to further examine the two popular deep architectures, CNN and RNN, using an open challenge corpus and compare them with the most state-of-the-art clinical NER systems. This study is one of the first studies to compare the two widely used deep learning models using an open clinical corpus – the i2b2 2010 clinical concept extraction dataset. We compared the two deep learning models with the state-of-the-art clinical NER systems with human designed features and demonstrated the superior performance of RNN model.
Methods

Data sets

In this study, we reused the word embeddings developed in our previous study\(^3\) using the unlabeled Multiparameter Intelligent Monitoring in Intensive Care (MIMIC) II corpus\(^9\). The MIMIC II corpus is composed of 403,871 notes from four different note types including discharge notes, radiology notes, ECG and ECHO notes. Table 1 shows the detailed information about the MIMIC II corpus. We used the labeled corpus developed for the Concept Extraction task of the 2010 i2b2 NLP challenge for model training and evaluation. We used the same training and test data sets as in the challenge, which consisted of 349 notes for training and 477 notes for testing. For each note, annotators manually extracted entities about Problem, Treatment, and Test. Table 1 shows detailed information for both the training dataset and the test data set.

<table>
<thead>
<tr>
<th>Data set</th>
<th>Notes</th>
<th>Entities</th>
<th>Entity types</th>
</tr>
</thead>
<tbody>
<tr>
<td>i2b2 2010</td>
<td></td>
<td></td>
<td>Problem, Treatment</td>
</tr>
<tr>
<td>Training</td>
<td>349</td>
<td>27,837</td>
<td></td>
</tr>
<tr>
<td>Test</td>
<td>477</td>
<td>45,009</td>
<td>Test</td>
</tr>
<tr>
<td>MIMIC II</td>
<td>N/A</td>
<td>403,871</td>
<td>N/A</td>
</tr>
</tbody>
</table>

Machine learning models

Baseline CRFs systems

The CRFs model approaches clinical NER as a sequence labeling problem that tries to find the best label sequence \(Y^* = y_1, y_2, ... , y_N\) for a given input sequence \(X = x_1, x_2, ... , x_N\). The clinical concept extraction task can be converted into a sequence labeling problem by assigning the annotated entities with appropriate tag representations. This study used the standard "BIO" schema to convert the named entity annotations into sequence labeling tags, in which each word is assigned to a label as following: B = beginning of an entity, I = inside an entity, and O = outside of an entity. As the i2b2 2010 concept extraction challenge focused on three types of clinical concepts, we used a total number of seven different tags (B-problem, B-test, B-treatment, I-problem, I-test, I-treatment, O). For comparison, we constructed three CRFs baseline systems using features extracted at different levels. The first system used only the context word and n-gram features. Based on the basic features, the second systems further added linguistic features and document level features (section names). Similarly, the third system further added knowledge features derived from general clinical NLP systems (MedLEE, MetaMap and KnowledgeMap) and dictionary match features from existing vocabularies (UMLS). All the baseline systems were developed using the CRFs implemented in the CRF++ package.

State-of-the-art NER systems

To challenge the deep learning models, we selected two state-of-the-art systems among all studies regarding the i2b2 2010 clinical concept extraction challenge dataset. The first system is the best-performed system developed by Debruijn et al.\(^1\). This system explored many typical NER features including word features from different linguistic levels, knowledge from existing clinical NLP systems, and Brown clustering. Authors proposed a semi-supervised Markov model solution and achieved the best F1 score during this challenge. The second system\(^2\) applied SSVMs model and further examined the distributional word representation feature generated by Random Indexing algorithm. To the best of our knowledge, this system achieved the best after-challenge performance on the i2b2 2010 dataset up until now.

Deep learning based NER system

Similar to the CRFs model, deep learning models formulate the clinical concept extraction task into a sequence labeling problem using the same “BIO” tagging schema. The input of deep learning models is quite different with conventional machine learning models. The input of CRFs model is human designed features represented in bag-of-word style vector. Whereas, the input for deep learning model is the raw sequence of words in the sentence without human engineering. We applied a word embedding layer to transform the sequence of words into densely-valued vectors, where most of the values are non-zero. Next, we train a deep neural network to learn high-level feature...
representations, capture long-term dependencies, and global features to help identify clinical entities. This study explored two deep architectures including CNN and RNN.

**CNN architecture:** This study adopted a popular CNN architecture using sentence level log-likelihood approach proposed by Dr. Ronan Collobert. This architecture consists of a convolutional layer, a non-linear layer using the hard version of the hyperbolic tangent (HardTanh), and several standard linear layers. This architecture achieved state-of-the-art NER performance in the general English domain and later was applied to many other NLP tasks later. Using word embeddings, each word in the input window can be mapped to an N-dimension vector (N is the embedding dimension). Then, a convolution layer captures the long-term dependencies (or global features) in the hidden nodes. Both the local features and the global features are then concatenated together and fed into a standard neural network trained using stochastic gradient descending. Finally, the classification layer utilized a sentence level log-likelihood approach to calculate the loss supervised by gold-standard annotations. Detailed information for this CNN architecture can be found from our previous study.

**RNN architecture:** RNN is an emerging new deep architecture for sequence data. Recent studies have shown that RNNs have good ability to capture long-term dependencies for sequence data. In this study, we adopted a RNN architecture implemented using the Long Short Term-Memory (LSTM) by Lample et al. The LSTM is the most popular implementation of RNN architecture. A basic LSTM unit is composed of three multiplicative gates, including an input gate to control the proportion of input information transferred to a memory cell; a forget gate to control the proportion of historical information from the previous state; and an output gate to control the proportion of output information to pass on to the next step. We also applied several standard deep learning techniques including character embedding and dropout. For the input layer, we combined the word embeddings and the character embeddings in an input vector. The word embeddings were pre-trained from the MIMIC II corpus and the character embeddings were initiated with random values. The final classification layer of the RNN used a CRFs loss function, which is similar to the CNN architecture examined in this study.

**Experiments and Evaluation**

This study used word embeddings with 50 dimensions, which was pre-trained from the MIMIC II corpus in one of our previous studies. We used a neural network with negative sampling to train the embeddings as our previous study showed that this embedding is better or at least comparable to the word2vec algorithm. For CNN, we used a Java implementation developed in one of our previous study. We compared several combinations of network parameters based on our previous study and finally used the following parameters for CNN: learning rate at 0.01, the word embedding dimension at 50, and hidden node number at 300. For RNN, we adopted a Python implementation using Theano package. Based on the parameters reported by Lample et al., we examined the character embedding sizes and learning rates. The final RNN model used the following parameters: word embedding dimension at 50; character embedding dimension at 25; the LSTM layer for word level is 100 and the LSTM layer for the character level is 25; learning rate at 0.005; dropout probability is 0.5. A Nvidia Tesla K40 GPU was used to train the RNN model. The official evaluation scripts provided by the i2b2 organizers were used to calculate the strict micro-averaged precision, recall, and F1-score.

**Results**

Table 2 compares the performances of the baseline CRFs, the Semi-Markov (best system during the challenge), the SSVMs (current best system developed after the challenge), the CNN and RNN using the i2b2 2010 test dataset. All evaluation scores were based on exact matching. For the baseline CRFs, the linguistic and document level features improved the baseline F1 score from 77.33% to 79.87%. The knowledge base features further boosted the score to 83.60%. For the state-of-the-art systems, the semi-Markov model developed by Debruijn et al. during the challenge further explored a semi-supervised word cluster feature using Brown clusters and applied a semi-supervised Markov model, which achieved the best F1 score at 85.23%. Later, Tang et al. explored the SSVMs model and distributional word representation from the Random Indexing algorithm and further improved the F1 score to 85.82%, which is the best performance score reported using the i2b2 2010 dataset. For deep learning models, the CNN with only word embeddings outperformed the baseline CRFs with word, linguistic and document level features. However, the performance of CNN did not outperform the baseline CRFs further combined with knowledge features. The RNN architecture implemented using bi-direction LSTM neurons outperformed the current best system and achieved the best F1 score (85.91%) on this data set using only the word/character embeddings.
Table 2. Performance comparison of all machine learning models.

<table>
<thead>
<tr>
<th>Approach</th>
<th>Feature</th>
<th>Precision (%)</th>
<th>Recall (%)</th>
<th>F1 Score (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CRFs baselines</td>
<td>Word</td>
<td>82.32</td>
<td>72.92</td>
<td>77.33</td>
</tr>
<tr>
<td></td>
<td>Word+Linguistic+Discourse</td>
<td>83.25</td>
<td>76.75</td>
<td>79.87</td>
</tr>
<tr>
<td></td>
<td>Word+Linguistic+Discourse+MedLEE+KnowledgeMap+DST</td>
<td>86.52</td>
<td>81.04</td>
<td>83.60</td>
</tr>
<tr>
<td>SSVMs by Tang et al.</td>
<td>All features in CRF baselines + Brown clustering + Random indexing</td>
<td>87.38</td>
<td>84.31</td>
<td>85.82</td>
</tr>
<tr>
<td>(Best in challenge)</td>
<td>Word+context+sentence+section+cTAKES + MetaMap+ConText+Brown clustering</td>
<td>86.88</td>
<td>83.64</td>
<td>85.23</td>
</tr>
<tr>
<td>CNN</td>
<td>Word embedding</td>
<td>84.91</td>
<td>80.73</td>
<td>82.77</td>
</tr>
<tr>
<td>RNN</td>
<td>Word embedding</td>
<td>85.33</td>
<td>86.56</td>
<td>85.94</td>
</tr>
</tbody>
</table>

Word: bag-of-word, orthographic such as capitalized letters and special symbols; Discourse: sections, note types; Linguistic: part of speech tags, prefix, and suffix; DST: Dictionary-based Semantic Tagger using UMLS

Discussion

This study examined two deep learning architectures to extract concepts from clinical texts. We constructed three baseline systems using the CRFs model with different levels of features. Two deep learning architectures, including a CNN and an RNN, were developed. We compared the deep learning architectures with the state-of-the-art clinical NER systems using i2b2 2010 corpus. The experimental results using the standard training, test and evaluation showed that the RNN model trained with only word embeddings achieved the new state-of-the-art performance for clinical NER, which outperformed the best system during the challenge and the current best system based on an SSVMs model. This study shows the advantage of using deep neural network architectures for information extraction from clinical texts. To the best of our knowledge, this is the first study comparing two popular deep learning architectures (CNN and RNN) for clinical concept extraction.

All deep learning models developed using only word embeddings outperformed the baseline CRFs with basic word level, linguistic level, and document level features, showing the efficiency of automatic feature learning from large unlabeled corpora. The performance improvements of the deep learning architectures are mainly from the recall (from 72.92% to 86.56% for RNN), showing that the unsupervised feature learning can capture extra features that did not exist in the training dataset to boost the recall. The RNN outperformed other systems with a new state-of-the-art F1 score at 85.94%. To the best of our knowledge, this is the best performance ever reported for the i2b2 2010 clinical concept extraction dataset. The RNN architecture outperformed CNN, another deep neural network architecture designed to learn high-level feature representations, showing that the RNN architecture is more efficient for sequence labeling tasks.

As an emerging technology, deep learning provides distributed word representation to replace the fragile bag-of-word model, automatic high-level feature learning to release researchers from time-consuming feature engineering, and deep architectures to capture long-term dependencies. We do observed differences between the general NLP and clinical NLP when applying deep learning models. For example, the clinical NLP have more knowledge bases (e.g., UMLS) with decent coverage. However, it is very hard to generate such comprehensive knowledge bases with decent coverage in the general NLP domain. Therefore, we can see that most of the top-performing clinical NER systems utilized the knowledge from dictionaries and integrated with other clinical NLP systems, which may make it hard for the deep learning models using only word embeddings to compete with the traditional clinical NER models in the clinical domain. For example, the CNN architecture outperformed the state-of-the-art NER systems on both the general English NER dataset and the Chinese clinical corpus, where there are limited knowledge bases. Whereas, the CNN model did not outperform the state-of-the-art clinical NER systems on the i2b2 corpus, where the top
systems utilized many knowledge bases and were hybrid systems using multiple general clinical NLP systems. The success of the RNN architecture shows the promise of using deep learning architectures for clinical NLP: machines can learn features better than humans, given the correct deep neural network architectures and advanced optimization algorithms.

This study has limitations. We only explored hidden layer dimension and learning rate but used arbitrarily selected network parameters according to previous studies. More robust and efficient parameter tuning strategy is needed. We compared RNN with the current best NER system and reported a new state-of-the-art performance on the i2b2 data set. Although the numeric improvement looks small, this study demonstrated the efficiency of RNN using the minimal feature engineering. Deep learning architectures provide a unified solution for clinical NLP, which may simplify the complex system architecture of current state-of-the-art clinical NLP systems to speed up their application in practical NLP systems. With the distributed word representation and high-level feature abstraction, deep learning may achieve the state-of-the-art performance without the requirement of combining with other models and systems. It worth further investigating on how the deep learning models can be applied to practical clinical NLP systems.

Conclusion

This study compared two deep learning models, including the CNN and RNN, to extract clinical concepts from clinical texts. We compared CNN and RNN with baseline CRFs using different levels of features as well as the state-of-the-art clinical NER systems using the i2b2 2010 clinical concept extraction corpus. The experimental results show that the RNN outperformed the best clinical NER system based on SSVMs with a new state-of-the-art F1 score at 85.94%. The RNN model can be adapted to other clinical NLP tasks to improve their performances and to release researchers from time-consuming feature engineering.

Acknowledgement

This study was supported by grants from the NLM 2R01LM010681, NIGMS 1R01GM103859, NCI U24 CA194215 and 1R01GM102282. We would like to thank the 2010 i2b2/VA challenge organizers for the development of the corpus used in this study. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Tesla K40 GPU used for this research.

References


Using EHR audit trail logs to analyze clinical workflow: A case study from community-based ambulatory clinics

Danny T. Y. Wu, PhD, MS1,8, Nikolas Smart, MS1,2, Elizabeth L. Ciemins, PhD, MPH, MA3,4, Holly J. Lanham, PhD, MBA5,6,7, Curt Lindberg, DMan4, Kai Zheng, PhD2,1,9

1School of Information, University of Michigan, Ann Arbor, MI, 2Department of Health Management and Policy, University of Michigan, Ann Arbor, MI; 3Billings Clinic Center for Clinical Translational Research, Billings, MT; 4American Medical Group Association, Alexandria, VA; 5University of Texas Health Science Center, San Antonio, TX; 6South Texas Veterans Health Care System, TX; 7University of Texas at Austin, TX; 8Department of Biomedical Informatics, University of Cincinnati, OH; 9Department of Informatics, University of California, Irvine, CA

Abstract

To develop a workflow-supported clinical documentation system, it is a critical first step to understand clinical workflow. While Time and Motion studies has been regarded as the gold standard of workflow analysis, this method can be resource consuming and its data may be biased due to the cognitive limitation of human observers. In this study, we aimed to evaluate the feasibility and validity of using EHR audit trail logs to analyze clinical workflow. Specifically, we compared three known workflow changes from our previous study with the corresponding EHR audit trail logs of the study participants. The results showed that EHR audit trail logs can be a valid source for clinical workflow analysis, and can provide an objective view of clinicians’ behaviors, multi-dimensional comparisons, and a highly extensible analysis framework.

Introduction

In recent years, clinical data capture and documentation has been increasingly supported by computer systems, especially Electronic Health Record (EHR) systems, as incentivized and regulated by the HITECH Act and the Meaningful Use1,2. While computer-based documentation (CBD) may be preferred due to its potential improvements in information availability and sharing as well as enhancement in patient care3,4, CBD may introduce negative effects if not designed properly. One significant negative effect is that a poorly designed CBD system contributes to cumbersome documentation activities and inaccurate and incomplete data capture when it fails to consider clinical workflow5.

In order to address this issue and design a workflow-supported CBD system, a critical first step is to understand and measure the current workflow and documentation behaviors, which has been frequently achieved by gold standard techniques such as Time and Motion (T&M) studies6. T&M studies, however, are often costly and incapable of collecting data in a large scale because of the challenges in recruiting numerous participants and observers as well as coordinating observation sessions. Meanwhile, T&M studies are inherently limited in accuracy due to human observers’ cognitive overload and biases, resulting in suboptimal quality of behavioral data such as incompleteness and misinterpretation.

An alternative method to make sense of clinical workflow is through the secondary use of EHR data, which has drawn significant attention from researchers. For example, Ozkaynak et al. (2015) characterize the workflow in pediatric emergency rooms by analyzing encounter and diagnosis data using Markov chain and visualization tools7. Kirkendall et al. (2014) understood the workflow of medication order-dosing alerts by simulating order activities in a testing EHR environment8. Redd et al. (2014) measured the efficiency of documentation by extracting the chart open and completion time in each patient encounter9. These studies, although suitable for their purposes, only utilized EHR data on a surface level. The data and methods in these studies did not have the ability to uncover more nuanced behavioral patterns of a large number of individual clinicians.

The clinicians’ behavioral patterns can be uncovered by a special type of EHR data, EHR audit trail logs, which automatically and minutely record clinicians’ EHR interactions for security and privacy purposes. These logs can be a cost-effective alternative to characterize clinicians’ documentation behaviors and have the potential to facilitate the understanding of clinical workflow and processes. For example, Hirsh et al. used audit files to understand time spending in primary care10. However, to our knowledge, there is no study evaluating EHR audit trail logs as a valid
source for clinical workflow analysis. This study therefore bridges this methodological gap by assessing the ability of these logs to provide supportive evidence to known workflow changes. Specifically, the study analyzed a sample of EHR audit trail log data extracted from an ambulatory healthcare organization. This organization launched an EHR enhancement project in mid 2013, resulting in subsequent workflow and behavioral changes as identified in our previous study. The present study chooses three findings to evaluate the ability of the EHR audit trail logs to provide supportive evidence. The implications of utilizing such log data for clinical workflow analyses are discussed. It is worth noting that these evaluation standards were validated based on the triangulation of our previous semi-structured interviews, ethnographic observations, and T&M studies. The qualitative analysis is out of the scope of this paper and will be published in another venue.

Methods

Study Site

The participating study site was a not-for-profit ambulatory healthcare organization located in an urban area in the western United States (Organization West). This organization contained more than 20 branches and affiliated locations including both primary care and specialty clinics. Organization West served several states over a widely rural area, with over 50% of patients being rural residents and around 50% of who are on Medicare/Medicaid. After nearly 10 years of use of a vendor EHR system, Organization West launched a significant update to enhance the system capability in 2013, including an electronic patient homepage, a standardized message center, a new computerized provider order entry (CPOE), and an e-prescribing functionality. During the EHR enhancement, one of the clinics experimented with a “Core Team Model,” in which a provider was teamed up with three clinical staff members to eliminate inefficiencies and patient wait times.

Data Collection

This study targeted three clinics in Organization West, including a primary care clinic in the main hospital (Primary Care 1), another primary clinic that experimented with the “Core Team Model” (Primary Care 2), and an Ear Nose Throat clinic (Specialty Care). A total of 24 clinicians were participated and equally distributed in each clinic. For each primary care clinic, the study participants included four providers, two nurses, one medical assistant, and one receptionist. For the specialty care clinic, the study participants included five providers, two nurses, one medical assistant, and no receptionist. The providers included in the study were physicians (MD), ophthalmologists (OD), nurse practitioners (NP), and physician assistants (PA). The participants’ EHR audit trail logs were extracted before, during, and after the EHR system changes. This data allowed the comparison of clinicians’ behavioral changes in many dimensions, e.g. primary care vs. specialty care; with and without the "Core Team Model"; and providers vs. non-providers. Since the quantitative analysis team (DW, NS, and KZ) was not physically located in the study sites, the log data were de-identified and exported as CSV files for analysis.

Data Manipulation

The log data were stored in a standalone SQLite database and manipulated using Structured Query Language. Specifically, the log data were categorized into three stages based on the timestamp: 1) PRE: log records comprised of a period of eight weeks prior to the launch date, 2) DURING: log records for a period of four weeks after the launch date, and 3) POST: log records of the eight weeks after the end of the DURING stage. The analysis in this study only focused on the data in the PRE and POST stages to examine the potential differences. The start and end of a week was set from Monday to Sunday. Moreover, the date and time information of the timestamps were parsed and stored separately, e.g. date, hour, and day of the week. The time information was further categorized into three types: 1) Regular hours (REG): time between 7am and 6pm from Monday to Friday; 2) off hours (OFF): time not in regular hours from Monday to Friday; and 3) weekend hours (WKN): anytime during Saturday and Sunday only. The start and end of the regular hours was determined based on our previous interviews and observations.

In addition, the clinical roles in the log data were marked as Provider (P), Nurse (N), Medical Assistant (A), and Receptionist (R). Note that the study created seven new event categories in addition to the original event names and tasks defined by the EHR vendor, including: 1) Communication, 2) Entering, 3) Login/Logout, 4) Processing, 5) Reading, 6) Printing, and 7) Other. Each event was first mapped to a category separately by the first two authors. Then, the mappings were discussed and an agreement was reached. If a consensus of an event mapping cannot be reached, this event will be categorized as “Other”.

1804
Hypotheses

The primary objective of this study was to evaluate the ability of EHR audit trail log data to provide supportive evidence to the known workflow changes. In particular, it aimed to reflect documentation-related behavioral changes among clinicians after the rollout of an EHR enhancement project. Three previously validated findings were chosen as the reference standards of the present study. The first evaluation involved a workaround order-placing behavior in Primary Care 1, where a provider likely had to prescribe a medication more than once in order to achieve the desired dose in the POST stage. For example, a provider would place two medication orders, one at 1.5 mg and the other at 0.5 mg, to equal the total intended dose of 2 mg. Since the order dosage was not captured in the log data, the providers in Primary Care 1 were alternatively hypothesized to have more number of order logs daily. Their patient volume in the PRE and POST stage was hypothesized to remain stable.

H1: After the EHR enhancement, the providers in Primary Care 1 had more number of order logs, but their patient volume remained stable.

The second evaluation was a shift of tasks in Primary Care 2, where providers spent less time on computer-related tasks after the EHR enhancement, while other clinicians consumed those tasks due to the implementation of the Core Team model. The providers in Primary Care 2 were hypothesized to have fewer “Computer Entering”-related logs in the POST stage, compared to the PRE stage. On the other hand, non-providers in the POST stage were hypothesized to have more “Computer Entering”-related logs.

H2: After the EHR enhancement, the providers in Primary Care 2 had fewer number of “Computer Entering”-related logs, while the non-providers had more such logs.

The third evaluation was the phenomenon of “deferred documentation” in Specialty Care clinic, highlighting differences in dictation. It is noted that providers in the PRE stage would dictate findings between patient visits, while in the POST stage they tended to document patient visits at the end of a day. It is hypothesized that providers in Specialty Care would have an increased number of logs in the off hours (and even weekends) in the POST stage.

H3: After the EHR enhancement, the providers in Specialty Care had more number of logs in the off hours.

Data Analysis

The data were analyzed through self-developed Python scripts, which were implemented using libraries such as pandas, numpy, scipy, and xlrd. The data analysis used two primary measures, namely, the number of logs and the percentage of logs. The analysis began with a statistical summary of all logs, followed by the statistical tests of the three hypotheses. For the first hypothesis (H1), the log records related to “adding an order” were selected based on keywords. The total number of orders placed per day was counted using the order identifiers stored in the logs. The days without any orders placed were dropped. On the other hand, the daily patient volume was estimated by the total number of patients accessed per day. For the second hypothesis (H2), the log records were manually coded into the seven high level categories, followed by the calculation of the number and percentage of logs in each category (e.g. Computer Entering). For the third hypothesis (H3), the same measures as H2 were used to quantify the off-hour and weekend logs. All the above measures in the PRE and POST stage were compared using a two-tailed, unequal variance t-test.

In addition, in H2 the fragmentation of the log events was assessed based on their categories. This measure calculated a fragmentation score representing the level of cognitive overload (12). The higher the score is, the more switches between categories the clinician had in the log data, and the more likely this person was suffered from cognitive overload. It is worth noting that a person may have multiple logs with varying categories but identical timestamp because multiple documentation behaviors may happen, or be triggered, at the same time. Fragmentation was detected if there was no overlap between any event categories of the current timestamp and its previous one. A daily fragmentation score was defined as the total number of detected fragmentation over the total number of timestamps on that day. The clinicians’ daily fragmentation scores in the PRE and POST stage were compared using the same t-test technique as the hypotheses.

Results

Statistical Summary

A total of 2.65 million EHR audit trail logs were extracted from Organization West in both PRE and POST stage, for a period of eight weeks each. Overall, the log volume of Primary Care 1 & 2 was approximately the same, whereas the log volume of Specialty Care was half of the previous two (Table 3.1). This is likely due to the differences of clinical practices between a primary and a specialty care clinic. In terms of the log volume by clinical roles, the
providers in Primary Care 2 had a much lower number than their nurses (87,600 vs. 141,090). In comparison, the average number of logs of the providers and nurses in Primary Care 1 were more comparable (111,351 vs. 116,078).

A further examination of the log volume considering the PRE and POST stage is listed in Table 3.2. This table shows that the nurses in Primary Care 1 had a 20% decrease of their log volumes and the providers had a 20% increase and in the POST stage. On the other hand, the nurses in Primary Care 2 had a 11% increase in their log volumes while the providers maintain a similar number of logs. This observation might be attributed to the implantation of the Core Team model in Primary Care 2, which increased the nurses’ documentation responsibility.

Table 3.1. Summary of the sample EHR audit trail log data.

<table>
<thead>
<tr>
<th>Clinic</th>
<th>PRE (8 wks)</th>
<th>Launch Date</th>
<th>POST (8 wks)</th>
<th>Total (# log)</th>
<th>Provider (avg. total)</th>
<th>Nurse (avg. # log)</th>
<th>Assistant (avg. # log)</th>
<th>Receptionist (avg. # log)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Primary Care 1</td>
<td>5/20/13</td>
<td>7/16/13</td>
<td>8/12/13</td>
<td>1,149,295</td>
<td>4 (111,351)</td>
<td>2 (116,078)</td>
<td>1 (215,347)</td>
<td>1 (256,386)</td>
</tr>
<tr>
<td>Primary Care 2</td>
<td>6/24/13</td>
<td>8/20/13</td>
<td>9/16/13</td>
<td>971,210</td>
<td>4 (87,600)</td>
<td>2 (141,090)</td>
<td>1 (143,758)</td>
<td>1 (194,868)</td>
</tr>
<tr>
<td>Specialty Care</td>
<td>8/5/13</td>
<td>10/1/13</td>
<td>10/28/13</td>
<td>529,999</td>
<td>5 (64,099)</td>
<td>2 (74,045)</td>
<td>1 (61,410)</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 3.2. The average number of logs in the PRE and POST stage (providers and nurses).

<table>
<thead>
<tr>
<th>Clinic/Role-Stage</th>
<th>Provider (4)</th>
<th>Nurse (4)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PRE</td>
<td>POST</td>
</tr>
<tr>
<td>Primary Care 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Primary Care 2</td>
<td></td>
<td>+1.6%</td>
</tr>
<tr>
<td>Specialty Care</td>
<td>-24.9%</td>
<td>-41.4%</td>
</tr>
</tbody>
</table>

Hypothesis One

As shown in Table 3.3, the statistical test indicated that the providers in Primary Care 1 placed significantly more orders in the POST stage (all p<0.05) and maintained the same number of patients seen, which confirms our hypotheses. The same statistical test was further extended to the providers in Primary Care 2, which didn’t show the same pattern. As shown in the second half of Table 3.3, two of the four providers in Primary Care 2 had no significant change in their order numbers, with one of them seeing more patients in the POST stage. This finding suggests that changes of EHR systems may impact clinicians’ behaviors differently even though they operate within the same health system and have similar clinical practice.

Table 3.3. Number of orders placed and patients seen for each provider in the two primary care clinics.

<table>
<thead>
<tr>
<th>Clinic</th>
<th>Measure</th>
<th>Provider</th>
<th>PRE</th>
<th>POST</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Primary</td>
<td>Number of Orders Placed</td>
<td>BC1-04</td>
<td>12.00</td>
<td>35.25</td>
<td>0.0 *</td>
</tr>
<tr>
<td>Care 1</td>
<td></td>
<td>BC1-05</td>
<td>42.35</td>
<td>53.15</td>
<td>0.0192 *</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BC1-09</td>
<td>44.68</td>
<td>53.53</td>
<td>0.0433 *</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BC1-10</td>
<td>15.94</td>
<td>30.14</td>
<td>0.0 *</td>
</tr>
<tr>
<td>Primary</td>
<td>Number of Patients Accessed</td>
<td>BC1-04</td>
<td>42.50</td>
<td>34.75</td>
<td>0.3491</td>
</tr>
<tr>
<td>Care 1</td>
<td></td>
<td>BC1-05</td>
<td>42.57</td>
<td>42.58</td>
<td>0.9971</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BC1-09</td>
<td>31.45</td>
<td>35.32</td>
<td>0.1939</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BC1-10</td>
<td>40.47</td>
<td>41.94</td>
<td>0.6158</td>
</tr>
<tr>
<td>Primary</td>
<td>Number of Orders Placed</td>
<td>BC2-02</td>
<td>23.04</td>
<td>32.11</td>
<td>0.0066 *</td>
</tr>
<tr>
<td>Care 2</td>
<td></td>
<td>BC2-03</td>
<td>16.69</td>
<td>16.07</td>
<td>0.8386</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BC2-04</td>
<td>28.89</td>
<td>38.12</td>
<td>0.0342 *</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BC2-16</td>
<td>27.38</td>
<td>33.84</td>
<td>0.1826</td>
</tr>
<tr>
<td>Primary</td>
<td>Number of Patients Accessed</td>
<td>BC2-02</td>
<td>31.77</td>
<td>28.34</td>
<td>0.1475</td>
</tr>
<tr>
<td>Care 2</td>
<td></td>
<td>BC2-03</td>
<td>29.47</td>
<td>35.11</td>
<td>0.0399 *</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BC2-04</td>
<td>36.39</td>
<td>34.46</td>
<td>0.5024</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BC2-16</td>
<td>23.14</td>
<td>22.04</td>
<td>0.6417</td>
</tr>
</tbody>
</table>

*p-value < 0.05
Hypothesis Two

The second evaluation began at coding the log events to the seven high-level categories. The first two authors achieved 81% agreement on coding the event categories initially, and resolved most of the discrepancies afterwards. Table 3.4 lists top ten frequent events that were assigned to “Computer Entering”, and their vendor defined names and types as well as their frequencies in Primary Care 2.

Table 3.4. Top 10 “Computer Entering” events

<table>
<thead>
<tr>
<th>EVENT TYPE</th>
<th>EVENT NAME</th>
<th>COUNT</th>
<th>PCT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maintain Person</td>
<td>Chart Access Log</td>
<td>141,616</td>
<td>42%</td>
</tr>
<tr>
<td>Maintain Person</td>
<td>Ensure</td>
<td>70,647</td>
<td>21%</td>
</tr>
<tr>
<td>Maintain Encounter</td>
<td>Patient-Provider Relations</td>
<td>22,068</td>
<td>7%</td>
</tr>
<tr>
<td>View Orders</td>
<td>Modify Details</td>
<td>16,719</td>
<td>5%</td>
</tr>
<tr>
<td>Maintain Encounter</td>
<td>Ensure</td>
<td>14,681</td>
<td>4%</td>
</tr>
<tr>
<td>Maintain Person</td>
<td>Patient</td>
<td>12,921</td>
<td>4%</td>
</tr>
<tr>
<td>Maintain Clinical Events</td>
<td>Write/Update Results</td>
<td>11,331</td>
<td>3%</td>
</tr>
<tr>
<td>Maintain Order</td>
<td>Add</td>
<td>9,770</td>
<td>3%</td>
</tr>
<tr>
<td>Maintain Order</td>
<td>Tasks</td>
<td>5,939</td>
<td>2%</td>
</tr>
<tr>
<td>Maintain Encounter</td>
<td>Prsnl Relationship</td>
<td>4,894</td>
<td>1%</td>
</tr>
</tbody>
</table>

The statistical test examined the differences of the “Computer Entering”-related events between the PRE and POST stage. The results (Table 3.5) showed that the participated nurses (BC2-11 & 19) and the medical assistant (BC2-22) had a significant increase (3-5%) in their percentage of computer entering logs as hypothesized. However, the receptionist (BC2-45) had an unexpected drop in their percentage of entering-related logs (53% -> 49%). On the other hand, the providers had a slight but not significant decrease in their number and percentage of logs, which is also unexpected.

Table 3.5. The average number and percentage of “Computer Entering” logs in the PRE and POST stage.

<table>
<thead>
<tr>
<th>Clinical Role</th>
<th>Participant ID</th>
<th>Number of Logs</th>
<th>Percentage of Logs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>PRE</td>
<td>POST</td>
</tr>
<tr>
<td>P</td>
<td>BC2-02</td>
<td>364.19</td>
<td>354.72</td>
</tr>
<tr>
<td>P</td>
<td>BC2-03</td>
<td>484.57</td>
<td>486.17</td>
</tr>
<tr>
<td>P</td>
<td>BC2-04</td>
<td>327.75</td>
<td>294.79</td>
</tr>
<tr>
<td>P</td>
<td>BC2-16</td>
<td>341.48</td>
<td>318.71</td>
</tr>
<tr>
<td>N</td>
<td>BC2-11</td>
<td>515.22</td>
<td>655.14</td>
</tr>
<tr>
<td>N</td>
<td>BC2-19</td>
<td>515.33</td>
<td>603.91</td>
</tr>
<tr>
<td>A</td>
<td>BC2-22</td>
<td>673.34</td>
<td>683.12</td>
</tr>
<tr>
<td>R</td>
<td>BC2-45</td>
<td>1375.36</td>
<td>1236.83</td>
</tr>
</tbody>
</table>

* 0.05 significance

The workflow fragmentation analysis of these providers (Table 3.6) indicated that three out of four providers had a lower level of fragmentation, although only one of them was statistically significant. The other providers had a slight but not significant increase in the level of fragmentation (0.398 -> 0.401). Together, the analysis of the second evaluation supports the hypotheses of shift clinical workflow, where the nurses and the medical assistant had an increased percentage of “Computer Entry”-related logs. While the providers didn’t have a significant increase, they had an improved level of fragmentation in the POST stage, which may help reduce their cognitive overload, leading to their positive attitude toward the EHR change (i.e. more time with patients).

Table 3.6. The workflow fragmentation analysis of the providers in the Primary Care 2. The higher the score is, the more switches between log event categories the clinician had, the more likely cognitive overload happened.

<table>
<thead>
<tr>
<th>Participant ID</th>
<th>PRE</th>
<th>POST</th>
<th>Difference</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>BC2-02</td>
<td>0.428</td>
<td>0.361</td>
<td>-0.067</td>
<td>0.0043*</td>
</tr>
<tr>
<td>BC2-03</td>
<td>0.388</td>
<td>0.365</td>
<td>-0.023</td>
<td>0.2669</td>
</tr>
<tr>
<td>BC2-04</td>
<td>0.358</td>
<td>0.306</td>
<td>-0.052</td>
<td>0.1305</td>
</tr>
<tr>
<td>BC2-16</td>
<td>0.398</td>
<td>0.401</td>
<td>0.003</td>
<td>0.8725</td>
</tr>
</tbody>
</table>

* 0.05 significance
Hypothesis Three

The initial statistical test of the providers’ weekly off-hour logs in Specialty Care, unfortunately, did not find any significant change at the 0.05 level. The statistical tests of the weekend logs also found no significant differences. A further exploratory analysis using visualization (line charts) suggested that the providers in Specialty Care had an increased number of logs right after 5pm (Figure 3.1). The data were therefore re-grouped and tested on an hourly basis. As shown in Table 3.8, four of five providers in Specialty Care had a significantly higher number of logs. They deferred documentation activities in at least one of the hours between 5pm and 10pm. The other provider (BC3-01) seemingly had already deferred the documentation activities in the PRE stage, which makes the EHR change less impactful in the POST stage.

Figure 3.1. The visualization of the number of off-hour logs of the provider BC3-04. X-axis represents the hours of a day and Y-axis represents the number of logs of all providers.

Table 3.7. The average number of daily off-hour log in the PRE and POST stage.

<table>
<thead>
<tr>
<th>Participant ID</th>
<th>Hour</th>
<th>PRE</th>
<th>POST</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>BC3-01</td>
<td>5pm</td>
<td>173.26</td>
<td>147.7</td>
<td>0.7476</td>
</tr>
<tr>
<td></td>
<td>6pm</td>
<td>118.96</td>
<td>186.5</td>
<td>0.3326</td>
</tr>
<tr>
<td></td>
<td>7pm</td>
<td>61.93</td>
<td>86.25</td>
<td>0.6350</td>
</tr>
<tr>
<td></td>
<td>8pm</td>
<td>11.07</td>
<td>0.50</td>
<td>0.1782</td>
</tr>
<tr>
<td></td>
<td>9pm</td>
<td>0</td>
<td>0</td>
<td>-</td>
</tr>
<tr>
<td>BC3-04</td>
<td>5pm</td>
<td>0.91</td>
<td>16.65</td>
<td>0.0263 *</td>
</tr>
<tr>
<td></td>
<td>6pm</td>
<td>0.36</td>
<td>19.10</td>
<td>0.0373 *</td>
</tr>
<tr>
<td></td>
<td>7pm</td>
<td>0.33</td>
<td>12.00</td>
<td>0.064 **</td>
</tr>
<tr>
<td></td>
<td>8pm</td>
<td>4.61</td>
<td>12.05</td>
<td>0.2927</td>
</tr>
<tr>
<td></td>
<td>9pm</td>
<td>3.33</td>
<td>0</td>
<td>0.4416</td>
</tr>
<tr>
<td>BC3-05</td>
<td>5pm</td>
<td>53.14</td>
<td>64.42</td>
<td>0.6525</td>
</tr>
<tr>
<td></td>
<td>6pm</td>
<td>17.05</td>
<td>13.50</td>
<td>0.8124</td>
</tr>
<tr>
<td></td>
<td>7pm</td>
<td>31.97</td>
<td>0.04</td>
<td>0.1661</td>
</tr>
<tr>
<td></td>
<td>8pm</td>
<td>12.11</td>
<td>0.29</td>
<td>0.3917</td>
</tr>
<tr>
<td></td>
<td>9pm</td>
<td>0.14</td>
<td>26.75</td>
<td>0.0393 *</td>
</tr>
<tr>
<td>BC3-10</td>
<td>5pm</td>
<td>71.74</td>
<td>88.26</td>
<td>0.6039</td>
</tr>
<tr>
<td></td>
<td>6pm</td>
<td>43.71</td>
<td>19.22</td>
<td>0.2146</td>
</tr>
<tr>
<td></td>
<td>7pm</td>
<td>0</td>
<td>0.35</td>
<td>0.0978 **</td>
</tr>
</tbody>
</table>
Discussion

This case study demonstrated the validity of using EHR audit trail logs to analyze clinical workflow. The results indicated that EHR audit trail logs contained rich information of clinicians’ documentation behaviors. The analyses base on the log data successfully produced supportive evidence to the three known behavioral and workflow changes. For the first hypothesis, the data confirmed that the providers in Primary Care 1 placed more orders and maintained the same patient volume. For the second hypothesis, the data supported that the nurses in Primary Care 2 had more “Computer Entering”-related activities, while the providers may have a lower cognitive overload. For the third hypothesis, the data indicated that the providers in Specialty Care deferred their documentation behaviors, although each provider had a different working timeframe.

The study showed several benefits of using EHR audit trail logs to analyze clinical workflow. First of all, since the log data were recorded automatically and consistently, an analysis of one clinic can be easily applied to another clinic with almost no extra cost. Take the first hypothesis for example, the changes of the order numbers and patient volumes was initially examined in Primary Care 1 and then directly applied to Primary Care 2 with nearly no extra effort. This benefit of easy extension makes this method more attractive than other workflow analysis tools, e.g. Time and Motion studies and work sampling.

Another benefit of using EHR logs for workflow analysis is to provide an objective view of clinicians’ EHR behaviors and to verify the findings derived from other methods. Take the second hypothesis for example, the previous study suggested that the providers in Primary Care 2 had more time for patient communication because they shifted documentation responsibilities to other clinicians as a result of the Core Team model. The analysis of the present study, however, showed that the providers’ EHR log volume did not drop significantly although other clinicians, especially nurses, indeed had a significant increase of the log volume. One possible explanation is that EHR audit trail logs may not fully capture the documentation process as they are not design for such intension. Another explanation of this phenomenon, as suggested by the workflow fragmentation analysis, is that the providers may have less fragmented documentation activities, leading to a smoother workflow and higher quality of patient interactions.

Still another benefit of using EHR logs for workflow analysis is the ability to compare and contrast results in multiple dimensions. In this study, the dimensions included clinical roles (providers vs. non-providers), care types (primary vs. specialty), care models (normal vs. the core team model), time (regular vs. off-hour), and event categories (Entering vs. Other). Moreover, EHR audit trail logs enable detailed comparison showing the differences of individual clinician’s behaviors. Take the third hypothesis for example, the analysis not merely supported the hypothesis of deferred documentation in Specialty Clinic, but further indicated which hour(s) the documentation activities were deferred to, which were varying among the five providers.

Although the rich information in EHR audit trail logs has several benefits, it can introduce analytical challenges. One such challenge is the high complexity and high level of noise in the log data. It could be very difficult to identify meaningful patterns solely based on EHR log data. The present study voided this problem by utilizing the findings from the previous study to shed light on the analysis. This suggests that using EHR logs for clinical workflow analysis may require mixed methods, e.g. integrating the qualitative results to the quantitative analysis. It may also require extensive effort on data exploratory, which can be supported by visual analytics techniques. Another challenge is that EHR logs may require high-level categorization to reduce their complexity and to make them more meaningful. For example, in the present study, the logs were manually categorized into seven groups in order to compare “Computer Entering” activities with other kinds of activities. Since each vendor has a different mechanism capturing audit trail logs, and since each study has different analysis goals, it seems that researchers would need to spend a significant amount of time labeling EHR logs. Data mining and machine learning techniques may help facilitating this process.
This study has several limitations. First, this study only analyzed EHR audit trail logs of one organization in a community-based, ambulatory setting. Since this is a proof-of-concept study, we believe that the scope is adequate. Future research can consider expanding the scope to analyze EHR logs in an inpatient setting, of different EHR systems, and/or from multiple organizations. Second, the study did not obtain the EHR audit trail logs of all clinicians in the targeted organization, but only the logs associated with the 24 participants in the three selected clinics. This limitation prevents the study from performing a comprehensive patient-centered analysis because the selected participants may not work on the same patients in the study period. Third, although we conducted interviews, observations, and T&M studies to understand workflow and behavioral changes in our previous study, we did not compare step-by-step how clinical behaviors are recorded to form audit trail logs. Rather, we understood the log data based on the description of the attributes, the sequences, and the timestamps of the logs. We formed an interdisciplinary research team to mitigate potential biases in this interpretation. Next, we only employed two volume-related measures in this study. We have been working on applying data mining techniques, especially in sequential pattern mining, to uncover unexpected but interesting workflow patterns from the EHR audit trail logs. Last but not least, many results were not significant at the 0.05 level, and even not significant at the 0.1 level. However, a statistically insignificant difference may still have clinical impact. These insignificant differences still showed changes (increase or decrease of the means) aligned with our hypotheses.

Conclusions
This study evaluated the validity of using EHR audit trail logs for clinical workflow analysis. The three evaluations successfully provided supporting evidences and justify the use of this type data. Future research includes developing a generalized framework capable of analyzing log data from any EHR systems, advanced computational and visual-based methods to uncover more nuance patterns of clinicians’ behavioral changes, and a real-time monitoring system to detect workflow bottlenecks, physician burnout, and opportunities to support medical education.

References
Learning Doctors’ Medicine Prescription Pattern for Chronic Disease Treatment by Mining Electronic Health Records: A Multi-Task Learning Approach

Eryu Xia, PhD¹, Jing Mei, PhD¹, Guotong Xie, PhD³, Xuejun Li, MD, PhD²,³, Zhibin Li, PhD³, Meilin Xu⁴

¹ IBM Research - China, Beijing, China; ² Department of Endocrinology and Diabetes, the First Affiliated Hospital, Xiamen University, Xiamen, China; ³ Xiamen Diabetes Institute, Xiamen, China; ⁴ Pfizer Investment Co. Ltd., Beijing, China

Abstract

Increasing learning ability from massive medical data and building learning methods robust to data quality issues are key factors toward building data-driven clinical decision support systems for medicine prescription decision support. Here, we attempted accordingly to address the factors using a multi-task neural network approach, benefiting from multi-task learning’s advantage in modeling commonalities to increase learning performance and neural network’s robustness to imprecise data. By mining electronic health record data, we learned medicine prescription patterns of multiple correlated antidiabetic agents in blood glucose control and antihypertensive drugs in blood pressure control scenarios. We achieved AUC increases of 0.02 to 0.06 in single drug prescription and an accuracy increase of 0.05 in prescription pattern prediction compared to logistic regression, demonstrating the efficacy of multi-task neural network approach in learning medicine prescription patterns.

Introduction

Clinical decision support systems (CDSSs) are health information technology systems that assist clinical decision-making tasks. CDSS can be broadly divided into two types of approaches: knowledge-based that relies on digitalized clinical guidelines, and data-driven that relies on gathering evidence from real-life medical data. Many publications convey the belief that CDSS, especially on electronic health record (EHR) platform, will provide decision support to achieve large gains in clinical performance, improving its quality, safety, and cost efficiency¹-³. The accumulation of EHR data has allowed us the opportunity to use big data analytics approaches for data-driven CDSS. Prior work on EHR data includes mining association rules and Bayesian networks between clinical orders and diagnoses⁴, ⁵. Data-driven CDSS from EHR can be regarded as learning from collective experience of many practitioners to automatically generate evidence for clinical decision support.

Though interest in data-driven CDSS has grown, its application remains limited. The reasons are primarily two-folds: the quality of data, and the ability to learn from data. EHR data suffers from three main data quality issues: incompleteness, inaccuracy, and inconsistency⁶. Incompleteness is reflected by missing values of many attributes. Inaccuracies are unavoidable during typing and are frequently caused by inappropriate use of ICD codes. Inconsistency is also reflected by the inconsistent ICD code use and the inconsistent value-unit match for laboratory tests. Apart from these EHR data quality issues, uncertainty and inconsistencies abound in clinical decision making due to a lack of consensus view on a same clinical condition and evidence to confirm or deny the efficacy of each view⁷, rendering each diagnosis, clinical order, and prescription less of a ground truth. Even with perfect data, the effectiveness of data-driven methods still depend on the capacity of machine learning algorithms to solve the problem.

In response to the limitations, efforts toward data-driven CDSS can be made from three perspectives: increasing data quality, increasing tolerance to data quality issues during learning, and increasing learning ability. The last two perspectives can be addressed in machine learning by using neural network. Neural networks have remarkable capability to derive meaning from imprecise or complicated data though suffering from a major disadvantage of difficulty in interpretation.

In this paper, we use neural networks to learn doctors’ medicine prescription patterns in chronic disease treatment by mining EHR data to benefit from neural network’s properties of robustness to imprecise data and capacity to learn
complex tasks. Blood glucose control medication and blood pressure control medication in type 2 diabetes patients were considered. Seven types of anti-diabetic agents were considered for blood glucose control: alpha-glucosidase inhibitors (AGIs), dipeptidyl peptidase-4 (DPP-4), thiazolidinediones (TZDs), biguanides, glinides, sulfonylureas, and insulin. Seven types of antihypertensive drugs were considered for blood pressure control: angiotensin-converting-enzyme inhibitors (ACEIs), alpha blockers, angiotensin II receptor blockers (ARBs), beta blockers, calcium channel blockers (CCBs), diuretics, and hydrochlorothiazide. In a typical medicine prescription for blood glucose control, a doctor can prescribe one to several of the seven kinds of drugs. Thus, to learn doctors’ prescription behavior, prescribing a certain kind of anti-diabetic agent or not can be formulated into a binary classification problem, resulting in seven binary classifiers that can each be modeled by a neural network. In our observation, the prescription of the seven drugs are not mutually independent. To benefit from their commonalities while preserving the differences, a multi-task learning approach was used to learn the prescription combination of the seven drugs. Multi-task learning has been applied in healthcare settings to solve various problems like disease risk prediction, disease onset prediction, and disease progression for its good representation learning capability. In anti-diabetic agent prescription prediction, our multi-task neural network has shown AUC increases of 0.02 to 0.06 compared to logistic regression in the prediction of each single drug, and an accuracy increase of 0.05 in the prescription pattern prediction. In antihypertensive drug prescription prediction, the multi-task neural network has shown AUC increases of 0.02 to 0.03 compared to logistic regression in single drug prediction, though no accuracy increase in prescription pattern prediction was observed.

Our contribution is that we used a multi-task neural network approach to learn doctors’ medicine prescription pattern of multiple correlated drugs in blood glucose control and blood pressure control scenarios by mining EHR data, and has demonstrated its usefulness in solving these prescription problems.

Methods

Scenarios. Two scenarios were considered in this work: learning anti-diabetic agent prescription for blood glucose control and learning antihypertensive drug prescription for blood pressure control. Data was each constructed and processed.

Data preparation. Deidentified EHR data of type 2 diabetes patients from Xiamen, China within a four-year period (August 2012 to April 2016) was used in this study. Each case is defined as a patient encounter that has a prescription of at least one of the drugs concerned. Features used are summarized in Table 1, including demographics, physical examinations, laboratory tests, recent prescriptions, and disease history. The prescription of each drug is concerned as a class label, resulting in seven class labels in each of the two scenarios.

Data imputation. In this work, missing values of categorical variables were imputed by randomly sampling a category from a discrete probability distribution calculated from non-missing values, while missing values of continuous variables were imputed by randomly sampling a number from a normal distribution fit with non-missing values.

Feature engineering. Demographics were first extracted and imputed. For physical examinations, the most recent ones before the patient encounter were used if multiple sets were available. The immediate ones after this patient encounter were used in the absence of previous ones. Missing values were imputed. In antihypertensive drug prescription prediction, cases with imputed blood pressures were excluded. Two laboratory tests were used for anti-diabetic agent prescription since blood glucose level is a major indicator of such prescriptions, but were not used for antihypertensive drug prescription due to the rare availability of these tests. Test results within three months before or after the case were taken as valid. Only cases with valid test results were used for the anti-diabetic agent prescription prediction. Disease history was constructed using ICD codes, where the ICD-10 coding hierarchy was collapsed to the first three characters. A case would be assigned a ‘1’ for a disease if the corresponding ICD code was observed within the past year.

Data summary. After feature engineering, only 4,480 cases remained for anti-diabetic agent prescription prediction due to vast missing of laboratory test values, while 153,470 cases were preserved for antihypertensive drug prescription prediction. Antidiabetic agent prescription frequency and antihypertensive drug prescription frequency were each summarized in Table 2 and Table 3. Here, the anti-diabetic agent prescription frequency is significantly different from that in the whole population before removing cases, as is evident from the high frequency of insulin use. This reflects a
Table 1: Summary of features used in doctor’s medicine prescription pattern learning.

<table>
<thead>
<tr>
<th>Category</th>
<th>Attribute</th>
<th>Type</th>
<th>Antidiabetic agent prescription</th>
<th>Antihypertensive drug prescription</th>
</tr>
</thead>
<tbody>
<tr>
<td>Demographics</td>
<td>Gender</td>
<td>Categorical</td>
<td>Included</td>
<td>Included</td>
</tr>
<tr>
<td></td>
<td>Age</td>
<td>Continuous</td>
<td>Included</td>
<td>Included</td>
</tr>
<tr>
<td></td>
<td>Ethnic group</td>
<td>Categorical</td>
<td>Included</td>
<td>Included</td>
</tr>
<tr>
<td></td>
<td>Education</td>
<td>Categorical</td>
<td>Included</td>
<td>Included</td>
</tr>
<tr>
<td></td>
<td>Marriage</td>
<td>Categorical</td>
<td>Included</td>
<td>Included</td>
</tr>
<tr>
<td></td>
<td>Occupation</td>
<td>Categorical</td>
<td>Included</td>
<td>Included</td>
</tr>
<tr>
<td>Physical examinations</td>
<td>Height</td>
<td>Continuous</td>
<td>Included</td>
<td>Included</td>
</tr>
<tr>
<td></td>
<td>Weight</td>
<td>Continuous</td>
<td>Included</td>
<td>Included</td>
</tr>
<tr>
<td></td>
<td>BMI</td>
<td>Continuous</td>
<td>Included</td>
<td>Included</td>
</tr>
<tr>
<td></td>
<td>Blood pressure, systolic</td>
<td>Continuous</td>
<td>Included</td>
<td>Included</td>
</tr>
<tr>
<td></td>
<td>Blood pressure, diastolic</td>
<td>Continuous</td>
<td>Included</td>
<td>Included</td>
</tr>
<tr>
<td>Laboratory tests</td>
<td>Serum creatinine level</td>
<td>Continuous</td>
<td>Included</td>
<td>Not included</td>
</tr>
<tr>
<td></td>
<td>Fasting plasma glucose level</td>
<td>Continuous</td>
<td>Included</td>
<td>Not included</td>
</tr>
<tr>
<td>Recent prescriptions</td>
<td>Antidiabetic agents</td>
<td>Boolean</td>
<td>AGIs</td>
<td>ACEIs</td>
</tr>
<tr>
<td></td>
<td>or Diabetic agents</td>
<td></td>
<td>DPP-4</td>
<td>Alpha blockers</td>
</tr>
<tr>
<td></td>
<td>or Antihypertensive drugs</td>
<td></td>
<td>TZDs</td>
<td>ARBs</td>
</tr>
<tr>
<td></td>
<td>or Antihypertensive drugs</td>
<td></td>
<td>Biguanides</td>
<td>Beta blockers</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Glinides</td>
<td>CCBs</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Sulfonylureas</td>
<td>Diuretics</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Insulin</td>
<td>Hydrochlorothiazide</td>
</tr>
<tr>
<td>Disease history</td>
<td>ICD codes</td>
<td>Boolean</td>
<td>862 codes</td>
<td>1,252 codes</td>
</tr>
</tbody>
</table>

bias that cases with blood glucose levels recorded in the system may have more serious conditions compared to others.

**Drug dependency study.** From statistics’ point of view, prescription of two drugs A and B are independent if their joint probability equals the product of their probabilities:

\[ P(A \cap B) = P(A)P(B) \]

The dependency of drug A and B in this work is thus measured as:

\[ dependency = \frac{P(A \cap B)}{P(A)P(B)} \]

It can be rewritten with conditional probabilities as:

\[ dependency = \frac{P(A|B)}{P(A)} = \frac{P(B|A)}{P(B)} \]

This is equivalent to the support A provides for B, or the support B provides for A in the Bayesian interpretation, where a value larger than 1 indicates positive correlation, while a value smaller than 1 indicates negative correlation.

**Baseline classifier.** Logistic regression (LR) was used as the baseline classifier. The prescription of each drug is modeled by an LR, resulting in seven LR models in each scenario.

**Multi-task learning.** Multi-task neural networks were constructed as in Figure 1 for multi-task learning (MTL). Two intermediate components were illustrated: representation learning layers and classifier layers. A consensus representation relevant to the prescription tasks was learned in the representation learning layers. Out of the consensus representation, classifier layers were constructed for each classification task to make a distinction between the tasks while optimizing jointly a cost function considering the cost of all classification tasks. After training, the MTL model can intake the features and output the prediction of the seven drugs at the same time.
### Table 2: Antidiabetic agents prescription frequency.

<table>
<thead>
<tr>
<th>Drug</th>
<th>Prescription (count)</th>
<th>Prescription (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGIs</td>
<td>500</td>
<td>11.16%</td>
</tr>
<tr>
<td>DPP-4</td>
<td>9</td>
<td>0.20%</td>
</tr>
<tr>
<td>TZDs</td>
<td>183</td>
<td>4.08%</td>
</tr>
<tr>
<td>Biguanides</td>
<td>1,067</td>
<td>23.82%</td>
</tr>
<tr>
<td>Glinides</td>
<td>234</td>
<td>5.22%</td>
</tr>
<tr>
<td>Sulfonylureas</td>
<td>741</td>
<td>16.54%</td>
</tr>
<tr>
<td>Insulin</td>
<td>2,435</td>
<td>54.35%</td>
</tr>
</tbody>
</table>

### Table 3: Antihypertensive drugs prescription frequency.

<table>
<thead>
<tr>
<th>Drug</th>
<th>Prescription (count)</th>
<th>Prescription (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACEIs</td>
<td>15,758</td>
<td>10.27%</td>
</tr>
<tr>
<td>Alpha blockers</td>
<td>2,365</td>
<td>1.54%</td>
</tr>
<tr>
<td>ARBs</td>
<td>52,698</td>
<td>34.34%</td>
</tr>
<tr>
<td>Beta blockers</td>
<td>29,855</td>
<td>19.45%</td>
</tr>
<tr>
<td>CCBs</td>
<td>90,068</td>
<td>58.69%</td>
</tr>
<tr>
<td>Diuretics</td>
<td>6,174</td>
<td>4.02%</td>
</tr>
<tr>
<td>Hydrochlorothiazide</td>
<td>9,401</td>
<td>6.13%</td>
</tr>
</tbody>
</table>

### Figure 1: A schematic representation of multi-task neural network. For each instance, a set of features and several class labels for the tasks are used for training. Input layer is followed by representation learning layers to learn the common representations, then followed by multiple classifiers to deal with the distinctions, after which an output layer is constructed for each classifier.

**Threshold selection.** For both LR and MTL, the output is produced as the probability of a case to be positive (or ‘1’). A threshold translating a probability to a class was such selected as to maximize the F1 score (explained in Table 4).

**Method assessment.** In a typical run of the experiment, the input data is first split into training data (70%) and testing data (30%). Either LR or MTL is trained on training data, with its performance assessed on testing data. For LR, 50 runs were conducted in each scenario with different training/testing splits to account for split randomness. For MTL, 150 runs were conducted for each task, 3 runs each on the 50 different training/testing splits to account also for the randomness in neural network training. A list of measurements used in this work is summarized in Table 4. The performance of a classifier for single drug prescription prediction is evaluated by AUC. The performance in combined prescription prediction is firstly measured by average Jaccard similarity between the predicted and the actual prescription pattern. The performance in combined prescription is also measured by prediction accuracy, where a predicted prescription pattern is counted accurate only if single drug prediction is accurate for all seven drugs.

**Statistical analysis.** Prediction by MTL is called AI (artificial intelligence) prescription in contrary to doctor prescription. To explore the justification for AI prescription, we conducted statistical analysis to evaluate the importance of each disease in disease history to the prediction. The importance is assessed for each disease in the prescription of each drug, where having this disease or not is counted towards having this drug prescribed or not. Odds ratios and Wald confidence intervals were calculated as in Table 4. A disease is considered significant in the prescription of this disease when the confidence interval does not include 1.
Table 4: Measurements used in this work.

<table>
<thead>
<tr>
<th>Measurement</th>
<th>Description</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>True positive rate (TPR)</td>
<td>The proportion of positives correctly identified as such.</td>
<td>$TPR = \frac{\text{true positives}}{\text{all positive cases}}$</td>
</tr>
<tr>
<td>False positive rate (FPR)</td>
<td>The proportion of negatives wrongly categorized as positive.</td>
<td>$FPR = \frac{\text{false positives}}{\text{all negative cases}}$</td>
</tr>
<tr>
<td>Receiver operating characteristic (ROC) curve</td>
<td>A curve created by plotting the TPR against the FPR at various thresholds.</td>
<td></td>
</tr>
<tr>
<td>Area under the curve (AUC)</td>
<td>Area under the ROC curve.</td>
<td></td>
</tr>
<tr>
<td>Precision</td>
<td>The proportion of predicted positives that are real positives.</td>
<td>$\text{Precision} = \frac{\text{true positives}}{\text{all predicted positives}}$</td>
</tr>
<tr>
<td>Recall</td>
<td>The same as TPR.</td>
<td>$\text{Recall} = \frac{\text{true positives}}{\text{all positive cases}}$</td>
</tr>
<tr>
<td>F1 score</td>
<td>A measure of a binary classifier’s accuracy.</td>
<td>$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$</td>
</tr>
<tr>
<td>Jaccard similarity</td>
<td>A measure of similarity between two finite sample sets.</td>
<td>$Jaccard(A, B) = \frac{</td>
</tr>
</tbody>
</table>
| Odds ratio (OR)                   | A measure of how strongly the presence one property (A) is associated with the presence of another (B). | $n_{11} = |A \cap B|, n_{00} = |\neg A \cap \neg B|$  
$n_{10} = |A \cap \neg B|, n_{01} = |\neg A \cap B|$ 
$OR = \frac{n_{00} \cdot n_{11}}{n_{10} \cdot n_{01}}$ |
| Wald confidence interval          | Confidence interval of the OR.                                              | $SE = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{10}} + \frac{1}{n_{00}} + \frac{1}{n_{01}}}$  
$Lower 95\% CI = \exp(\log(OR) - SE)$ 
$Upper 95\% CI = \exp(\log(OR) + SE)$ |

Results

Drug dependency. In antidiabetic agent prescription, around 13% cases have more than one drug prescribed (Table 2). Dependency of antidiabetic agents is shown in Table 5. Many drug pairs show dependency other than 1, suggesting correlation between the drugs. As an example, the support glinides provides for AGIs is 2.46, suggesting elevated level of combined use. Also, the support sulfonylureas provides for insulin is 0.28, suggesting less combined use. DPP-4 has some large dependency values, probably a bias caused by its rare use. Similarly as in antihypertensive drug prescription, around 28% cases have more than one drug prescribed (Table 3). Dependency of antihypertensive drugs are summarized in Table 6, showing many dependency other than 1. These suggest the relatedness in prescribing the drugs, forming the basis for multi-task learning to model information sharing.

Table 5: Dependency of antidiabetic agent prescription.

<table>
<thead>
<tr>
<th></th>
<th>AGIs</th>
<th>DPP-4</th>
<th>TZDs</th>
<th>Biguanides</th>
<th>Glinides</th>
<th>Sulfonylureas</th>
<th>Insulin</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGIs</td>
<td>3.26</td>
<td>1.18</td>
<td>0.80</td>
<td>2.46</td>
<td>1.42</td>
<td>0.89</td>
<td></td>
</tr>
<tr>
<td>DPP-4</td>
<td></td>
<td>4.08</td>
<td>1.07</td>
<td>8.35</td>
<td>1.11</td>
<td>0.83</td>
<td></td>
</tr>
<tr>
<td>TZDs</td>
<td>1.18</td>
<td>4.08</td>
<td>1.53</td>
<td>1.61</td>
<td>1.32</td>
<td>0.62</td>
<td></td>
</tr>
<tr>
<td>Biguanides</td>
<td>0.80</td>
<td>1.07</td>
<td>1.53</td>
<td>0.70</td>
<td>1.16</td>
<td>0.38</td>
<td></td>
</tr>
<tr>
<td>Glinides</td>
<td>2.46</td>
<td>8.35</td>
<td>1.61</td>
<td>0.70</td>
<td>0.59</td>
<td>1.79</td>
<td></td>
</tr>
<tr>
<td>Sulfonylureas</td>
<td>1.42</td>
<td>1.11</td>
<td>1.32</td>
<td>1.16</td>
<td>0.59</td>
<td>0.28</td>
<td></td>
</tr>
<tr>
<td>Insulin</td>
<td>0.89</td>
<td>0.83</td>
<td>0.62</td>
<td>0.38</td>
<td>1.79</td>
<td>0.28</td>
<td></td>
</tr>
</tbody>
</table>
Table 6: Dependency of antihypertensive drug prescription.

<table>
<thead>
<tr>
<th></th>
<th>ACEIs</th>
<th>Alpha blockers</th>
<th>ARBs</th>
<th>Beta blockers</th>
<th>CCBs</th>
<th>Diuretics</th>
<th>Hydrochlorothiazide</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACEIs</td>
<td>0.56</td>
<td>0.87</td>
<td>2.04</td>
<td>1.55</td>
<td>0.83</td>
<td>0.75</td>
<td></td>
</tr>
<tr>
<td>Alpha blockers</td>
<td>0.56</td>
<td>1.21</td>
<td>1.19</td>
<td>0.67</td>
<td>2.68</td>
<td>1.73</td>
<td></td>
</tr>
<tr>
<td>ARBs</td>
<td>0.87</td>
<td>1.21</td>
<td>1.59</td>
<td>1.50</td>
<td>0.99</td>
<td>3.42</td>
<td></td>
</tr>
<tr>
<td>Beta blockers</td>
<td>2.04</td>
<td>1.19</td>
<td>1.59</td>
<td>1.90</td>
<td>1.36</td>
<td>1.01</td>
<td></td>
</tr>
<tr>
<td>CCBs</td>
<td>1.55</td>
<td>0.67</td>
<td>1.50</td>
<td>1.90</td>
<td>0.51</td>
<td>0.82</td>
<td></td>
</tr>
<tr>
<td>Diuretics</td>
<td>0.83</td>
<td>2.68</td>
<td>0.99</td>
<td>1.36</td>
<td>0.51</td>
<td>8.46</td>
<td></td>
</tr>
<tr>
<td>Hydrochlorothiazide</td>
<td>0.75</td>
<td>1.73</td>
<td>3.42</td>
<td>1.01</td>
<td>0.82</td>
<td>8.46</td>
<td></td>
</tr>
</tbody>
</table>

Performance comparison in antidiabetic agent prescription prediction: MTL vs LR. In LR, the prescription pattern is predicted by constructing an LR classifier for each drug and joining the predictions from them. MTL is capable of predicting prescription pattern together. Their performances are compared in Table 7. AUC is used to evaluate the prediction of each single drug, with its mean, standard deviation (SD), and maxima in all runs listed. For all the seven drugs, MTL has higher AUC than LR, with the highest increase in glinides, a mean AUC increase from 0.788 to 0.850. In the prediction of combined prescription, MTL has a mean Jaccard similarity of 0.600 compared to 0.569 achieved by LR. In terms of accuracy (percentage of predictions that have all seven drugs correctly predicted), MTL has an accuracy of 0.493 compared to 0.445 of LR.

Table 7: Performance comparison of LR and MTL in antidiabetic agent prescription prediction.

<table>
<thead>
<tr>
<th></th>
<th>Single drug: AUC</th>
<th>Combined prescription</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AGIs</td>
<td>DPP-4</td>
</tr>
<tr>
<td>LR</td>
<td>Mean 0.756</td>
<td>0.878</td>
</tr>
<tr>
<td></td>
<td>SD 0.020</td>
<td>0.214</td>
</tr>
<tr>
<td></td>
<td>Max 0.802</td>
<td>1</td>
</tr>
<tr>
<td>MTL</td>
<td>Mean 0.793</td>
<td>0.903</td>
</tr>
<tr>
<td></td>
<td>SD 0.021</td>
<td>0.089</td>
</tr>
<tr>
<td></td>
<td>Max 0.857</td>
<td>1</td>
</tr>
</tbody>
</table>

Performance comparison in antihypertensive drug prescription prediction: MTL vs LR. Similar comparison as in antidiabetic agent prescription prediction was conducted for antihypertensive drug prescription prediction, with performances summarized in Table 8. In single drug prediction, MTL has about 0.02 increase in AUC compared to LR in all seven drugs. However, MTL and LR have similar performance in combined prescription prediction, both mean Jaccard similarity and accuracy. This may be explained that though AUC is an important measure of binary classifier performance, it is difficult to be translated to a measure while solving a real-life problem since a threshold is needed to make a prediction based on the probability given. In this work, we selected the threshold that maximized the F1 score, bearing our intension to balance precision and recall. Under such selection, the accuracy of combined prescription prediction may well be the same between MTL and LR.

Disease history relevant to insulin prescription. Disease onsets (denoted by ICD codes) within a year from the time of case were evaluated for their relevance to insulin prescription both by doctor and by AI. A list of ICD codes was identified as significant for doctor prescription and AI prescription. We first tried to figure out AI’s ability in capturing doctors’ intension for insulin use. AI made a successful capture if an ICD code significant to doctors’ prescription was also significant to AI’s prescription. We thus present Figure 2 for visualization. From the figure, the majority of the diseases significant to doctors’ prescription were also identified as significant in AI prescription (OR >1 in Figure 2A or OR <1 in Figure 2B). A few exceptions were identified and summarized in Table 9. Out of the 121 ICD codes significant to increased doctor insulin prescription (OR >1), 15 (12.4%) were missed by AI prescription,
Table 8: Performance comparison of LR and MTL in antihypertensive drug prescription prediction.

<table>
<thead>
<tr>
<th></th>
<th>ACEIs Mean</th>
<th>ACEIs SD</th>
<th>ACEIs Max</th>
<th>Alpha blockers Mean</th>
<th>Alpha blockers SD</th>
<th>Alpha blockers Max</th>
<th>ARBs Mean</th>
<th>ARBs SD</th>
<th>ARBs Max</th>
<th>Beta blockers Mean</th>
<th>Beta blockers SD</th>
<th>Beta blockers Max</th>
<th>CCBs Mean</th>
<th>CCBs SD</th>
<th>CCBs Max</th>
<th>Hydrochlorothiazide Mean</th>
<th>Hydrochlorothiazide SD</th>
<th>Hydrochlorothiazide Max</th>
<th>LR Mean</th>
<th>LR SD</th>
<th>LR Max</th>
<th>MTL Mean</th>
<th>MTL SD</th>
<th>MTL Max</th>
<th>Mean Jaccard similarity</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single drug: AUC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Single drug: AUC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Combined prescription</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

suggesting AI’s ability in capturing most diseases. However, some potentially important diseases were missed, like high blood pressure, lipid disorder, and diseases related to eye, limb, and bone, raising more expectations on AI methods. Out of the 47 ICD codes significant to decreased doctor insulin prescription (OR <1), 10 (21.3%) were not significant to AI prescription, suggesting AI has denied the association between some diseases and decreased insulin use. Some of these diseases may not be related to decreased insulin use, like embedded and impacted teeth (K01), which suggests AI’s capability in reducing impact of less related features. We also tried to figure out AI’s inherent logic in prediction after learning from doctors’ prescription. AI prescription identified 147 ICD codes associated with increased insulin prescription, and 97 ICD codes associated with decreased insulin prescription. This number is greater than in doctor prescription, indicating that AI is inflating the importance of features in prediction to make a greater distinction between different predictions. Diseases significant only in AI prescription are listed in Table 9. However, the interpretation of the results are complicated, which are discussed in the discussion section.

Table 9: Comparison of the differential disease history-insulin prescription associations between doctor prescription and AI prescription.

<table>
<thead>
<tr>
<th></th>
<th>Odds ratio &gt; 1</th>
<th>Odds ratio &lt; 1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Significant only in doctor prescription</td>
<td>Significant only in AI prescription</td>
</tr>
<tr>
<td>Diabetes</td>
<td>E13</td>
<td></td>
</tr>
<tr>
<td>Heart-related</td>
<td>I50, I51</td>
<td></td>
</tr>
<tr>
<td>Cardiovascular disease</td>
<td>G45, I63, I64</td>
<td></td>
</tr>
<tr>
<td>Eye-related</td>
<td>H20</td>
<td></td>
</tr>
<tr>
<td>Kidney disease</td>
<td>N04, N05, N28</td>
<td></td>
</tr>
<tr>
<td>Limb-related</td>
<td>L97</td>
<td></td>
</tr>
<tr>
<td>Bone-related</td>
<td>M81</td>
<td>M24, M75, M77, M91, S52, S63</td>
</tr>
<tr>
<td>Infection</td>
<td>B35, J15</td>
<td>M19, M25, M47, M48, M51</td>
</tr>
<tr>
<td>High blood pressure</td>
<td>I10</td>
<td></td>
</tr>
<tr>
<td>Lipid disorder</td>
<td>E75, E78</td>
<td></td>
</tr>
<tr>
<td>Neoplasm</td>
<td>C20, C61, Z85</td>
<td>D23</td>
</tr>
<tr>
<td>Local codes</td>
<td>AVY</td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>D72, I84, J01, K04, K56, L73, L98, N30</td>
<td>D84, E03, E79, E87, G12, K08, K26, K58, K74, K81, L29, L30, M10, M32, M33, M35, M60, R04, R06, R21, R42, R53, S39, Z01, Z71</td>
</tr>
</tbody>
</table>

1817
Figure 2: Characterizing ICD codes significant to doctor prescription of insulin and their corresponding significance in AI prescription. Odds ratios and confidence intervals of ICD codes that are significant to doctor prescription of insulin are plotted in A (odds ratio $>1$) and B (odds ratio $<1$). Each pink row represents an ICD code’s statistics, with the odds ratio plotted as a black dot and confidence interval plotted as a pink line. The blue line immediately below a pink line denotes the ICD code’s odds ratio (black dot) and confidence interval (blue line) in AI prescription of insulin. The red vertical lines represent the line of $x=1$. In A, confidence intervals beyond 75 are not explicitly plotted.

Discussion

Application of AI methods in clinical practices. While AI may never be able to act the role of human doctors completely, it can learn from doctors’ behavior to provide support for further actions. In this work, we present an effort toward learning doctor medicine prescription patterns. The potential application of such efforts in clinical practices is supporting doctors who have less experience, providing them with clinical decisions learned from other doctors’ experience. We are aware that a major critique of such AI methods is a lack of interpretability compared to clinical guidelines that can explicitly be formatted into a series of if-then rules.

Data quality. As mentioned in the introduction section, data quality is a limiting factor of the application of data-driven CDSS. Data quality issue in our work can be viewed from two perspectives. First, missing data. We encountered plenty of features with missing values in EHR. Mostly, we either imputed the missing values or discarded the feature. However, this is not a proper choice for features vital to drug prescription, blood glucose levels, for example. In our antidiabetic agent prescription problem, cases without blood glucose levels tested were discarded, resulting in extremely reduced number of cases. Second, it may be problematic to learn from doctors’ prescriptions without assessing effectiveness. The prescription of one doctor may not be proper, or may not be the only proper prescription. Some drugs may have highly overlapping functions and the choice is highly doctor-specific or hospital-specific. Under such circumstances, AI can hardly tell the differential use of the drugs. And unfairly in performance evaluation, the prediction of AI may be judged as wrong if it predicts a drug while the doctor used its counterpart.
Disease history as features for prediction. The use of demographics, physical examinations, laboratory tests and recent prescriptions in learning doctors’ prescription is less debatable as they are more or less self-evident. However, how to make use of disease history remains an open question. While many traditional studies pick only a few clinically relevant diseases in analysis, we include nearly all, benefiting from big data approaches’ merit of being able to learn feature importance and in the hope of finding hidden factors to better explain doctors’ behavior. Here we collapsed the ICD-10 code hierarchy to the first three characters since in real-life settings, the more detailed the ICD code goes, the less accurate it is. However, this collapse from the nomenclature point of view may not actually capture the hierarchy and relatedness of diseases, which, in itself, is a complex task. Moreover, since ICD-10 contains codes for not only diseases, but also signs and symptoms, abnormal findings, complaints, social circumstances, and external causes of injury or diseases, it is a rather mixed representation, which is currently not treated with special care.

Caveats of disease relevance to drug prescription. We have discussed about disease history relevant to insulin prescription in the results section. Several ICD codes were associated with increased or decreased insulin prescription either by doctor or by AI. However, the interpretation of the association should be made with caution. Some diseases associated with increased insulin use are potential complications of type 2 diabetes (N18: chronic kidney disease, and N19: unspecified kidney failure, for example), thus can be used as an indication of exacerbating disease condition and an indication of insulin use. Some diseases negatively associated with insulin use may not be a reason not to use insulin but a result of no previous insulin use. Taking elevated blood glucose level (R73) as such an example. Previous insulin use can decrease the onset of elevated blood glucose level, making the onset of it an indication of no previous insulin use and thus an indication of no insulin prescription. Some may have rather complex reasons. Vitamin B12 deficiency anemia (D51) is significantly correlated with decreased insulin use in both doctor and AI prescription. Existing studies have well demonstrated that long term metformin treatment increases the risk of vitamin B12 deficiency\textsuperscript{11}. This is also observed in our data that D51 is strongly associated with increased biguanides use (OR=12.88, 95% CI = [2.73, 60.76]). Besides, in our drug dependency study, use of biguanides decrease the use of insulin. These altogether lead D51’s association with decreased insulin use. As the same with all association studies, it is hard to infer causation from such results. After all, it is not necessary for big data analytics to know causation to infer decision.

Conclusion

In this paper, we use a multi-task neural network approach to learn doctor medicine prescription pattern of multiple correlated anti diabetic agents in blood glucose control and antihypertensive drugs in blood pressure control scenarios by mining EHR data. We have demonstrated that for single drug prescription, multi-task learning achieves 0.02 to 0.06 increases in AUC compared with logistic regression. In prescription pattern prediction, we achieved an accuracy increase of 0.05 in antidiabetic agent prediction, showing the efficacy of multi-task learning approach in learning doctor prescription patterns.

References


Predicting Changes in Pediatric Medical Complexity using Large Longitudinal Health Records

Yanbo Xu, MS¹, Mohammad Taha Bahadori, PhD¹, Elizabeth Searles, RN, MBA², Michael Thompson, BS, MBA², Javier Tejedor-Sojo, MD², Jimeng Sun, PhD¹
¹Georgia Institute of Technology; ²Childrens Healthcare of Atlanta

Abstract

Medically complex patients consume a disproportionate amount of care resources in hospitals but still often end up with sub-optimal clinical outcomes. Predicting dynamics of complexity in such patients can potentially help improve the quality of care and reduce utilization of hospital resources. In this work, we model the change prediction of medical complexity using a large dataset of 226K pediatric patients over 5 years from Children's Healthcare of Atlanta (CHOA). We compare different classification methods including logistic regression, random forest, gradient boosting trees, and multilayer perceptron in predicting whether patients will change their complexity status in the last year based on the data from previous years. We achieved an area under the ROC curve (AUC) of 88% for predicting non-complex patients becoming complex and 74% for predicting complex patients staying complex. We also identify the factors associated with the change in complexity of patients.

1 Introduction

Medically complex or fragile children need intense medical care due to multisystem dysfunction, technology dependence, or complex medication needs¹. It is estimated that children with special health care needs constitute about 18% (near 12.6 million) of all US children². Given the large portion of complex children in consumption of hospital resources³, the Institute of Medicine has identified Children with Special Health Care Needs (CSHCN) as a priority population for study⁴.

Medically complex pediatric patients consume a significant disproportionate amount of care resources in hospitals³. For example, Simon et al.³ report that complex patients “accounted for 8.9% of US pediatric admissions in 1997 and 10.1% of admissions in 2006. These admissions used 22.7% to 26.1% of pediatric hospital days, used 37.1% to 40.6% of pediatric hospital charges, accounted for 41.9% to 43.2% of deaths, and (for 2006) used 73% to 92% of different forms of technology-assistance procedures.” Advances on several fronts of research in medically complex children have improved outcomes, increasing survival rate and decreasing healthcare cost¹, 5.

With the massive amount of Electronic Health Record (EHR) data collected at Children’s Healthcare of Atlanta (CHOA), it is possible to develop predictive models to gain insights into underlying complexity risks. It is important to identify high risk patients who are rapidly becoming complex so that early intervention could be applied. Furthermore, identifying the key risk factors contributing to the complexity of the patients could lead to a better understanding of medical complexity.

Statistical models such as logistic regression have been used to identify the medical codes that are associated with complexity¹, 6–10. Often, manual features are also constructed to help the predictive analysis⁷. Logistic regression¹¹ and decision trees¹² have been used for quantifying the odds of readmission of medically complex patients. Performance for other machine learning algorithms such as methods based on sparse logistic regression and boosting have also been assessed for predicting readmission¹³, and more involved models such as hierarchical mixed-effects logistic regression models have been used to study contributing factors to medical complexity¹⁴.

In contrast to majority of the previous works which study the factors associated with being complex, in this work we study the dynamics of complexity in pediatric setting. That is, we develop predictive models to accurately forecast the change in complexity for children using a large EHR dataset from Children’s healthcare of Atlanta. To this end, we design two cohorts: (i) a dataset including patients that are initially not complex and either stayed non-complex or become complex, and (ii) a dataset with patients that are complex in the initial period. To formally quantify medically complexity, we use three criteria: (i) the clinical risk group’s definition of complexity (CRG level 5b or above), (ii) the hospital cost (top 5% high utilizers per year), and (iii) the number of specialities required for patients (3 or more...
specialities). We intuitively consider a patient complex if he or she satisfies all the above criteria.

Our methodology includes two steps to fulfill our prediction goal on the two cohorts. First, we use $L_1$ regularized Logistic regression to filter out irrelevant features and reduce feature dimensionality. Second, we apply four state-of-the-art predictive models, including logistic regression, random forests, gradient boosted decision trees, and multilayer perceptron. From our results, multilayer perceptron is found to be the best classifier (0.88) on cohort (i) in terms of ROC-AUC and random forest is found to be the best classifier (0.74) on cohort (ii). However, the gradient boosted trees method provides the best predictions on both cohorts (0.16 for (i) and 0.74 for (ii)) in terms of PRC-AUC, a metric that measures the accuracy in only predicting the change to complex status. Lastly, random forests gives the best $F_1$ score on both cohorts (0.56 for (i) and 0.67 for (ii)) by thresholding the predictive probabilities with 0.5.

For the rest of the paper, we will first describe the data collected from CHOA in Section 2, introduce our methodology in Section 3, report our results in Section 4 and finally conclude in Section 5.

2 Data

In this work, we analyze a large longitudinal EHR dataset from Children’s Healthcare of Atlanta (CHOA). We extract structured fields from the visit records in the dataset, which include various medical codes such as diagnosis, medication, and procedure. In total, the data consists of 951,885 patients with total number of 3,760,691 visits from year 2010 to 2015. Below we describe in details how we construct the cohort and the features for the purpose of our study.

2.1 Cohort Construction

We are interested in predicting the dynamics of medically complexity in patients. Specifically, we divide each patient’s visits into an observation window and a prediction window, then aim to predict if a patient who was not medically complex (MCP negative) in the observation window becomes complex (MCP positive) in the prediction window — and vice versa. Here, as shown on top of Figure 2, the prediction window is defined to be the 1-year duration before a patient’s last visit recorded in the data and the observation window is the 1-year duration before the prediction window.

![Figure 1: Medically complex patients defined within the observation window in the cohort](image)

We consider medically complex patients as the group of patients who have more serious health conditions, need more medical specialty services, and therefore incur higher medical charges. In the CHOA dataset, the severity level of a patient’s health condition can be captured by the pre-defined Clinical Risk Grouping (CRG) score, ranged in {1, 2, 3, 4, 5a, 5b, 6, 7, 8, 9} with higher orders denoting higher severity levels. For specialties, as listed in Figure 1(b), we only consider the 10 specialties that were evaluated by CHOA as essential services to Multiple Chronic Condition
Therefore, to construct the cohort, we include the patients who have at least one visit in the observation window and at least one record of CRG score, one visit of specialty and one hospital charge in both windows. As a result, we obtain a cohort containing 228,128 patients with total number of 2,150,256 visits.

**Figure 2:** Number of patients from the cohort in a 2-by-2 transition table of MCP status

**Definition of Medically Complex Patients:** Formally, we define medically complex patients (MCP) as the patients who have CRG $\geq 5$, number of specialties $\geq 3$, and total hospital charge $\geq$ $31,027$ (ranked in the top 5% for charges in the cohort) within 1-year windows. Figure 1(a) illustrates our MCP definition within the observation window. There are in total 5,094 MCP in the observation window and 5,378 in the prediction window.

In addition, Figure 2 shows the number of patients in the cohort that are contained in each group of MCP as they transition from the observation to prediction window. In summary, about 50% of the patients who were observed MCP positive become negative in the following year, and about 1.3% of the patients who were observed MCP negative become positive in the following year. More statistics of the cohort prior to the prediction window are summarized in Table 1.

**Table 1:** Basic statistics of the cohort.

<table>
<thead>
<tr>
<th></th>
<th>MCP Positive</th>
<th>MCP Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td># of patients</td>
<td>5,094</td>
<td>223,034</td>
</tr>
<tr>
<td># of visits</td>
<td>64,894</td>
<td>524,940</td>
</tr>
<tr>
<td>Avg. # of visits per patient</td>
<td>12.7</td>
<td>2.3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>1st Qu.</th>
<th>Median</th>
<th>3rd Qu.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>2.3</td>
<td>7.3</td>
<td>14.3</td>
</tr>
<tr>
<td>CRG</td>
<td>6</td>
<td>6</td>
<td>9</td>
</tr>
<tr>
<td>Number of specialties</td>
<td>3</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Total hospital charge</td>
<td>$54.0K</td>
<td>$98.0K</td>
<td>$215.4K</td>
</tr>
</tbody>
</table>

$2.2$ Feature construction

We extract demographic features per patient and medical codes per visit from the data. Demographic features include age, gender and race, where gender is binary, race is categorical with 6 categories, and age is continuous valued by the index time (years since birth to the beginning of the prediction window). Medical codes recorded in the data include ICD-9 diagnosis codes, Current Procedural Terminology (CPT) codes and PHARMA CLASS medication codes that are provided by CHOA. In our cohort, there are total number of 4,497,616 medical codes recorded prior to the prediction window, including 1,521,956 ICD-9 codes, 1,934,116 CPT codes, and 1,041,544 PHARMA CLASS codes. For the sake of dimension reduction, we group the raw diagnosis and procedure codes by using the Clinical
Classification Software (CCS) codes. As a result, the total number of unique codes reduce from 8,643 to 9,49. We list the top 10 most frequent raw codes and group codes that were prescribed for the MCP positive and MCP negative respectively in Table 2. More basic statistics of the features prior to the prediction window are summarized in Table 3.

Table 2: Top 10 most frequent codes prescribed for MCP positives vs. MCP negatives

<table>
<thead>
<tr>
<th>Raw Code</th>
<th>Description</th>
<th>Freq.</th>
<th>Raw Code</th>
<th>Description</th>
<th>Freq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPT 85025</td>
<td>Complete Blood Counts</td>
<td>16276</td>
<td>CPT 99283</td>
<td>Level 3 Under New or Established Patient Emergency Department Services</td>
<td>107289</td>
</tr>
<tr>
<td>CPT 99213</td>
<td>Level 3 Established Office Visit</td>
<td>14648</td>
<td>PHARMA 10</td>
<td>NSAIDS CYCLOOXYGENASE INHIBITOR - TYPE</td>
<td>79521</td>
</tr>
<tr>
<td>PHARMA 701</td>
<td>HEPARIN AND RELATED PREPARATIONS</td>
<td>14095</td>
<td>CPT 99213</td>
<td>Level 3 Established Office Visit</td>
<td>78582</td>
</tr>
<tr>
<td>CPT 99214</td>
<td>Level 4 Established Office Visit</td>
<td>12277</td>
<td>CPT 99214</td>
<td>Level 4 Established Office Visit</td>
<td>69228</td>
</tr>
<tr>
<td>PHARMA 1156</td>
<td>SODIUM/SALINE PREPARATIONS</td>
<td>11976</td>
<td>CPT 99284</td>
<td>Level 4 Under New or Established Patient Emergency Department Services</td>
<td>66415</td>
</tr>
<tr>
<td>PHARMA 1603</td>
<td>ANALGESICS/ANTIPYRETICS NON-SALICYLATE</td>
<td>11740</td>
<td>PHARMA 1603</td>
<td>ANALGESICS/ANTIPYRETICS NON-SALICYLATE</td>
<td>63668</td>
</tr>
<tr>
<td>PHARMA 11</td>
<td>ANTIEMETIC/ANTIVERTIGO AGENTS</td>
<td>11111</td>
<td>ICD9 780.60</td>
<td>Fever NOS</td>
<td>58132</td>
</tr>
<tr>
<td>CPT 99285</td>
<td>Level 5 Under New or Established Patient Emergency Department Services</td>
<td>9744</td>
<td>CPT 87880</td>
<td>Group A Rapid Strep Screen - IRL Only</td>
<td>47521</td>
</tr>
<tr>
<td>PHARMA 674</td>
<td>ANALGESICS NARCOTICS</td>
<td>9615</td>
<td>PHARMA 473</td>
<td>BETA-ADRENERGIC AGENTS</td>
<td>43385</td>
</tr>
<tr>
<td>PHARMA 477</td>
<td>TOPICAL LOCAL ANESTHETICS</td>
<td>8516</td>
<td>PHARMA 13</td>
<td>PENICILLINS</td>
<td>43123</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Group Codes</th>
<th>Description</th>
<th>Freq.</th>
<th>Group Codes</th>
<th>Description</th>
<th>Freq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCS_PR 227</td>
<td>Other diagnostic procedures (interview; evaluation; consultation)</td>
<td>192731</td>
<td>CCS_PR 227</td>
<td>Other diagnostic procedures (interview; evaluation; consultation)</td>
<td>1204519</td>
</tr>
<tr>
<td>CCS_PR 233</td>
<td>Laboratory - Chemistry and hematology</td>
<td>170108</td>
<td>CCS_PR 233</td>
<td>Laboratory - Chemistry and hematology</td>
<td>649223</td>
</tr>
<tr>
<td>CCS_PR 206</td>
<td>Microscopic examination (bacterial smear; culture; toxicology)</td>
<td>86179</td>
<td>CCS_PR 206</td>
<td>Microscopic examination (bacterial smear; culture; toxicology)</td>
<td>454464</td>
</tr>
<tr>
<td>CCS_PR 235</td>
<td>Other laboratory</td>
<td>72275</td>
<td>CCS_DX 126</td>
<td>Other upper respiratory infections</td>
<td>198305</td>
</tr>
<tr>
<td>CCS_PR 235</td>
<td>Other therapeutic procedures</td>
<td>52200</td>
<td>CCS_PR 235</td>
<td>Other laboratory</td>
<td>172837</td>
</tr>
<tr>
<td>CCS_DX 58</td>
<td>Other nutritional; endocrine; and metabolic disorders</td>
<td>23506</td>
<td>CCS_PR 231</td>
<td>Other therapeutic procedures</td>
<td>152714</td>
</tr>
<tr>
<td>CCS_DX 95</td>
<td>Other nervous system disorders</td>
<td>22070</td>
<td>CCS_PR 226</td>
<td>Other diagnostic radiology and related techniques</td>
<td>150081</td>
</tr>
<tr>
<td>CCS_DX 155</td>
<td>Other gastrointestinal disorders</td>
<td>21370</td>
<td>CCS_DX 246</td>
<td>Fever of unknown origin</td>
<td>133087</td>
</tr>
<tr>
<td>CCS_PR 217</td>
<td>Other respiratory therapy</td>
<td>21074</td>
<td>CCS_PR 213</td>
<td>Physical therapy exercises; manipulation; and other procedures</td>
<td>127505</td>
</tr>
<tr>
<td>CCS_PR 213</td>
<td>Physical therapy exercises; manipulation; and other procedures</td>
<td>20885</td>
<td>CCS_DX 133</td>
<td>Other lower respiratory disease</td>
<td>122270</td>
</tr>
</tbody>
</table>

Table 3: Basic statistics of the features in the cohort.

<table>
<thead>
<tr>
<th></th>
<th>MCP Positive</th>
<th>MCP Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td># of unique raw medical codes</td>
<td>8,154</td>
<td>11,742</td>
</tr>
<tr>
<td>- # of unique ICD-9 diagnosis codes</td>
<td>4,992</td>
<td>7,263</td>
</tr>
<tr>
<td>- # of unique CPT procedure codes</td>
<td>2,670</td>
<td>3,908</td>
</tr>
<tr>
<td>- # of unique PHARMA medication codes</td>
<td>489</td>
<td>571</td>
</tr>
<tr>
<td># of unique CCS group codes</td>
<td>460</td>
<td>490</td>
</tr>
<tr>
<td>- # of unique CCS diagnosis codes</td>
<td>260</td>
<td>275</td>
</tr>
<tr>
<td>- # of unique CCS procedure codes</td>
<td>200</td>
<td>215</td>
</tr>
</tbody>
</table>

Further, we count the encounters of each CCS and PHARMA code per patient preceding the prediction window. Since, for each patient, most of the codes appear less than 100 times (shown as the blue line in Figure 3) and rarely appear more than 200 times, we transform the counts into $\log_{10}$ space (shown as the orange line in Figure 3) and group them into 15 bins ranging from $\log_{10}(1)$ to $\log_{10}(200)$. Using this transformation over the original counts gives us a prediction gain of 2% in terms of ROC-AUC in general for all the prediction methods we describe in the next section.
3 Methodology

Our analysis has two steps: we perform feature selection (Section 3.1) followed by prediction (Section 3.2). For implementation, we use the publicly available machine learning toolkit scikit-learn\(^1\) in Python.

3.1 Feature Selection

Feature selection—i.e., filtering out the irrelevant factors from the predictive analysis—is a common approach in data analysis procedures. An efficient feature selection algorithm not only helps with interpretability of the results by excluding spurious factors, but it also reduces the dimensionality of the input and allows us to run more complex predictive models.

Feature selection algorithms can be divided to two categories: (i) Univariate feature selection techniques rank the features based on their relevance to the target label using a non-parametric score such as mutual information. They allow us to select the top \(k\) relevant features using the ranked list. (ii) Multivariate feature selection techniques assume a multivariate regression model (e.g., logistic regression) and impose sparsity inducing regularization such as \(L_1\) regularization to allow only a small fraction of features to be included in the model. Multivariate feature selection is usually more accurate than the univariate approach as it performs joint feature selection and prediction; thus, the ensemble of features selected by the multivariate approach are often more predictive of the target.

In this paper, we use \(L_1\) regularized logistic regression for multivariate feature selection. We control the number of selected features by varying the regularization parameter \(c\) ranging from 0.001 to 1.

3.2 Prediction

Given the features selected, we investigate four classification methods for two prediction tasks: (i) predicting among all the regular patients who will become medically complex patients, i.e., the change of status from MCP negative to MCP positive/negative (Neg2(Pos/Neg)), and (ii) predicting among existing MCP patients who will stay MCP and who will become non-MCP, i.e., the change of status from MCP positive to MCP positive/negative (Pos2(Pos/Neg)).

We start with a simple generalized linear model Logistic Regression (LR) and two ensemble classification methods Random Forest (RF) and Gradient Boosting Decision Tree (GBDT). The ensemble methods fit a number of weak classifiers and classify a new patient by taking a (weighted) average of their predictions. These methods usually make better predictions than a single method, and result at better performance when the number of weak learners increases. However, a large number of weak learners can lead to over-fitting. In our setting, we set the number of trees in RF as 300 for both tasks, and in GBDT as 300 for the smaller task Pos2(Neg/Pos) and 800 for the much larger task Neg2(Neg/Pos). In addition, we use 3-fold cross validation to grid search the hyper-parameters used in each method. Particularly, we use gini importance as the criteria for RF splitting, and grid search min sample split in

\(^1\)http://scikit-learn.org/stable/
Figure 4: Feature selection using $L_1$-Logistic regression for the two prediction tasks. Only small number of features (252 for Neg2(Pos/Neg) and 73 for Pos2(Pos/Neg)) are sufficient to construct great predictive models \(\{3, 5, 15, 18, 20\}\) and \(\text{min samples leaf} \in \{2, 5, 7, 10, 15\}\); we use a \textit{learning rate} of 0.01 for GBDT, and grid search the \textit{max depth} in \(\{1, 3, 5\}\). Moreover, there exists imbalance problem in the task Neg2(Pos/Neg): only 1.3\% of the negative patients change to be positive in the prediction window. To address this, we adjust the class weights inversely proportional to class frequencies in the training data.

\textit{Multilayer Perceptron (MLP)} is another approach to increase the learning capacity of the logistic regression. It is based on stacking layers of linear transformation followed by a non-linear activation function (rectified linear units in our case). MLP is able to transform the input data multiple times and learn a form of input that allows more accurate prediction of the target label. For training the neural network, we use both dropout\(^{21}\) and batch normalization\(^{22}\) and train the network using the Adamax\(^{23}\) stochastic optimization algorithm.

4 Results

To evaluate our methodology, we randomly pick 90\% of the cohort as training patients and the remaining 10\% as test patients. For task Neg2(Pos/Neg), this results at 200,581 patients for training and 22,453 patients for testing. For task Pos2(Pos/Neg), it gives 4,580 patients for training and 514 patients for testing. We report the prediction results in terms of three metrics: Receiver Operating Characteristic (ROC) curves and its area-under-curve (AUC), Precision-Recall Curves (PRC) and its area-under-curve (AUC), and $F_1$ score with thresholding 0.5 on the predictive probabilities. We report PRC-AUC because the Neg2(Pos/Neg) task is extremely challenging due to the imbalance and PRC-AUC provides a better metric for evaluating predictions on the positive class.

\textbf{Feature selection.} We first plot the ROC-AUC scores obtained from LR, RF, and GBDT given different values of the regularization parameter $c$ used by the $L_1$-Logistic regression in Figure 4. The number of features that are selected given different $c$’s are also provided at the bottom of the figure. We can see that the prediction methods can achieve optimal performance on both tasks by only selecting a small number of features, that is 252 features for Neg2(Pos/Neg) and 73 for Pos2(pos/Neg). Moreover, we investigate the CCS diagnosis codes from the top predictive features that are selected for predicting Pos2(Pos/Neg) (see the biclustering heat map of Figure 6 in Appendix). There we observe several interesting diagnosis groups including metabolic disorders, immunity disorders, epilepsy, and cancer.

\textbf{Prediction.} We report the optimal accuracy results (ROC-AUC, PRC-AUC, and $F_1$) obtained from the four classification methods for the two tasks in Table 4. Overall, the three more complicated methods outperform the single linear method LR: MLP gives the best ROC-AUC score for predicting Neg2(Pos/Neg) and RF gives the best score for predicting Pos2(Pos/Neg); GBDT gives the best PRC-AUC score for both tasks; and RF gives the best $F_1$ score for the two tasks. More specifically, we plot the ROC and PRC curves respectively in Figure 5.
Table 4: Prediction Results (Percentage ROC-AUC, PRC-AUC and $F_1$)

<table>
<thead>
<tr>
<th></th>
<th>Neg2(Pos/Neg)</th>
<th></th>
<th></th>
<th>Pos2(Pos/Neg)</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ROC-AUC</td>
<td>PRC-AUC</td>
<td>$F_1$</td>
<td>ROC-AUC</td>
<td>PRC-AUC</td>
<td>$F_1$</td>
</tr>
<tr>
<td>LR</td>
<td>86.90</td>
<td>12.65</td>
<td>54.14</td>
<td>69.03</td>
<td>67.98</td>
<td>63.75</td>
</tr>
<tr>
<td>RF</td>
<td>87.59</td>
<td>12.74</td>
<td><strong>56.40</strong></td>
<td><strong>74.11</strong></td>
<td>73.62</td>
<td><strong>66.73</strong></td>
</tr>
<tr>
<td>GBDT</td>
<td>86.88</td>
<td><strong>15.94</strong></td>
<td>52.54</td>
<td>74.06</td>
<td><strong>74.13</strong></td>
<td>66.70</td>
</tr>
<tr>
<td>MLP</td>
<td><strong>88.11</strong></td>
<td>15.14</td>
<td>50.48</td>
<td>71.52</td>
<td>68.92</td>
<td>65.53</td>
</tr>
</tbody>
</table>

Figure 5: Prediction Results (ROC and PRC)

Different metrics provide different perspectives for different purpose of evaluations. ROC evaluates the performance of the classifier with its discrimination threshold varied. Specifically, it measures the trade-off between *true positive rate* (TPR, known as *sensitivity* and *recall*)—the rate of MCP positive patients correctly classified as positive—and the *false positive rate* (FPR, known as 1-specificity)—the rate of MCP negative patients incorrectly classified as positive. An ROC curve indicates better performance (reported as a larger area under the curve (AUC)) when the curve gets closer to the top-left corner. However, sensitivity and specificity individually measure the correct classification rate within each group of the positive patients and the negative patients, so combining these two measurements has ignored the information between the two groups. As a result, the ROC curve can still obtain a high AUC score even if the classification result is not precise enough. In such cases, PRC becomes a better metric. Especially, when the purpose of classification is focused on identifying the MCP positive patients, who occupy only a small portion of the cohort (e.g., 1.3% in the Neg2(Neg/Pos) task), PRC becomes a more useful metric. Specifically, PRC measures the trade-off between *precision*—the rate of the classified MCP positive patients actually labeled as positive—and the *recall*. A PRC curve indicates a better performance (reported as a larger area under the curve (AUC)) when the curve gets closer to the top-right corner. Lastly, $F_1$ score is a similar measure as *accuracy* that requires a threshold on the prediction probabilities, but instead it calculates an harmonic mean of *precision* and *recall*. 
Note that we obtain consistently high ROC-AUC scores for predicting the Neg2(Neg/Pos) task due to the heavily skewed data (98.7% of the cohort are labeled MCP negative). That is, even though all classifiers get much more false positives than true positives (therefore they get low PRC-AUC scores), the FPR becomes small after being divided by the large denominator; similarly the TPR becomes large after being divided by a small denominator (therefore they get high ROC-AUC scores). Thus, for evaluating the Neg2(Neg/Pos) task, PRC-AUC becomes more useful since it provides the precision in predicting the positive events.

5 Conclusion

We study the dynamics of medically complexity in children by applying machine learning techniques on a massive observational dataset collected at Children’s Healthcare of Atlanta. We formalize two prediction tasks: (i) predicting the risk of a non-complex patient becoming complex, and (ii) predicting the risk of a complex patients staying complex. We first use a statistical model to automatically identify important medical codes for the two prediction tasks, and follow up with a comparison of four state-of-the-art classification models for completing the prediction. We report the results using three different metrics that can be used for different purposes in evaluation and demonstrate that the four investigated methods can predict the dynamics of complexity with reasonable accuracy.

ACKNOWLEDGEMENT

This work was supported by the National Science Foundation, award IIS-#1418511 and CCF-#1533768, Children’s Healthcare of Atlanta, Google Faculty Award and UCB.

References


Appendix: CCS diagnosis codes identified in feature selection.

Figure 6: CCS diagnosis codes selected by $L_1$ Logistic regression ($c = 0.02$) for prediction task Pos2(Pos/Neg)
Identifying High Health Care Utilizers Using Post-Regression Residual Analysis of Health Expenditures from a State Medicaid Program

Chengliang Yang¹, Chris Delcher², Ph.D., Elizabeth Shenkman², Ph.D., and Sanjay Ranka¹, Ph.D.
¹Dept. of Computer & Information Science & Engineering
²Dept. of Health Outcomes & Policy
University of Florida, Gainesville, FL 32611, USA
ximen14@ufl.edu, cdelcher@ufl.edu, eshenkman@ufl.edu, ranka@cise.ufl.edu

Abstract

We propose an approach to identify high health care utilizers using residuals from a regression-based health care utilization adjustment model to analyze the variations in health care expenditures. Using a large administrative claims dataset from a state public insurance program, we show that the residuals can identify a group of patients with high residuals whose demographics and categorization of comorbidities are similar to other patients but who have a significant amount of unexplained health care utilization. Additionally, these high utilizers persist from year to year. Correlation analysis with 3M™ Potentially Preventable Events (PPE) software shows that a portion of this utilization may be preventable. In addition, these residuals can be useful in predicting future PPEs and hence may be useful in identifying impactable high utilizers.

1 Introduction

The Agency for Healthcare Research and Quality (AHRQ) reports that in 2012, the top 10% of the health care-utilizing population accounted for 66% of overall health care expenditures in the United States. This highly disproportionate spending pattern frequently is interpreted as a sign of inefficient health care delivery. Furthermore, when such spending is driven by health care events believed to be avoidable, health care utilization is considered preventable and/or unnecessary. In this context, stakeholders have argued the need for higher efficiency health care for such patients, sometimes referred to as "high utilizers". For example, the deployment of managed care organizations (MCOs) and the capitation payments system in the United States public health programs provide incentives for health care providers to deliver services in a more cost-effective way.

The Centers for Medicare and Medicaid Services (CMS) recommends in their report that state Medicaid programs determine the extent to which expenditures driven by high utilizing populations represent impactable costs. Nationally, in 2010, potentially avoidable emergency department (ED) encounters accounted for $64.4 billion, 19.6% of ED episodes, and 2.4% of national health expenditures. As an example at the state-level, the Minnesota Department of Public Health found that nearly 1.3 million visits to hospitals in the state in 2012, representing nearly $2 billion in associated costs, could have potentially been prevented.

Most existing studies identify high utilizers based on the total number of visits or total expenditures or some combination thereof. While using such methods may be a relatively straightforward starting point for Medicaid programs with limited analytic resources, the approach can result in the inclusion of patient populations with health conditions unresponsive to efforts to reduce inappropriate spending. To illustrate the problem of relying on these criteria alone to identify impactable patient populations, consider the following examples. First, during flu season, elderly patients with little-to-no access to primary care may generate a large number of ED visits resulting in relatively inexpensive treatments. Second, patients with serious conditions, such as cancer or traumatic injuries, may require expensive medical treatments that are entirely appropriate and necessary.

An alternative cost-reduction strategy may involve more sensitive data-driven approaches that identify more impactable subpopulations of high utilizers even though they may not incur the highest absolute expenditures. We assume that this group of patients is "over-utilizing" the health care system because they spend more than their expected costs based on adjustments for health care-utilizing factors. Though clinical diagnostic classification tools such as the New York University ED profiling algorithm and 3M™ Potentially Preventable Events (PPE) are designed to improve
the ability to identify preventable or avoidable conditions, they have a number of limitations, including a limited population scope, lack of transparency due to commercial considerations, and others.

This paper proposes an approach to address such limitations. Briefly, using claims data typically available to state Medicaid programs, our data-driven approach calculates expenditure expectations for specific health care conditions and other measurable factors first; we then identify patients whose expenditures are higher (i.e. overutilization) than expected. The magnitude of overutilization is quantified by the degree to which cost residuals deviate from the model for each patient. Thus, those with higher residuals are hypothesized to represent potentially preventable high utilizers. Our models cannot account for all factors (e.g. genetic variations, socioeconomic status) expected to be predictive of health care utilization. However, we hypothesize that a non-trivial fraction can be potentially preventable.

Briefly, the key contributions of our paper are as follows:

- Using a large administrative claims dataset from a public insurance program with millions of patients and spanning for 4 years, we identify a best-fitting regression model to account for patients’ acute and chronic conditions loads and demographic characteristics.
- We identify populations with the highest deviations from expected after adjustment and examine their characteristics. Furthermore, we examine whether the model identifies patients that overutilize consistently through time. The results suggest that our approach identifies a significant proportion of health care costs that are consistent from year to year.
- We stratify the model by service settings (inpatient hospital, emergency department). In each setting, we compare our approach with an existing clinical tool, the 3MTM PPE.

The rest of this paper is organized as follows: In Section 2, we introduce the dataset and methods used for the study. Specifically, we find the best fitting model and define how residuals are used to identify high utilizers. Section 3 examines the demographics, clinical characteristics, utilization profiles and temporal consistency of the high-utilizing population. In Section 4, we compare the residuals approach with the 3MTM PPE tool. Section 5 concludes the paper.

2 Data and Methods

Study Population In this study, we use an administrative claims and encounter dataset from the Medicaid program of the State of Texas. As this is a repeated cross-sectional study, we set the inclusion criteria at a yearly basis. For each year from 2011 through 2014, we include adult (18-60 years old) Texas Medicaid beneficiaries, excluding pregnant women, with nonzero expenditures. The resulting size of the study population for 2011 to 2014 is 464572, 530242, 514601, and 535422 respectively. *

Data Preprocessing We normalize the health care expenditures to a per-member per-month dollar amount (total medical expenditure divided by number of months enrolled in Medicaid). This measure is commonly used for expenditure analyses in Medicaid programs12. To more meaningfully summarize the 21,374 unique diagnosis codes (International Classification of Diseases, Ninth Revision, Clinical Modification, ICD-9-CM) identified from the dataset, we group ICD-9-CM codes into 283 categories using AHRQs Clinical Classification Software13.

Model Regression-based adjustment models3 have been widely used in Medicare capitation payments as they systematically account for spending associated with specific health care conditions. Generally, we can write the model into the equation below:

\[ y = \beta x + \epsilon \] (1)

where \( y \), \( x \), and \( \beta \) represent expenditure, a vector of exogenous health care utilization factors and their linear coefficients respectively. If all the factors associated with health-care expenditures are exogenous, inclusion of these factors

*This study is approved by the IRB of the University of Florida.
would perfectly explain all the variations in health care expenditures except for a normally distributed error \( \epsilon \). If this is the case, the residuals, which are the observed error term \( \epsilon \) in (1), should follow the normal distribution. This is also an assumption of classical linear regression. If the residuals have a longer right tail than the normal distribution, we would expect that unmeasured factors are driving health care utilization. We set up the linear regression as specified in Table 1 to account for the patients’ health care conditions, demographics and health insurance plan differences. Note that we adjust for only the exogenous variables largely considered non-modifiable, such as health conditions, demographics, residence county, etc., but not for the endogenous variables like frequency of health care visits. We use ordinary least squares (OLS) to find the best fit for the health care utilization adjustment model (the model).

**Fitting the Model**  As described, the residuals analysis starts by setting up a model to account for exogenous factors. The models are highly significant for each year from 2011 to 2014 \((p < 0.001)\) for the \( F - test \). We also run multiple standard statistical diagnostics to check if the residuals meet the assumptions described previously.

We examine the distribution of our dependent variable and identify significant skewness. As a result, the residuals are not well aligned with the normal distribution as assumed (Figure 1a). Therefore, we log-transform the per-member per-month expenditure with base 10. Consequently, normality improves as shown in Figure 1b. Also, the R-squared of the model improves from 0.27 to 0.57 on average. Thus, all results discussed are based on the log-transformed model. This model shows a long tail on the right hand side, implying higher-than-expected value of residuals. We will later proceed to examine whether this represents potentially preventable utilization.

Although multi-collinearity is detected among the independent variables of the model, we retain all correlated variables in the model for the following reasons. First, our goal is not to interpret any individual independent variable or estimate its effects. Second, from the adjustment point of view, as long as the independent variables are not completely linearly correlated, we deem them acceptable for inclusion.

Heteroscedasticity also could affect the distribution of residuals. If strong heteroscedasticity exists among the patients, the high residuals could be those with error terms \( \epsilon \) of larger variances. We investigate the residuals versus predicted plot (Figure 2) and find that the residuals are distributed along the x-axis with similar variances. Although the Breuch Pagan test does suggest that heteroscedasticity exists \((p < 0.05)\), after we adopt the heteroscedasticity-consistent standard errors and re-run the model, it remains statistically significant. We conclude that heteroscedasticity is not a major issue for our dataset.

**Identifying the High Residuals Population**  We define an empirical threshold to discriminate the high residuals population from the rest of the patients. We select the residuals value as the cutoff threshold where the long tail begins to consistently deviate from the normal distribution. The process is described graphically on the Q-Q plot in Figure 3 based on our hypothesis that unexpected high utilizers are contained in the right long tail. In the following discussion, the term “high utilizer” always refers to the high residuals population.

Overall, we observe higher-than-expected total cost residuals on the right tail. We hypothesize that this may represent overutilization in the health care system and is worth investigating. We compare the high utilizer group to other patients using demographics, utilization patterns and overall chronic health condition burden using the Charlson Comorbidity

<table>
<thead>
<tr>
<th>Model</th>
<th>Classical linear regression</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dependent Variable</td>
<td>Per-member per-month expenditure</td>
</tr>
<tr>
<td>Independent Variables</td>
<td>Disease categories: ICD-9-CM codes grouped into Clinical Classification Software categories (CCS)(^{13})</td>
</tr>
<tr>
<td></td>
<td>Demographics: age, sex, race, and disabled status</td>
</tr>
<tr>
<td></td>
<td>Geographical information: county of residence</td>
</tr>
<tr>
<td></td>
<td>Health insurance programs and plans: fee-for-service, managed care organization plans</td>
</tr>
</tbody>
</table>

**Table 1: Model Specification**
Figure 1: Quantile-to-quantile plot (Q-Q plot) of standard normal distribution and 1,000 sample residuals obtained from the model for year 2014. In the left panel, the dependent variable, per-member per-month expenditures, is not transformed. Due to its strong skewness, the residuals significantly deviate from normal distribution. In the right panel, we log-transform the dependent variable with base 10 and re-run the model. The Q-Q plot better fits a normal distribution. The right and left ends are heavy-tailed, indicating over/under utilization.

Index. We examine temporal consistency using the following two-step process to test the Pearson product-moment correlation between residuals rank percentiles in different years from 2011 through 2014.

Step 1: Rank order all the patients in year 1 and year 2 based on residuals magnitude.

Step 2: Compute the Pearson product-moment correlation coefficient between the rank percentiles in the two years.

Stratified Model We stratify the adjustment model to specific service settings of inpatient hospital and emergency department (ED). For this purpose, in each setting, the study population is limited to those who had nonzero expenditures in the corresponding setting. In addition, we construct dependent variables based on the setting in which the expenditure occurred. Otherwise, our modeling approaches are the same as above. We compare our results with 3MTM PPE as a quasi-validation tool. To be specific, we examine if there are differences in PPE statistics between high residuals patients and all others. Also, we compute the Pearson correlation coefficient between PPE measures and residuals. Additionally, we test if the residuals are predictive for future PPEs.

3 Characterizing the High Utilizers

In this section, we take a closer look at the high residuals population. We first descriptively show their characteristics and then study their temporal consistency.

Demographics, Health Conditions, and Utilization For 2014, we summarize the demographics, health conditions, and health care utilization of high utilizers and other patients in Table 2. These results do not vary significantly by year. In terms of demographics, there are more male and Hispanic patients in the high-utilizer group. We use an integrated score, Charlson Comorbidity Index, to represent patients’ overall disease burden. The two populations vary little on this measure, which means the categorization of comorbidities is similar. Thus, we argue that there are no significant differences in demographics and categorization of comorbidities. This is not surprising because we expect the model to account for all these factors.

However, the levels of health care utilization in the two groups are completely different. In general, the high utilizer group has annual expenditures about four times higher than other patients, and they consistently have more expen-
ditures in each service category. The only exception is that high utilizers visit the ED less frequently than the other patients do. But because ED visits comprise a small fraction of high utilizers’ overall costs, ED expenses have less impact on the overall cost. In the next section, our stratified models examine costs incurred in the ED more specifically.

The findings in this section support our hypothesis that a group of patients whose demographics and comorbidities looks similar to other patients has a significantly higher amount of unexplained health care utilization.

Temporal Consistency of Residuals  
Next, we examine the extent to which this excessive utilization persists through time. Generally, if high unexplained variance occurs at random from year to year, then preventing such health care events would be extremely challenging. Otherwise, non-random correlation from year to year may suggest the persistence of health care-utilization factors that might be discoverable. To examine, we compute the Pearson product-moment correlation coefficient between the rank percentiles of residuals for any two subsequent years from 2011 to 2014, as shown in Table 3. Results show significant correlation between residuals from year to year. As expected, the correlation is strongest to the immediate subsequent year and decreases with time. The strong, significant correlation is visually well-represented as a dense diagonal on the scatter plot of rank percentiles of residuals of 2013 and 2014 in Figure 4. We observe an even stronger consistency for the high utilizers because the upper right area is denser. To be specific, the high utilizers of 2011, 2012 and 2013 on average rank 72%, 78% and 79% respectively in residuals percentile for the next year. In comparison, other patients has an average rank of 50%.

This correlation structure indicates that expenditures do not vary randomly from year to year and implies that unobserved factors drive high utilization. In conclusion, the subpopulation corresponding to the high residuals shows an excessive amount of utilization that is temporally consistent and driven by unknown factors.

4 Stratified Models by Service Settings

In the previous section, we have shown that the residuals indicate an unexplained high amount of health care utilization. Next, we examine whether this utilization is associated with preventable costs. The 3M™ PPE software identifies potentially preventable health care events in the inpatient hospital and ED settings. The software is proprietary but its validity has been verified and widely accepted. In this section, we examine the correlation between the 3M™ PPE approach and our own. 3M™ Potentially Preventable Readmissions (PPR) Grouping Software identifies clinically

Figure 2: Predicted versus residuals plot for 2,000 observation random sample for 2014. No obvious heteroscedasticity is observed from the plot.

Figure 3: Decide threshold of high utilizer from Q-Q plot: The red dashed line indicates where the long right tail of residuals consistently deviates from the normal distribution quantiles. The analysis identifies 10,121 high utilizers for 2014.
### Characteristics

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>High Utilizers (High Residuals)</th>
<th>Other Patients</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Number of Patients</strong></td>
<td>10,121</td>
<td>525,301</td>
</tr>
<tr>
<td><strong>Demographics</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean age (years)</td>
<td>33.84</td>
<td>35.19</td>
</tr>
<tr>
<td>Sex</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female, %</td>
<td>56.70</td>
<td>68.19</td>
</tr>
<tr>
<td>Race</td>
<td></td>
<td></td>
</tr>
<tr>
<td>White, non-Hispanic, %</td>
<td>21.32</td>
<td>26.40</td>
</tr>
<tr>
<td>Black, non-Hispanic, %</td>
<td>16.58</td>
<td>21.79</td>
</tr>
<tr>
<td>Hispanic, %</td>
<td>44.79</td>
<td>37.35</td>
</tr>
<tr>
<td>American Indian or Alaskan, %</td>
<td>0.21</td>
<td>0.19</td>
</tr>
<tr>
<td>Asian, Pacific Islander, %</td>
<td>1.90</td>
<td>1.59</td>
</tr>
<tr>
<td>Unknown/Other, %</td>
<td>15.20</td>
<td>12.68</td>
</tr>
<tr>
<td>Disabled, %</td>
<td>50.12</td>
<td>45.54</td>
</tr>
<tr>
<td><strong>Health Conditions</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Charlson Comorbidity Index(^{14})</td>
<td>1.68</td>
<td>1.61</td>
</tr>
<tr>
<td><strong>Utilization</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average total medical expenditures, $</td>
<td>25,544.18</td>
<td>6,882.04</td>
</tr>
<tr>
<td>Average professional expenditures †, $</td>
<td>4,979.69</td>
<td>2,271.29</td>
</tr>
<tr>
<td>Average institutional expenditures ‡, $</td>
<td>10,265.39</td>
<td>2,543.53</td>
</tr>
<tr>
<td>Average pharmacy expenditures, $</td>
<td>10,225.37</td>
<td>2,054.03</td>
</tr>
<tr>
<td>Average number of emergency department visits</td>
<td>0.35</td>
<td>1.01</td>
</tr>
<tr>
<td>Average number of inpatient hospital visits</td>
<td>0.49</td>
<td>0.21</td>
</tr>
</tbody>
</table>

† Professional expenditures represent paid claims generated for work performed by physicians, suppliers, and other non-institutional providers for all medical services.

‡ Institutional expenditures represent paid claims generated for work performed by hospitals, skilled nursing facilities, and other institutions for all medical services.

### Table 2: Comparison of demographics, health conditions and health care utilization of high utilizers with other patients, 2014

related and potentially preventable inpatient hospital readmissions. 3M\(^{TM}\) Potentially Preventable Emergency Visits (PPV) Grouping Software identifies ED visits that relates to ambulatory-sensitive conditions and may result from lack of access to primary care. Thus, we can identify potentially preventable utilization in two important service settings, inpatient hospital and ED. We modify the dependent variable to reflect expenditures incurred only in those settings. By comparing the PPE statistics and resulting residuals from the stratified model, we examine our hypothesis that our approach can effectively identify preventable health-care utilization.

### Residuals and Potentially Preventable Readmissions (PPR)

Table 4 presents the PPR statistics for the high utilizer group versus other patients. The high utilizer group has a significantly (MannWhitney U test, \(p < 0.05\)) higher amount of PPR events and PPR expenditures than all others. The difference is consistent from 2011 to 2014. This suggests that the high residuals are associated with a substantial amount of potentially preventable hospitalizations. The Pearson correlation coefficients between the residuals and PPR expenditures tell the same story. PPR expenditures are log-transformed to scale with the residuals. We compute the correlation coefficient with the residuals obtained from the stratified model using inpatient hospital expenditures as the dependent variable, as shown in Table 5. The correlation is always significant through the years, suggesting a strong association between residuals and PPRs.
<table>
<thead>
<tr>
<th>Year/Correlation</th>
<th>2012</th>
<th>2013</th>
<th>2014</th>
</tr>
</thead>
<tbody>
<tr>
<td>2011</td>
<td>0.396</td>
<td>0.326</td>
<td>0.287</td>
</tr>
<tr>
<td>2012</td>
<td>-</td>
<td>0.457</td>
<td>0.380</td>
</tr>
<tr>
<td>2013</td>
<td>-</td>
<td></td>
<td>0.472</td>
</tr>
</tbody>
</table>

Table 3: The Pearson correlation coefficient between the rank percentiles of residuals for any two subsequent years from 2011 to 2014.

Figure 4: Scatter plot of rank percentiles of residuals: 2013 and 2014.

(a) Scatter plot for overall population

(b) Scatter plot for high utilizers

Residuals and Potentially Preventable Emergency Department Visits (PPV) We replicate the analysis for potentially preventable ED visits. The population is limited to patients with nonzero ED expenditures. The dependent variable is changed to per-member per-month ED expenditures and log-transformed. The statistics of PPVs of high utilizers versus other patients are presented in Table 6. The correlation test results are shown in Table 5. All results imply that the residuals also identify a significant amount of preventable ED utilization. However, comparing the level of differences in mean statistics in Table 4 and 6, the relationship for PPVs and residuals is weaker compared to PPRs and residuals. This generally makes sense because ED visits are more incidental than inpatient hospital visits.

Residuals and Future Potentially Preventable Events In Table 4 and 6, we also show the PPR and PPV statistics for the year following the year from which we compute the residuals (index year). Although the differences in PPR and PPV utilization amounts between high utilizers and other patients are narrowed from the index year, the high-utilizers group still has significantly more PPRs and PPVs, as well as associated costs. The correlation coefficients of the next year in Table 5 also indicate a decreased but still sizable temporal correlation. Thus, the residuals are also good predictors for future PPRs and PPVs. This finding is consistent with the strong temporal correlation of residuals we present in the previous section.

To summarize the section, using the 3M™ PPE software, we have demonstrated that the residuals from the models, when stratified to inpatient hospital and ED settings, are strongly associated with potentially preventable utilization of
Table 4: Potentially Preventable Readmissions (PPR) statistics of index year and next year for high utilizers versus other patients. Residuals and high utilizers are identified from index year. Per-member per-month inpatient hospital expenditures is the dependent variable of the model.

<table>
<thead>
<tr>
<th></th>
<th>Index year</th>
<th>High Utilizers (High Residuals)</th>
<th>Other Patients</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Average number of PPRs in index year</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2011</td>
<td>1.24</td>
<td></td>
<td>0.15</td>
</tr>
<tr>
<td>2012</td>
<td>1.12</td>
<td></td>
<td>0.20</td>
</tr>
<tr>
<td>2013</td>
<td>1.10</td>
<td></td>
<td>0.21</td>
</tr>
<tr>
<td>2014</td>
<td>0.86</td>
<td></td>
<td>0.19</td>
</tr>
<tr>
<td><strong>Average PPR expenditures in index year, $</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2011</td>
<td>7,793.23</td>
<td>1,200.63</td>
<td></td>
</tr>
<tr>
<td>2012</td>
<td>6,807.52</td>
<td>1,259.59</td>
<td></td>
</tr>
<tr>
<td>2013</td>
<td>6,229.76</td>
<td>1,237.03</td>
<td></td>
</tr>
<tr>
<td>2014</td>
<td>5,241.54</td>
<td>1,112.03</td>
<td></td>
</tr>
<tr>
<td><strong>Average number of PPRs in next year</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2011</td>
<td>0.70</td>
<td>0.17</td>
<td></td>
</tr>
<tr>
<td>2012</td>
<td>0.50</td>
<td>0.18</td>
<td></td>
</tr>
<tr>
<td>2013</td>
<td>0.37</td>
<td>0.14</td>
<td></td>
</tr>
<tr>
<td><strong>Average PPR expenditures in next year, $</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2011</td>
<td>3,698.67</td>
<td>1,113.45</td>
<td></td>
</tr>
<tr>
<td>2012</td>
<td>2,549.41</td>
<td>1,001.92</td>
<td></td>
</tr>
<tr>
<td>2013</td>
<td>2,072.29</td>
<td>758.82</td>
<td></td>
</tr>
</tbody>
</table>

Table 5: The Pearson correlation coefficient between the residuals from the stratified model (inpatient hospital and emergency department) with Potentially Preventable Readmissions (PPR) expenditures and Potentially Preventable Emergency Department Visits (PPV) expenditures respectively from 2011 to 2014. Residuals and high utilizers are identified from index year.

<table>
<thead>
<tr>
<th></th>
<th>Index Year</th>
<th>Correlation Coefficient with Residuals</th>
</tr>
</thead>
<tbody>
<tr>
<td>log_{10}(Per-member per-month PPR expenditures of index year)</td>
<td>2011</td>
<td>0.2974</td>
</tr>
<tr>
<td></td>
<td>2012</td>
<td>0.2638</td>
</tr>
<tr>
<td></td>
<td>2013</td>
<td>0.2755</td>
</tr>
<tr>
<td></td>
<td>2014</td>
<td>0.2664</td>
</tr>
<tr>
<td>log_{10}(Per-member per-month PPV expenditures of index year)</td>
<td>2011</td>
<td>0.2452</td>
</tr>
<tr>
<td></td>
<td>2012</td>
<td>0.3214</td>
</tr>
<tr>
<td></td>
<td>2013</td>
<td>0.3423</td>
</tr>
<tr>
<td></td>
<td>2014</td>
<td>0.3165</td>
</tr>
<tr>
<td>log_{10}(Per-member per-month PPR expenditures of next year)</td>
<td>2011</td>
<td>0.0990</td>
</tr>
<tr>
<td></td>
<td>2012</td>
<td>0.0972</td>
</tr>
<tr>
<td></td>
<td>2013</td>
<td>0.0948</td>
</tr>
<tr>
<td>log_{10}(Per-member per-month PPV expenditures of next year)</td>
<td>2011</td>
<td>0.0908</td>
</tr>
<tr>
<td></td>
<td>2012</td>
<td>0.1051</td>
</tr>
<tr>
<td></td>
<td>2013</td>
<td>0.1121</td>
</tr>
</tbody>
</table>

the same year as well as the next year, implying promising potential to identify impactable high utilizers.

5 Conclusion and Discussion

In this paper, we propose a novel approach to analyze variations in Medicaid health care expenditures based on using higher-than-expected values of the residuals from health care utilization adjustment models. We conducted our analyses on a large administrative claims dataset from a state public insurance program. Our approach identifies a
<table>
<thead>
<tr>
<th></th>
<th>Index year</th>
<th>High Utilizers (High Residuals)</th>
<th>Other Patients</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Average number of PPVs in index year</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2011</td>
<td>1.51</td>
<td>1.37</td>
<td></td>
</tr>
<tr>
<td>2012</td>
<td>3.10</td>
<td>1.89</td>
<td></td>
</tr>
<tr>
<td>2013</td>
<td>3.23</td>
<td>1.95</td>
<td></td>
</tr>
<tr>
<td>2014</td>
<td>2.99</td>
<td>1.88</td>
<td></td>
</tr>
<tr>
<td><strong>Average PPV expenditures in index year, $</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2011</td>
<td>1,815.21</td>
<td>747.62</td>
<td></td>
</tr>
<tr>
<td>2012</td>
<td>3,444.05</td>
<td>698.43</td>
<td></td>
</tr>
<tr>
<td>2013</td>
<td>3,653.24</td>
<td>703.12</td>
<td></td>
</tr>
<tr>
<td>2014</td>
<td>3,986.26</td>
<td>664.19</td>
<td></td>
</tr>
<tr>
<td><strong>Average number of PPVs in next year</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2011</td>
<td>1.87</td>
<td>1.75</td>
<td></td>
</tr>
<tr>
<td>2012</td>
<td>2.46</td>
<td>1.71</td>
<td></td>
</tr>
<tr>
<td>2013</td>
<td>2.48</td>
<td>1.69</td>
<td></td>
</tr>
<tr>
<td><strong>Average PPV expenditures in next year, $</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2011</td>
<td>1,083.07</td>
<td>750.32</td>
<td></td>
</tr>
<tr>
<td>2012</td>
<td>1,453.91</td>
<td>653.90</td>
<td></td>
</tr>
<tr>
<td>2013</td>
<td>1,495.37</td>
<td>647.75</td>
<td></td>
</tr>
</tbody>
</table>

Table 6: Potentially Preventable Emergency Department Visits (PPV) statistics of index year and next year for high utilizers versus other patients. Residuals and high utilizers are identified from index year. Per-member per-month emergency department expenditures is the dependent variable of the model.

significant amount of unexplained health care utilization. This utilization fraction is shown to be largely preventable as it produces results similar to the 3MTM PPE software. Thus, other state Medicaid programs could potentially use this approach to identify impactful high utilizers and possibly reduce inappropriate health care spending.

This study is limited in several dimensions. First, though our model tries to include a large number of exogenous variables in the claims dataset, it is not comprehensive. For example, important variables that contribute to health care expenditures, such as health condition severity measures, are missing. The residuals we obtain from our current model are likely to be affected by these unadjusted exogenous variables. Second, we have not conduct case studies to follow up on specific clinical details about the high utilizers. Such an investigation could reveal more information about clinical factors contributing to the variation. Third, though the residuals prove to be highly associated with potentially preventable utilization, they do not point to specific pathways to inform policy. Thus, to impact high utilizers, we need to decompose the residuals and identify the source of variation.

Our future work will try to address these limitations. We will gather more data, such as social determinants of health, to adjust as many exogenous factors as possible. We also plan to conduct medical record reviews with clinical and health care policy experts to identify components of care that could be addressed to reduce preventable utilization. More importantly, we will extend this exclusively data-driven approach into an iterative process16 between health care practitioners and informatics researchers to better understand impactable health care conditions and progress toward interventions to reduce inappropriate health care utilization.

ACKNOWLEDGMENTS

This work was supported in part by Texas Health and Human Services (HHS) and in part through the Patient-Centered Outcomes Research Institute (PCORI) (PCO-COORDCTR2013) for development of the National Patient-Centered Clinical Research Network, known as PCORnet. The 3MTM company provided software support for the PPE tools. The views, statements and opinions presented in this work are solely the responsibility of the author(s) and do not necessarily represent the views of Texas HHS and the Patient-Centered Outcomes Research Institute (PCORI), its board of governors or methodology committee or other participants in PCORnet.
References


Classification of hepatocellular carcinoma stages from free-text clinical and radiology reports

Wen-wai Yim, PhD1,2, Sharon W Kwan, MD3, Guy Johnson, MD, PharmD3, Meliha Yetisgen, PhD4,5
1Veterans Affairs Palo Alto Health Care System, 2Biomedical Informatics Research, Stanford University, Palo Alto, CA; 3Radiology, 4Biomedical and Health Informatics, 5Linguistics, University of Washington, Seattle, Washington

Abstract
Cancer stage information is important for clinical research. However, they are not always explicitly noted in electronic medical records. In this paper, we present our work on automatic classification of hepatocellular carcinoma (HCC) stages from free-text clinical and radiology notes. To accomplish this, we defined 11 stage parameters used in the three HCC staging systems, American Joint Committee on Cancer (AJCC), Barcelona Clinic Liver Cancer (BCLC), and Cancer of the Liver Italian Program (CLIP). After aggregating stage parameters to the patient-level, the final stage classifications were achieved using an expert-created decision logic. Each stage parameter relevant for staging was extracted using several classification methods, e.g. sentence classification and automatic information structuring, to identify and normalize text as cancer stage parameter values. Stage parameter extraction for the test set performed at 0.81 F1. Cancer stage prediction for AJCC, BCLC, and CLIP stage classifications were 0.55, 0.50, and 0.43 F1.

Introduction
Cancer staging is the categorization of the severity of cancer progression based on patient and disease characteristics. Clinically, cancer stage can be used to determine the most appropriate treatment option or to predict prognosis. In research, cancer stage facilitates outcomes analysis by allowing comparisons between patients with similar severity of disease. Numerous stage classification schemes may exist for each cancer type and are sometimes quite complex. Further, cancer stages are not reliably recorded in the electronic medical record (EMR) in either structured or unstructured form [1]. Even when collected, the accuracy of recorded stages may be incomplete or inaccurate [2–6]. Automatic classification of staging has many benefits in cancer research. A successful system would circumvent the need for separate data abstractors to perform work that is resource-intensive, expensive, and error-prone. An automatic system would be scale-able to large volumes of data. Free text narrative is the preferred modality of choice in the clinical setting and an automated system would give the benefits of classifying patients into cancer stages without altering clinical workflow. Additionally, an automatic classification system could be deployed to assign stage using multiple staging systems or could be adapted for changes in staging systems.

This study focuses on automatic cancer stage classification for hepatocellular carcinoma (HCC), also known as primary liver cancer, using free-text clinical and radiology notes and four laboratory values. The three HCC staging systems we used, American Joint Committee on Cancer (AJCC), Barcelona Clinic Liver Cancer (BCLC), and Cancer of the Liver Italian Program (CLIP), depend on a mixture of tumor characteristics, e.g. tumor size, assessment of liver dysfunction, e.g. ascites, and patient characteristics, e.g. performance status, as factors in its staging algorithms [7]. The purpose of this study was to extract each of those factors in the staging algorithms and use that information to classify cancer stages.

Background
There are more than six different liver cancer staging systems world-wide, reflecting various ways to assess liver function as well as cancer spread [7]. We pooled the common factors used for the staging of the three commonly used staging systems (AJCC, BCLC, and CLIP) and provide common categories for each parameter that can be used to calculate the final stage classifications. A brief description of each stage classification system and the final stage parameter follows. Figure 1 shows an example of how stage parameter values translate to a stage classification for BCLC.

Figure 1. Example of stage parameters used for BCLC staging

1841
Stages

AJCC has classifications (I, II, IIIA, IIIB, IIIC, IVA, IVB) and is based on the TNM stage framework that primarily addresses tumor characteristics and spread (tumor size, number, metastasis) but not liver functioning statuses.

BCLC has classifications (A1, A2, A3, A4, B, C, D) and is the only staging scheme that takes into account overall patient performance status, using Eastern Cooperative Oncology Group’s (ECOG) classification system.

CLIP has classifications (0, 1, 2, 3, 4, 5, 6) and is the only staging scheme that takes into account the relative size of the tumor to the liver. This stage classification takes into account alpha fetoprotein level, a lab value.

Stage parameters

Ascites: accumulation of fluid in the peritoneal cavity (e.g., “no significant ascites”, “does not endorse abdominal swelling”) with values (None, Mild, Moderate-Severe)

Child-Pugh: a measurement of liver cirrhosis severity (e.g., “Child’s B”, “his CTP score would be 5”) with values (A, B, C)

ECOG (Eastern Cooperative Oncology Group) Performance Status: a scaled measure of general well-being where 0 is fully active and 5 is dead (e.g., “ECOG 0”, “She notes good energy”) with values (0, 1, 2, >=3)

Extrahepatic invasion: direct spread of cancer outside of the liver (e.g., “No evidence of extrahepatic extension”, “the tumor may […] extend from the liver to the right ribs or muscular wall”) with values (No, Yes)

Hepatic encephalopathy: confusion or altered consciousness attributed to liver cirrhosis (e.g., “patient denies confusion, forgetfulness, or other symptoms of hepatic encephalopathy”, “lactulose” in the medication list) with values (None, Mild, Severe)

Macrovascular invasion: spread of cancer to nearby blood vessels (e.g. “vascular invasion: possible involvement of middle hepatic vein branches”, “no evidence of portal vein thrombosis”) with values (No, Yes-minor branch, Yes-major branch)

Metastasis: spread of cancer to outside the liver, such as to lymph nodes (e.g., “lymph nodes suspicious for metastatic involvement: none”, “no lymphadenopathy”) with values (No, Yes-regional, Yes-distal)

Portal hypertension: elevation of hepatic venous pressure gradient to greater than 5 mm Hg (e.g., “no evidence of cirrhosis or portal hypertension”, “patient had an endoscopy which showed small varices”) with values (No, Yes)

Tumor morphology: number and size of tumor relative to the liver (e.g., “small segment 7 hepatic mass”, “two new liver lesions noted […] suggesting hepatomas”) with values (Massive >= 50% of liver, Multinodular < 50% of liver, Uninodular < 50% of liver)

Tumor number: number of liver tumors (e.g., “1 lesion measuring 2.1 x 1.7 cm […] HCC”) with values (<3 cm, 3-5 cm, >5 cm).

Though many stage parameters factor into the stage classifications of at least two staging systems, some parameters are specific to a single staging system. For example, BCLC is the only system that incorporates ECOG, meanwhile only CLIP uses tumor morphology. It is also important to note that classification decisions for a stage category may depend on only a single stage parameter, or at other times many stage parameters. For example, in BCLC staging, an ECOG score of 3 automatically places the patient in Stage D regardless of other variables.

Related work

There are few published work in automatic cancer stage classification from free text reports where stages are not explicitly written. Several systems that do exist include the lung cancer staging work by Nguyen et al. [8,9] and McCowan et al. [1] and the prostate cancer staging from pathology notes by Martinez et al. [10,11]. The most evolved systems required division of classifications to handle several stage parameters, mixed method modalities (e.g. sentence versus document classification), and some rules. For example, McCowan et al. [1] classified sub-stages for Tumor-Node-Metastasis (TNM) staging from least to most severe, e.g. classifying for T0 before T1 and so on. Most of these classifications were sentence classifications with two steps, the first for relevance to the sub-stage (yes or no for T0) the second to identify positives. In Nguyen et al. [9], concepts were identified and normalized through the Systematized Nomenclature of Medicine - Clinical Terms vocabulary with stage classified through symbolic logic rules. Martinez et al. [10] used a sentence classification for numeric stage parameters and a document classification for nominal stage parameters for TNM staging and Australian clinico-pathological staging (ACPS) staging. Other works focus on identifying and structuring specific aspects of cancer information, which could be used for cancer staging after processing [11–14]. Cancer stage classification, in the cervical cancer domain, from structured data using neural network methods have also been explored [15,16].
To our knowledge, this is the first reported system developed for HCC staging. HCC staging presents unique challenges in that staging parameters require incorporation of free-text derived from multiple data sources, including radiology reports, clinic visit notes, and laboratory results. Further, some staging systems such as BCLC require derivation of parameter values for complex concepts such as severity of hepatic encephalopathy. This work is unique in its integration of multiple text extraction systems with component- and application-level evaluations.

**Methods**

**Creation of a gold standard**

Our proposed methods sought to use several information extraction methods to identify and normalize free text to each stage parameter and its severity, e.g. *ascites - mild*. Afterwards, the patient value for each stage parameter would be determined and used to classify final patient-level stages for each staging system. In order to train and evaluate these extraction techniques, we needed to create a gold standard incorporating text level and patient level annotations.

**Dataset**

Following Institutional Review Board approval, a retrospective cohort was drawn from new patients visiting the University of Washington Medical Center liver cancer clinic from 1/2011-12/2013. Data for each patient comprised of: (1) all clinical notes from the day of clinic visit (2) all laboratory results within 30 days prior to and following the visit day, and (3) radiology reports within 3 months prior to and 1 month following the visit day. This study focused on the subset of patients that had at least one clinical report, at least one radiology report, and the full set of laboratory results needed for staging (Albumin, Alpha Fetoprotein, Bilirubin, Prothrombin INR). After manually excluding patients with irrelevant diagnosis, the final dataset included 200 patients with 545 documents.

**Annotation**

During annotation, text evidence was identified and normalized to stage parameters and patient-level stage parameter and stage values were identified and labeled. This corresponded to roughly two annotation phases. Phase I consisted of text annotation, similar to those shown in Figure 1 in which parts of text were highlighted and assigned to a stage parameter label and value.

![Figure 2. Text annotation](image)

![Figure 3. Patient-level annotation. Highlighted text from Phase I annotations appears in the left panel, where as laboratory values appear on the right panel. Annotators choose the patient-level 11 stage parameters and 3 stage classifications with the drop-down menu buttons on the bottom panel.](image)
We used a sparse annotation approach in which annotation of the repeated information within a document were omitted. The exact rules for which text to mark were specified in annotations guidelines (phekb.org/phenotype/568). Phase II consisted of patient-level annotation, shown in Figure 2, in which patient-level stage parameter and stages were marked. Patient-level annotations were performed by the consensus of two HCC domain experts. The relevant stage parameter and stage description and values were described previously. Figure 1 illustrates how stage parameters can be used to calculate BCLC stage. Phase I annotation agreement was at 0.73 F1 for partial text overlap and 0.91 F1 at the patient level. Phase II annotations were done in consensus therefore annotator agreement is not applicable. The details of this annotation may be read in a previous paper [17], however we will emphasize several subtle points of interest here.

Firstly, Child-Pugh and ECOG are themselves stages. Child-Pugh is a classification for liver cirrhosis with values that range from A-C; ECOG is a classification for general well-being with values that range from 0-4. Both require multiple sets of criteria in their staging algorithms. We define explicit text evidence as those text clues that do not require using the Child-Pugh or ECOG staging calculations to arrive at the proper normalization, e.g. “Patient is Child-Pugh A”. Non-explicit Child-Pugh stages were not annotated, but they were calculated from ascites and hepatic encephalopathy (both text stage parameters) as well as 3 laboratory values, according to the logic of Table 1. Meanwhile, non-explicit ECOG stage evidence, e.g. “patient is confined to a wheelchair”, were annotated with explicit ECOG. Secondly, the tumor-related stage parameters may sometimes be clearly written as text, e.g. “1 lesion suspicious of HCC, 6.0 cm”, however, oftentimes this is not the case. In fact, taking the radiology report extract from Figure 4 as an example, only after reading the full report can we ascertain that tumor morphology Uninodular and < 50% of liver of the liver, there is a single malignant tumor, and the tumor size is <3 cm.

Finally, while some of the text evidence of the stage parameters may be more explicit, e.g. “no ascites”, this is not always the case. Often medical knowledge or inference is required. For example, some “free fluid” in the abdomen, by definition, is considered ascites-mild. The medication “lactulose” implies hepatic encephalopathy – mild, as it is used to treat this condition. Macrovascular invasion requires reasoning not only about whether a tumor invaded blood vessels but also whether or not said blood vessels are considered major or otherwise. Signs and symptoms can also give clues regarding certain conditions, for example “no lymphadenopathy” (enlarged lymph nodes) implies there are no signs of local metastasis – though the converse is not true; enlarged lymph nodes does not imply that there is metastasis, regional or distal.

Training and test sets were separated according to a 160/40 patient split. A subset of 101 radiology reports, randomly sampled from the training set, was further annotated for tumor-related entity and relations, reference resolution, and tumor characteristics to train a tumor characteristics extractor, which is described previously [18].

### System design

Stage parameter text evidence passages were identified from documents using one of three ways: (1) a two-step regular expression extraction module, (2) a sentence classification module, and (3) a tumor characteristics extraction module. Then (3) the patient-level stage parameters were determined using a rule-based heuristic. Finally, (4) the stage classifications were labeled using a rule-based expert logic. Figure 5 shows a diagram of the system and each subsystem are described in the following sections.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Points</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variable</td>
<td>1</td>
</tr>
<tr>
<td>Albumin (g/dL)</td>
<td>&gt;3.5</td>
</tr>
<tr>
<td>Ascites</td>
<td>None</td>
</tr>
<tr>
<td>Bilirubin (Total) (mg/dL)</td>
<td>&lt;2</td>
</tr>
<tr>
<td>Hepatic Encephalopathy</td>
<td>None</td>
</tr>
<tr>
<td>Prothrombin INR</td>
<td>&lt;1.7</td>
</tr>
</tbody>
</table>

Within hepatic segment II/III there is a 14 x 9 mm hypervascular lesion is isointense to liver parenchyma on portal venous phase ...

This lesion is suspicious for hepatocellular carcinoma

**Impression:**
1. Hypervascular lesion in hepatic segment II/III with imaging features suspicious for hepatocellular carcinoma.

**Figure 4.** Radiology report excerpt

---

**Table 1.** Child-Pugh score. Each variable is given points depending on the patient condition. The sum from each variable is used to calculate Child-Pugh according to the following mapping. 5-6 points: Child-Pugh A, 7-9: Child-Pugh B, and 10-15 points: Child-Pugh C.
(1) Two-step regular expression extraction module:
This module was used to find text evidence for Child-Pugh class and ECOG parameters from clinical notes. The two-step regular expression extraction module was used to find explicit Child-Pugh and ECOG stage parameter text evidence only. The first step regular expression identifies a trigger term, e.g. “Child-Pugh” or “ECOG”. Once a trigger term had been identified, a second regular expression looks at the remaining part of the sentence for a stage cue, e.g. the “6” in “Child-Pugh score 6”, and maps it to the appropriate stage value, e.g. “A”.

(2) Sentence classification module:
Six stage parameters label and values (ascites, extrahepatic invasion, hepatic encephalopathy, macrovascular invasion, metastasis, and portal hypertension) were identified and normalized to the described predetermined categories using a sentence classification method. The sentence classification applied a multi-label classification approach. Features included 1-, 2-, 3-grams, their asserted versions (whether or not the term is negated, probable, etc), and asserted UMLS concepts, identified using MetaMap [19,20]. For example, in “Ascites: None”, a 1-gram “ascites” feature would be accompanied by “[NEGATED]-ascites” and “[NEGATED]-[UMLS:C0003962-ascites]”. We employed statistical feature selection using several measures: t-test, chi-squared, and point-wise mutual information for the n-grams and asserted UMLS concepts. A configurable number, N, was used to select the top scoring features. Only those features associated with the top N significant values for 1-, 2-, and 3-grams (with its asserted values) and asserted UMLS concepts were used as final features in a maximum entropy classifier. This module was applied to clinical notes for hepatic encephalopathy, to all note types for ascites, and only to radiology notes for remaining parameters.

(3) Tumor characteristics extraction module:
Because the stage parameters tumor morphology, tumor size, and tumor number required reasoning over the entire report, we needed to build a separate module that processed radiology reports. The overall sub module is represented in Figure 6. First template extraction, which identifies tumor-related entities, attributes, and relations, was

---

**Figure 5.** System design

**Figure 6.** The tumor characteristics extraction module does several processing steps, before distilling information to the tumor morphology, size, and number stage parameters.
performed on the Findings or Impression section of each radiology report. Entities were identified through a conditional random field (CRF) entity extractor. Relations and attributes were identified using a maximum entropy relation classifier [18,21]. Sections were identified using our statistical in house section identifier [22]; we used CRFSUITE [23] for CRF classification and MALLET [24] for maximum entropy classification.

Tumor reference resolution was performed to identify coreference cases and parent-child cases of tumor mentions. Using the above elements and reference resolution information, a rule-based characteristic annotator classified documents for whether 50% of the liver was invaded, the largest tumor size, and the number of tumors represented in a report. A rule-based transformation outputted the final values for tumor morphology, size, and number. This is described in more detail in another paper [25].

(4) Patient-level stage parameter classifier:

Patient-level stage parameter classifications were determined using a simple heuristic: given text evidence for stage parameter labels and values, choose the most severe case. For example, if there were mentions of both Child-Pugh A and Child-Pugh B, the patient value was assigned to Child-Pugh B. If no values were mentioned, the patient’s value defaulted to the least severe value. Child-Pugh values were taken from explicit regular expression extraction, as well as a rule-based calculation of Child-Pugh using ascites and hepatic encephalopathy text evidence.

(5) Patient-level stage classifier:

Lookup tables were created that accounted for all stage parameter combinations. Ambiguous guideline cases were adjudicated by an interventional radiology expert. Our rule-based patient-level stage classifier used the logic stored in our lookup tables for stage classification.

Training for (1) and (2), (all stage parameters except tumor morphology, tumor number, and tumor size) were performed with 5-fold cross validation of the 160 training patients. Training for tumor-related stage parameters were performed on the 101 radiology reports randomly sampled from the training set, then applied to the entire corpus.

Performance evaluation

We implemented several baselines for text extraction (1)-(3), the patient level stage parameter classification (4), and for patient level stage classification (5) evaluation. Since these baselines were used as intermediate evaluations for building the final system classification algorithm, they were only applied on the training set in cross-validation.

Text extraction baseline: We used a previously published document-level classification baseline, which used the same five-fold validation splits here, to compare text extraction results [17]. (Because of our sparse annotation, it was not appropriate to use a sub-document evaluation as baseline.) The baseline consisted of the best out of five machine learning classifiers (c4.5 decision tree, binary decision tree, maximum entropy, naïve bayes, and support vector machine) using lower-cased unigram, bigram, and trigram counts after tokenization and punctuations removed. The classification was a binary value {positive, negative} value for each label and value, e.g. ascites-none.

Patient level stage parameter and stage classification baselines: To evaluate against the simple patient-level stage parameter text evidence to patient-level classification (4), the baseline labels every instance with the most prevalent class (if ascites-none appears the most frequently in training, label all patients with this class). Patient level stage classifications (5) were measured against two baselines: a straight classifier the uses the most prevalent class (B1) and a maximum entropy 1-, 2-, 3- gram, and UMLS concepts document classifier with frequency features (B2).

Results

Table 2 shows the results for the regular expression, sentence classifier, and tumor characteristics extraction modules resolved for the document-level for the training and testing sets, where TP=true positive, P=positive, R=recall, F1= 2*P*R/(P+R). Reflecting the complexity of stage parameters, expectedly, the regular expression variables (Child-Pugh and ECOG) were the best-performing, whereas tumor-related parameters were the least. Table 3 shows the performance of the rule-based stage parameter aggregation heuristic against the majority class classifier baseline.

Patient-level stage classifications are shown in Table 4 for the 3 staging schemes, with the majority class classifier baseline (B1) and document classification baseline (B2). Relaxing evaluation where sub-stages are merged, e.g. AJCC stage IVA and AJCC stage IVB are mapped to AJCC stage IV, the relaxed system performance for training was at 0.68, 0.69, 0.55 F1 for AJCC, BCLC, and CLIP training, and 0.60, 0.50, and 0.43 F1 for test.
Table 2. Stage parameter document-level classification results.

<table>
<thead>
<tr>
<th>Label</th>
<th>TP</th>
<th>P</th>
<th>R</th>
<th>F1</th>
<th>Baseline F1</th>
<th>TP</th>
<th>P</th>
<th>R</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ascites</td>
<td>187</td>
<td>0.52</td>
<td>0.89</td>
<td>0.66</td>
<td>0.41</td>
<td>52</td>
<td>0.58</td>
<td>1.00</td>
<td>0.73</td>
</tr>
<tr>
<td>Child-Pugh</td>
<td>81</td>
<td>0.86</td>
<td>0.95</td>
<td>0.91</td>
<td>0.53</td>
<td>26</td>
<td>0.87</td>
<td>0.84</td>
<td>0.85</td>
</tr>
<tr>
<td>ECOG</td>
<td>135</td>
<td>0.98</td>
<td>0.69</td>
<td>0.81</td>
<td>0.61</td>
<td>28</td>
<td>0.97</td>
<td>0.61</td>
<td>0.75</td>
</tr>
<tr>
<td>Extrahepatic invasion</td>
<td>53</td>
<td>0.90</td>
<td>0.87</td>
<td>0.88</td>
<td>0.81</td>
<td>6</td>
<td>0.86</td>
<td>0.55</td>
<td>0.67</td>
</tr>
<tr>
<td>Hepatic encephalopathy</td>
<td>111</td>
<td>0.63</td>
<td>0.85</td>
<td>0.73</td>
<td>0.72</td>
<td>26</td>
<td>0.63</td>
<td>0.76</td>
<td>0.69</td>
</tr>
<tr>
<td>Macrovascular invasion</td>
<td>145</td>
<td>0.72</td>
<td>0.94</td>
<td>0.81</td>
<td>0.78</td>
<td>33</td>
<td>0.83</td>
<td>0.80</td>
<td>0.81</td>
</tr>
<tr>
<td>Metastasis</td>
<td>103</td>
<td>0.74</td>
<td>0.85</td>
<td>0.79</td>
<td>0.69</td>
<td>25</td>
<td>0.81</td>
<td>0.86</td>
<td>0.83</td>
</tr>
<tr>
<td>Portal hypertension</td>
<td>83</td>
<td>0.87</td>
<td>0.93</td>
<td>0.90</td>
<td>0.78</td>
<td>27</td>
<td>0.84</td>
<td>0.93</td>
<td>0.89</td>
</tr>
<tr>
<td>Tumor morphology</td>
<td>114</td>
<td>0.65</td>
<td>0.68</td>
<td>0.66</td>
<td>0.58</td>
<td>25</td>
<td>0.63</td>
<td>0.60</td>
<td>0.61</td>
</tr>
<tr>
<td>Tumor number</td>
<td>102</td>
<td>0.58</td>
<td>0.63</td>
<td>0.60</td>
<td>0.60</td>
<td>23</td>
<td>0.58</td>
<td>0.56</td>
<td>0.57</td>
</tr>
<tr>
<td>Tumor size</td>
<td>129</td>
<td>0.83</td>
<td>0.75</td>
<td>0.79</td>
<td>0.50</td>
<td>32</td>
<td>0.86</td>
<td>0.74</td>
<td>0.80</td>
</tr>
<tr>
<td>ALL</td>
<td>1243</td>
<td>0.70</td>
<td>0.80</td>
<td>0.75</td>
<td>0.62</td>
<td>303</td>
<td>0.73</td>
<td>0.76</td>
<td>0.74</td>
</tr>
</tbody>
</table>

Table 3. Stage parameter patient-level F1 classification results (Total(training)=160, Total(testing)=40)

<table>
<thead>
<tr>
<th>Label</th>
<th>System TP</th>
<th>System F1</th>
<th>Baseline TP</th>
<th>Baseline F1</th>
<th>System TP</th>
<th>System F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ascites</td>
<td>130</td>
<td>0.81</td>
<td>120</td>
<td>0.75</td>
<td>29</td>
<td>0.73</td>
</tr>
<tr>
<td>Child-Pugh</td>
<td>139</td>
<td>0.87</td>
<td>99</td>
<td>0.62</td>
<td>35</td>
<td>0.88</td>
</tr>
<tr>
<td>ECOG</td>
<td>135</td>
<td>0.84</td>
<td>96</td>
<td>0.60</td>
<td>32</td>
<td>0.80</td>
</tr>
<tr>
<td>Extrahepatic invasion</td>
<td>156</td>
<td>0.98</td>
<td>156</td>
<td>0.98</td>
<td>40</td>
<td>1.00</td>
</tr>
<tr>
<td>Hepatic encephalopathy</td>
<td>145</td>
<td>0.91</td>
<td>136</td>
<td>0.85</td>
<td>35</td>
<td>0.88</td>
</tr>
<tr>
<td>Macrovascular invasion</td>
<td>134</td>
<td>0.84</td>
<td>138</td>
<td>0.86</td>
<td>34</td>
<td>0.85</td>
</tr>
<tr>
<td>Metastasis</td>
<td>145</td>
<td>0.91</td>
<td>147</td>
<td>0.92</td>
<td>39</td>
<td>0.98</td>
</tr>
<tr>
<td>Portal hypertension</td>
<td>131</td>
<td>0.82</td>
<td>96</td>
<td>0.60</td>
<td>34</td>
<td>0.85</td>
</tr>
<tr>
<td>Tumor morphology</td>
<td>114</td>
<td>0.71</td>
<td>90</td>
<td>0.56</td>
<td>23</td>
<td>0.58</td>
</tr>
<tr>
<td>Tumor number</td>
<td>106</td>
<td>0.66</td>
<td>102</td>
<td>0.64</td>
<td>23</td>
<td>0.58</td>
</tr>
<tr>
<td>Tumor size</td>
<td>128</td>
<td>0.80</td>
<td>74</td>
<td>0.46</td>
<td>33</td>
<td>0.83</td>
</tr>
<tr>
<td>ALL</td>
<td>1463</td>
<td>0.83</td>
<td>1254</td>
<td>0.71</td>
<td>357</td>
<td>0.81</td>
</tr>
</tbody>
</table>

Table 4. Stage classifications results (Total(training)=160, Total(testing)=40)

<table>
<thead>
<tr>
<th>Stage</th>
<th>System TP</th>
<th>System F1</th>
<th>B1 TP</th>
<th>B1 F1</th>
<th>B2 TP</th>
<th>B2 F1</th>
<th>System TP</th>
<th>System F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>AJCC</td>
<td>103</td>
<td>0.64</td>
<td>87</td>
<td>0.54</td>
<td>86</td>
<td>0.54</td>
<td>22</td>
<td>0.55</td>
</tr>
<tr>
<td>BCLC</td>
<td>98</td>
<td>0.61</td>
<td>57</td>
<td>0.36</td>
<td>57</td>
<td>0.36</td>
<td>20</td>
<td>0.50</td>
</tr>
<tr>
<td>CLIP</td>
<td>88</td>
<td>0.55</td>
<td>53</td>
<td>0.33</td>
<td>48</td>
<td>0.30</td>
<td>17</td>
<td>0.43</td>
</tr>
</tbody>
</table>

In addition, we include a sensitivity analysis (Table 5) in which each patient-level stage parameter was substituted with the gold-standard value and the impact on F1 score was assessed. Performance increases in vital cancer parameter (e.g. metastasis, macrovascular invasion, tumor morphology, tumor number, tumor size) substitutions led to overall improvements in the F1 scores, (e.g. CLIP, tumor morphology improves +0.19 and +0.37 F1 for train and test). We also show the results of the opposite substitutions, where gold-standard values of stage parameters were
used excepting one patient-level system value. Tumor morphology and number were shown to degrade stage classifications the most (e.g. CLIP, tumor morphology degrading by -0.31 and -0.42 F1 for train and test).

Table 5. F1 stage classification results with sensitivity analysis. Stage parameter columns represent substitution with 1 gold/system standard with otherwise 10 other system/gold stage parameters used to calculate stage).

<table>
<thead>
<tr>
<th>REFERENCE</th>
<th>Ascites</th>
<th>Child-Pugh</th>
<th>ECOG</th>
<th>Extrabdominal invasion</th>
<th>Hepatic encephalopathy</th>
<th>Macrovacular invasion</th>
<th>Metastasis</th>
<th>Portal hypertension</th>
<th>Tumor morphology</th>
<th>Tumor number</th>
<th>Tumor size</th>
</tr>
</thead>
<tbody>
<tr>
<td>SYS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AJCC</td>
<td>0.64</td>
<td>--</td>
<td>--</td>
<td>+0.01</td>
<td></td>
<td>+0.10</td>
<td>+0.09</td>
<td>--</td>
<td>--</td>
<td>+0.12</td>
<td>+0.03</td>
</tr>
<tr>
<td>BCLC</td>
<td>0.61</td>
<td>--</td>
<td>+0.12</td>
<td>--</td>
<td>+0.06</td>
<td>+0.05</td>
<td>--</td>
<td>+0.08</td>
<td>+0.03</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CLIP</td>
<td>0.55</td>
<td>+0.05</td>
<td>--</td>
<td>--</td>
<td></td>
<td>+0.09</td>
<td>--</td>
<td>--</td>
<td>+0.19</td>
<td>--</td>
<td>--</td>
</tr>
</tbody>
</table>

| GOLD      |         |            |      |                        |                       |                       |            |                   |                 |             |           |
| AJCC      | 1.00    | -0.02      | -0.02| -0.02                  | -0.12                 | -0.11                 | -0.02      | -0.15             | -0.06           |             |           |
| BCLC      | 1.00    | -0.03      | -0.06| -0.13                  | -0.09                 | -0.07                 | -0.03      | -0.12             | -0.05           |             |           |
| CLIP      | 1.00    | -0.05      | -0.16| -0.05                  | -0.16                 | -0.05                 | -0.31      | -0.05             | -0.05           |             |           |

Error analysis

(1) Two-step regular expression extraction module:

The largest challenge with the regular expression module were hedge cases. For example, in “child-pugh score of 6 or 7” only “child-pugh 6” is captured, but “child-pugh 7” is not, though labeled in the gold standard. This similarly occurred for ECOG, such as in “ECOG 0-1”. The effect of these would cause underestimation of liver cancer stage. Child-Pugh errors included non-capitalized version of “child” such as in text evidence “child a”. We also miss cases for ambiguous acronyms such as “CPA”, which may stand for “cardiopulmonary arrest”, or “childhood physical abuse”. ECOG errors were mainly due to non-explicit evidence, e.g. “He [...] needs a wheelchair to walk”.

(2) Sentence classification module:

The sentence classification module had the most problems with lowly populated subcategories such as ascites-moderate/severe, as these categories often simultaneously had the most variation. For example, there are relatively fewer ways to say “no” in “no ascites” or “ascites none” compared to adjectives for severe cases such as “gross”, “extensive”, or “refractory”. In macrovascular invasion, the two other populated cases had more semantic variations expressing the blockages of blood vessels, e.g. “involvement”, “infiltrated”, “distending” or “occluded”. Disambiguation between variations and word-orders was another hurdle for normalize between major and minor blood vessels, e.g. “the portal vein” versus “superior mesenteric vein” or “anterior branches of the right portal vein”, which was necessary for classifying between macrovacular yes-major or yes-minor. Another problem was with correlated concepts. For example, though “diuretics” and “cirrhosis” are related to ascites (the former is used to treat, the latter is a cause of it), they are not exact surrogates to ascites. Other challenges include evidence that
relies on more complex inferences. For example, “he only complains of mild, occasional confusion” or “has well-compensated cirrhosis […] has not had any significant complications” informs about hepatic encephalopathy (mild in the first case, none in the second case).

(3) Tumor characteristics extraction module:
Several transformations were required for tumor morphology, tumor number, and tumor size factors, which allowed several opportunities for error introduction, e.g. errors due to relation classification or reference resolution, etc. We found the most challenging piece was the need for reference resolution. For example, determining the number of malignant tumors for tumor number requires reasoning over multiple sentences to, firstly, identify the malignancy status of each tumor and to, secondly, sum unique mentions of the same tumors. Tumor morphology, which depends on both determining the number of tumors and if >50% of the liver is invaded, was limited by the poor performance of ascertaining tumor number. Tumor size required automatically identifying which tumor mentions in radiology reports were malignant, which may not be expressed in a single sentence, before finding the maximum size.

(4) Patient-level stage parameter classifier:
Given the automatic annotations resulting in the previous extraction systems, there were several options to arrive at patient-level stage parameter values, such as taking the highest frequency versus highest value. For our system, we used the simple heuristic of taking the most severe finding for each category. Lower performance with gold-standard annotations of ECOG is because experts preferred to take least severe values during disambiguation.

(5) Patient-level stage classifier:
Using gold-standard annotation inputs for stage parameters for stage classification, we found performances of 0.98, 0.97, 0.96 F1 for AJCC, BCLC, and CLIP stages respectively. This signals that our rule-based algorithm was highly accurate. Unfortunately, cascading errors from upstream modules led to significant degradation of performances. Reviewing the sensitivity analysis in Table 6 reveals that improvements in certain stage parameters led to large gains in stage classification. For example, tumor number and tumor morphology were simultaneously crucial cancer stage parameters and were the most challenging stage parameters. Substituting gold-standard annotations for these classifications, performance improved tremendously. Conversely, errors would drop significantly after substitution of system values.

Conclusions
HCC staging is a complex task that requires reasoning over multiple sources for and with many variables. This work (1) systematically identifies information extraction subtasks for overall patient stage prediction, such as with stage parameter normalization; (2) provides several classes of solutions to identify and normalize stage parameters and stage patients, for example sentence classification vs. regular expression; and (3) characterizes challenges to the problem, most salient of which were reference resolution and patient-level classification. We have shown that several modules performed well, e.g. Child-Pugh and portal hypertension extractions and patient-level stage parameter prediction. Limitations of this work include that the data came from a relatively small group of patients from a single institution. Because some parts of the system included hand-crafted heuristics, our models and overall algorithms may not be generalizable to other datasets. Inclusion of other centers’ data would address unwarranted over-training of our algorithms for this particular study. Our data was skewed towards patients in the early and intermediate stages as patients with late stages were less likely to be referred for treatment. The strength of this work includes is trace-ability and modular design. We envision applications of this algorithm to support accelerated staging by highlighting supporting evidence in text and outputting classifications along with the uncertainty of its predictions. Human reviewers, given highlighted information, may then review uncertain classifications. Additional labeled data and improvements in various components, in turn, will lead to improvements in the overall system. In this work, we contributed insights into various issues with complex patient phenotype classification using primarily text data. Future work includes expanding the dataset and testing our algorithms on other institution datasets.

Acknowledgements
This project was funded by the National Institutes of Health, National Center for Advancing Translational Sciences (KL2 TR000421 and the University of Washington Institute of Translational Health Sciences (UL1TR000423).

References


13. Ashish N, Dahm L, Boicey C. University of California, Irvine-Pathology Extraction Pipeline: The pathology extraction pipeline for information extraction from pathology reports. Health Informatics J. 2014 Aug 25;


Talking About My Care: Detecting Mentions of Hormonal Therapy Adherence Behavior in an Online Breast Cancer Community

Zhijun Yin, MS¹, Wei Xie, MS¹, and Bradley A. Malin, PhD¹
¹Vanderbilt University, Nashville, Tennessee, USA

Abstract

Hormonal therapy adherence is challenging for many patients with hormone-receptor-positive breast cancer. Gaining intuition into their adherence behavior would assist in improving outcomes by pinpointing, and eventually addressing, why patients fail to adhere. While traditional adherence studies rely on survey-based methods or electronic medical records, online health communities provide a supplemental data source to learn about such behavior and often on a much larger scale. In this paper, we focus on an online breast cancer discussion forum and propose a framework to automatically extract hormonal therapy adherence behavior (HTAB) mentions. The framework compares medical term usage when describing when a patient is taking hormonal therapy medication and interrupting their treatment (e.g., stop/pause taking medication). We show that by using shallow neural networks, in the form of word2vec, the learned features can be applied to build efficient HTAB mention classifiers. Through medical term comparison, we find that patients who exhibit an interruption behavior are more likely to mention depression and their care providers, while patients with continuation behavior are more likely to mention common side effects (e.g., hot flashes, nausea and osteoporosis), vitamins and exercise.

Introduction

Breast cancer is the most prevalent cancer among American women¹ and the second leading cause of death among women with cancer (just behind lung cancer)². It is estimated that close to 12% of American women will eventually develop invasive breast cancer during their lifetime³. A common initial treatment for breast cancer is surgical intervention (e.g., lumpectomy or mastectomy), while adjuvant therapy (i.e., treatment after surgical intervention) is often invoked to reduce the risk of cancer recurrence⁴. In particular, hormonal adjuvant therapy is a popular treatment with a proven track record of significantly improving the long-term survival rate of patients with hormone-receptor-positive breast cancer⁵. This is notable because this disease subtype comprises 75% of all breast cancer cases¹. To maximize this benefit of hormonal therapy, patients are prescribed a regimen of medication that is expected to continue for a minimum of five years⁶. For instance, taking tamoxifen (an oral hormonal therapy drug) for five years reduces breast cancer mortality by 33% in the decade after initial treatment⁷. Moreover, more recent evidence⁸ suggests that maintaining a tamoxifen regimen for an additional five years can further reduce mortality by approximately 50%.

Despite the benefit of hormonal therapy for women, only around half complete a full five-year treatment⁹. There can be various reasons why breast cancer patients would fail to complete the regimen, ranging from adverse side effects¹⁰ to progression of the disease into a terminal form¹¹,¹². Still, there are many women who fail to stay on a recommended regimen for less obvious reasons¹⁰. As such, learning the factors associated with why women choose to stop (as well as stay on) hormonal therapy is critical to improving a patient’s treatment experience. While there have been various investigations into regimen adherence¹¹,¹³–¹⁷, most studies rely on traditional clinical resources and methodology, such as formal survey-based studies¹³,¹⁵,¹⁶,¹⁸, electronic medical records (EMRs) and other clinical resources¹¹,¹⁷,¹⁹. Though such traditional methods and data are valuable in healthcare research, there are certain drawbacks that should be recognized. In particular, survey-based methods are limited in that they typically incur high costs in time and money, often restricting a study to a smaller number of participants. Moreover, considering that breast cancer patients with hormonal therapy generally only have follow-up with their doctors every six months, this leads to a large information and time gap in traditional EMR systems about the patients’ treatment (e.g., their feelings and experiences) between two visits.

The Internet, and social media in particular, has provided patients with the opportunity to seek and share treatment experiences in online environments. For instance, the BreastCancer.org website maintains an online discussion board for breast cancer patients to discuss any aspect of their daily lives they deem relevant. This includes, but is not limited to, their concerns, diagnoses, treatments, side effects and social support structure. This self-reported information
provides a new opportunity to learn about breast cancer patients’ treatment adherence - and on a much larger scale. With thousands of patients posting and interacting regularly and accumulating tens of thousands or greater (up to millions) of posts on discussion boards such as BreastCancer.org, one immediate research challenge that arises is how to efficiently leverage such rich text, ideally in an automated and less labor-intensive manner. More concretely, notable research challenges in this domain include: 1) mapping behavioral and health research questions to ad hoc self-reported information and 2) extracting useful knowledge from an environment that lacks formal mediation where true signals are often hidden in a vast amount of noise (e.g., patients can discuss anything they wish, including topics that are not directed related to cancer).

In this paper, we aim to develop a machine learning framework to distinguish mentions of hormonal therapy adherence behavior (HTAB) from other less relevant free-text contents in online health forums. In particular, we are interested in studying patient behaviors (and their associated factors), such as taking a prescribed medication or interrupting treatment (e.g., stopping or pausing a regimen, or switching to a different medication). In our framework, the task of distinguishing mentions and non-mentions of HTAB is cast as a classification problem. To maximize the predictive performance of our framework, we extensively adapt and compose preprocessing and feature engineering techniques, as well as to interpret their effects. Our framework demonstrates that, through applying natural language processing and machine learning techniques, we can obtain an effective classifier to automatically detect mentions (and non-mentions) of hormonal therapy treatment adherence behaviors. Finally, we perform content analysis (through medical terms) to gain insight into the factors affecting how people communicate taking medication behavior and interrupting medication behavior.

Our work contributes to the field of user (or patient) generated online data (e.g., in social platforms and discussion communities), specifically where it is applied to supplement traditional data sources (e.g., EMRs) to study health related problems. In this research domain, we acknowledge that there is a growing collection of studies that cover a range of areas, including flu trends20, mental health11,12, privacy issues about health mentions21,22 as well as how to build online communities to provide local cancer support23. Further, regarding this specific research topic, Freedman et al.24 studied a large number of posts mentioning cancer treatments (including hormonal therapy) and identified treatment barriers that manifest from various aspects, including emotions, preferences and religious belief. Mao et al.25 found that joint pain is the main reason patients stop taking Aromatase Inhibitors (AIs) treatment in online discussions of drug side effects. There have also been several studies that focus on BreastCancer.org, as discussed in a recent review26, though the focus has been on different prediction problems.

**Methods**

Our goal is to build an automatic framework to distinguish HTAB status (mentions and non-mentions) and learn their associated factors. Figure 1 shows the three main components of the proposed framework: 1) data preparation, 2) classifier building and 3) content analysis. Specifically, free-text data from users’ postings are first collected from the hormonal therapy forum in breastcancer.org online discussion board. This yields a large amount of unlabeled text. Next, a subset of sentences containing at least one of seven common hormonal therapy medication keywords (e.g., Tamoxifen) are manually labeled based on their contents through a majority voting model. The labeled sentences are then applied to fit several candidate classifiers and the model with the best performance is applied to boost the number of labeled data. Finally, after extracting different HTAB, a regression analysis is applied to study associated factors.

**Data sources**

The online breast cancer community studied in this paper is maintained by Breastcancer.org, a non-profit organization that provides information about breast cancer. There are more than 170,000 registered members and 80 main forums in this online health community. Each forum is organized into different threads. Each thread has an initial post indicating a general topic (e.g., a question) that will be discussed (asked), with zero or more posts that follow the initial post. However, the posts that follow do not necessarily respond the initial post and could simply respond to each other. Additionally, certain threads may have a very small number of posts, while others could span years and consist of hundreds of posts. In this paper, we focus on one particular forum, Hormonal Therapy - Before, During and After1.

---

1https://community.breastcancer.org/forum/78/
Figure 1: Framework for studying HTAB through an online breast cancer forum data. Three core components are highlighted in the figure: 1) data preparation, 2) HTAB mention classifier and 3) HTAB comparison.

We collected all the posts published in this hormonal therapy forum before June 22, 2016. There are 9,996 users who participated in 5,995 threads with more than 130,000 posts over a 9-year period.

Data Annotation

In this study, we rely on supervised machine learning to distinguish mentions from non-mentions of HTAB. To engineer efficient classifiers, we manually annotated a gold standard dataset. Considering the rich context of the posts (e.g., posts in this forum can consist of multiple paragraphs), we focus on features at the sentence-level for the classification task. This is because we observe that most sentences used in this forum follow general grammatical rules and are sufficiently verbose to convey information of interest. However, there still exists a substantial number of irrelevant sentences and building a classifier on such an unbalanced dataset would seriously affect its performance. As such, we further constrain our classification objects to those sentences containing at least one of seven common hormonal therapy medication keywords (Arimidex, Aromasin, Femara, Tamoxifen, Evista, Fareston and Faslodex). We also extend the keywords to include their corresponding generic names (Anastrozole, Exemestane, Letrozole, Tamoxifen, Raloxifene, Toremifene, and Fulvestrant, respectively) to account for variation in communication.

The goal of annotation is to obtain binary labels that indicate whether a given sentence communicates an HTAB or not. However, directly providing such labels to annotators may not be sufficient for them to understand the task. Thus, we adopted the hierarchical labeling strategy that was applied in21. Specifically, based on our observation of how patients discuss their treatments in the forum and guidance in a decision making codebook introduced by Beryl and colleagues in a hormonal therapy survey16, we provide seven options for annotators to choose from to best describe whether a given sentence, containing at least one of the seven common medication keywords as mentioned above, conveys information about: 1) taking medication action, 2) stopping taking medication, 3) switching medication action, 4) taking medication intent, 5) not taking medication intent, 6) not decided yet, and 7) none of the above. The first three options correspond to mentions of HTAB, while the other four options belong to non-mentions of HTAB. We randomly select 1000 sentences and assign them to each of two annotators who are familiar with the online discussion forum to obtain labels. A third annotator is called upon if the first two annotators do not agree in their labels. The final label of each sentence is based on a majority rule.

HTAB Classification

N-gram features. There are two important steps in a typical classifier fitting process to note: feature extraction and algorithm tuning. While both steps can influence the performance of a classifier, in this task we focus on feature
extraction with several common off-the-shelf learners. In text mining and natural language processing, term frequency-inverse document frequency (TF-IDF) is a common technique to help extract features. As the name suggests, the intuition behind this concept is that if a term is frequent in a document (high TF) but scarce in other documents (high IDF), then the term will be distinctive to this document. As a result, the higher the TF-IDF value that a term exhibits, the more important the term is to distinguish the document from others. A natural choice for a term is a single word in a document. However, using words as features to represent a document loses the power in the ordering of the words. An alternative solution is to apply an n-gram word combination (i.e., linking n adjacent words together) as a term (e.g., "hot flash" as a 2-gram). Further, to reduce the impact of sparsity in natural language, the term could also be an n-gram of characters. Compared to other traditional high dimensionality reduction techniques (e.g., PCA), the advantage of this technique is that it maximizes interpretability.

**Low dimensional representation features.** Recently, another popular technique aimed at high dimensionality reduction in text mining is based on learning a low-dimensional representation of a word. These techniques make use of large amounts of data to discover the semantic similarities between terms through an embedding. For instance, word2vec is a technique where each word is represented as a vector of a predefined dimension (e.g., 100). By building a model to predict terms through their neighbors, the model has been shown to learn vector representations of terms quite well. For example, when applying our collected hormonal therapy forum data to fit a word2vec model for tamoxifen, the ten most similar words provided by the model (based on cosine similarity) were: taxmox (0.97), tam (0.91), arimidex (0.89), tamoxifin (0.88), tamo (0.87), femara (0.85), aromasin (0.79), anastrozole (0.78), letrozole (0.77) and armidex (0.77). The model provides semantically similar words (e.g., other hormonal therapy medications such as femara, aromasin and anastrozole), as well as its common misspellings (e.g., taxmox, tam and tamo). This is notable from a natural language processing perspective because they serve as a different type of feature to fit a classifier, while accounting for misspellings and abbreviations, which is common in online environments.

**Classification Design.** In this paper, we extract both n-gram words and n-gram characters from sentences as features (2 ≤ n ≤ 5). We also introduce the mean of the vector representation (word2vec) of words in a sentence as another type of feature. For classification, we rely on logistic regression (LR), a SVM classifier (SVC) and a random forest classifier (RFC). To compare the performance of different classifiers, we rely on stratified 10-fold cross-validation (CV), whereby for each 10 times of randomly data shuffling, we use 900 labeled sentences (out of a total of 1000) for training and the remaining 100 for testing. Note that CV can be applied for either 1) model parameter tuning to avoid overfitting or 2) performance evaluation to mitigate the effect of randomness. While it is rigorous to create another test data set from all of the collected corpus to evaluate the models performance, we did not apply CV to tune the hyperparameters in the model. Instead, we directly applied it to test the model. We report the average accuracy, precision, recall, F1 score and area under the receiver operating characteristic curve (AUC), along with their standard deviations. We also applied a t-test to ascertain if there is a significant difference in the performance of the classifiers.

We trained the word2vec model by feeding all of the posts we collected (after removing the labeled sentences) into an implementation of gensim (version 0.31.0). We empirically set the dimensionality of the model to 100 and removed all of the words with frequency smaller than 5. The LR, SVC and RFC models were trained using scikit-learn (version 0.18), where the hyperparameters were left as the default in the package.

**HTAB Comparison**

In this task, we aim to boost the amount of the labeled data using a machine learning technique. As such, we can continue to extract HTAB and conduct further behavior analysis. We will subsequently apply the classifier with the best precision to detect HTAB mentions from the remaining unlabeled sentences. In this paper, we mainly focus on two types of HTAB: 1) taking: where a user claimed that she is under treatment with some hormonal therapy medication, and 2) interruption: where a user claimed that she stops (or pauses) a regimen, or switches to another different medication. We use patterns that we observed during labeling process to filter these two types of HTAB. For instance, we apply patterns that include (but are not limited to) started, been on and stay on to extract taking HTAB, and apply patterns such as switched/switch, took me off, back on and stopped taking to extract interruption HTAB.

After filtering the sentences, we retained the posts mentioning either a taking behavior or an interruption behavior (but not both). While people may have different reasons for interrupting a treatment, and they may even talk about the same
behavior multiple times in the forum, in this study we do not aggregate posts published by the same user on the same type of behavior. This is because, when discussing a behavior such as taking, a person’s feelings and health conditions may change with time. Moreover, a person may discuss their taking behavior with respect to different medications. Thus, to obtain meaningful concepts, we rely on ADEPT, a conditional random field (CRF) based classifier aimed to identify medical terms from patient-authored text. Due to space limitations, we refer the reader to the original publication for details on model performance. As such, we removed non-medical related terms from further analysis. However, even if we focus on medical concepts only, more than 20,000 medical terms would remain. Thus, we applied two strategies to mitigate the high dimensionality problem: 1) we empirically retain the top 2,000 medical terms based on their TF-IDF values and 2) we use LR with lasso regularization, where posts mentioning an interruption behavior form the positive class and posts mentioning a taking behavior form the negative class. While lasso can serve as a feature selection approach by forcing the coefficients of many terms to zero, it has been found to be unstable. This is because, given a different sample and random state, it may result in different coefficients. To resolve this problem, we adopt stable selection. The basic idea behind this technique is to subsample the training data and fit a lasso regularized LR model, whereby the features that are repeatedly selected across multiple runs of randomization receive higher scores (with a range from 0 to 1, the higher, the better). However, the scores cannot communicate which behavior the corresponding terms are informative for. Thus, to obtain directionality, we rely on the Pearson biserial correlation between the terms and the two behaviors. We apply the implementation of stable selection sklearn (version 0.18) and the implementation of Pearson biserial correlation in scipy (version 0.81), with all possible hyperparameters set to their default.2

Results

Data Annotation

Table 1 shows the distribution of labeling results. There is an approximately even number of HTAB mentions and non-mentions. The two primary annotators exhibited a very good agreement level (Cohen’s $\kappa = 0.80$) at the mentions vs. non-mentions level and a good agreement (Cohen’s $\kappa = 0.72$) at the seven detailed options level. For the sentences designated as an action-related option, approximately 80% are about taking hormonal therapy medication, while only about 8% are about discontinuing hormonal therapy medication and 12% are about switching hormonal therapy medications.

Model Performance

Table 2 shows the performance of models fitted with various features and algorithm combinations. For each measure, the mean values and the standard deviations are reported. The maximal mean value of each measure is highlighted in blue. There are several notable findings to recognize.

First, models fit with word2vec features have a significant improvement over models fit with either single word or word n-gram features for almost all of the five metrics. For instance, in comparison to RFC fit with word features, RFC fit with word2vec had a 13.9% improvement on AUC, 11.9% improvement on precision, 21.5% improvement on recall, 16.2% improvement on F1 score, and 12.8% improvement on accuracy. This suggests word2vec features have a positive influence on HTAB mention detection.

2 Recently, open-vocabulary approaches has become popular in natural language processing. In particular, while topic modeling is appealing and has the potential to complete a similar task, its results are often difficult to interpret. Further, the number of topics has to be determined ahead. Thus, we focused on medical terms because they have domain-specific meaning and their interpretation is relatively straightforward.

---

<table>
<thead>
<tr>
<th>Option</th>
<th>HTAB Mention</th>
<th>No Mention of HTAB</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Action</td>
<td>Intent</td>
</tr>
<tr>
<td></td>
<td>Taking</td>
<td>Stop</td>
</tr>
<tr>
<td></td>
<td>Switch</td>
<td>Take</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Not Taking</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Undecided</td>
</tr>
<tr>
<td></td>
<td></td>
<td>None-of-Above</td>
</tr>
<tr>
<td>#Sent.</td>
<td>403</td>
<td>41</td>
</tr>
<tr>
<td></td>
<td>62</td>
<td>40</td>
</tr>
<tr>
<td></td>
<td></td>
<td>25</td>
</tr>
<tr>
<td></td>
<td></td>
<td>33</td>
</tr>
<tr>
<td></td>
<td></td>
<td>396</td>
</tr>
</tbody>
</table>

Table 1: The distribution of options in the 1000 labeled sentences (after the third annotator broke ties).
First, hormonal therapy medications, such as Aromatase Inhibitors (AI) (e.g., femara, exemestane (aromasin),
and arimidex (anastrozole)), are more likely to be mentioned in posts related to an interruption behavior.

Second, models fit with character n-gram features exhibited significant improvement in AUC over models
fitted with either single word features or word n-gram features ($p < 0.01$). As an example, the improvement for
RFC was $8 \sim 9\%$. Further, the LR model fitted with character n-gram features significantly improved on the
LR model fitted with single word features on all measures ($p < 0.01$).

Third, while the LR model fitted with character n-gram features have a similar performance to the LR model fitted
with word2vec features, the latter improved both RFC and SVC on AUC, precision, F1 score and accuracy.

However, given the same type of features (e.g., either n-gram or word2vec features), there is no significant difference
among the algorithms on the majority of the measures. It should also be noted that SVC with either single word features
or character n-gram features obtained a perfect recall but with a severe cost of precision (almost equivalent to
a random guess).

Based on Table 2, we can conclude that, for HTAB detection, word2vec features result in the best performance,
followed by character n-gram features, and then word related features. Based on this finding, we chose to apply RFC
fit with word2vec (with the highest average precision) to classify all of the remaining unlabeled sentences for further
analysis. Moreover, the smaller standard deviation suggests that RFC is not overfitting.

### Comparing Medical Terms Between HTAB

After filtering with the taking and interruption behavior patterns and extracting the medical terms, we obtained
19,174 posts published by 5,251 users. Among those posts, 13,461 (70.2\%) discuss a taking behavior (the negative class),
while 5,713 (29.8\%) discuss an interruption behavior (the positive class). Among these users, 4,548 (86.7\%) men-
tion a taking behavior, 1,961 (37.3\%) mention an interruption behavior, and 1,258 (24.0\%) mention both taking and
interruption behaviors.

After applying stable selection through 200 randomly generated subsamples, Table 3 shows the 50 medical terms
that were most important for distinguishing between posts that mention a taking behavior and an interruption behavior.
The Pearson biserial correlation between all of the terms was significant at the 0.01 level ($p < 0.01$). The goodness of
fit measured with AUC is 0.765 \pm 0.004, suggesting that the lasso models used in stable selection fit the observations
well. Here, there are several notable results to highlight.

First, hormonal therapy medications, such as Aromatase Inhibitors (AI) (e.g., femara, exemestane (aromasin),
and arimidex (anastrozole)), are more likely to be mentioned in posts related to an interruption behavior. It should also be
noted that some users may switch between different AIs, as stated by one user:

**Example 1** “... I had this problem with Femara and was taken off and switched to Aromasin ...”
<table>
<thead>
<tr>
<th>Rank</th>
<th>Score</th>
<th>Term</th>
<th>Behavior</th>
<th>Rank</th>
<th>Score</th>
<th>Term</th>
<th>Behavior</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.0</td>
<td>femara</td>
<td>interruption</td>
<td>26</td>
<td>0.515</td>
<td>patient</td>
<td>interruption</td>
</tr>
<tr>
<td>2</td>
<td>1.0</td>
<td>exemestan</td>
<td>interruption</td>
<td>27</td>
<td>0.5</td>
<td>recurr</td>
<td>interruption</td>
</tr>
<tr>
<td>3</td>
<td>1.0</td>
<td>drug</td>
<td>interruption</td>
<td>28</td>
<td>0.495</td>
<td>carpel</td>
<td>interruption</td>
</tr>
<tr>
<td>4</td>
<td>1.0</td>
<td>aromasin</td>
<td>interruption</td>
<td>29</td>
<td>0.49</td>
<td>sweat</td>
<td>taking</td>
</tr>
<tr>
<td>5</td>
<td>1.0</td>
<td>arimidex</td>
<td>interruption</td>
<td>30</td>
<td>0.48</td>
<td>menopaus</td>
<td>interruption</td>
</tr>
<tr>
<td>6</td>
<td>1.0</td>
<td>ai</td>
<td>interruption</td>
<td>31</td>
<td>0.48</td>
<td>induc</td>
<td>interruption</td>
</tr>
<tr>
<td>7</td>
<td>0.985</td>
<td>pain</td>
<td>interruption</td>
<td>32</td>
<td>0.47</td>
<td>vacat</td>
<td>interruption</td>
</tr>
<tr>
<td>8</td>
<td>0.95</td>
<td>chemo</td>
<td>taking</td>
<td>33</td>
<td>0.47</td>
<td>tylenol</td>
<td>taking</td>
</tr>
<tr>
<td>9</td>
<td>0.945</td>
<td>radiat</td>
<td>taking</td>
<td>34</td>
<td>0.465</td>
<td>med</td>
<td>interruption</td>
</tr>
<tr>
<td>10</td>
<td>0.81</td>
<td>calcium</td>
<td>taking</td>
<td>35</td>
<td>0.46</td>
<td>exercis</td>
<td>taking</td>
</tr>
<tr>
<td>11</td>
<td>0.76</td>
<td>flash</td>
<td>taking</td>
<td>36</td>
<td>0.455</td>
<td>lumpectom</td>
<td>taking</td>
</tr>
<tr>
<td>12</td>
<td>0.745</td>
<td>hot</td>
<td>taking</td>
<td>37</td>
<td>0.45</td>
<td>trigger</td>
<td>interruption</td>
</tr>
<tr>
<td>13</td>
<td>0.735</td>
<td>hormon</td>
<td>interruption</td>
<td>38</td>
<td>0.445</td>
<td>improv</td>
<td>interruption</td>
</tr>
<tr>
<td>14</td>
<td>0.68</td>
<td>sever</td>
<td>interruption</td>
<td>39</td>
<td>0.435</td>
<td>wellbutrin</td>
<td>interruption</td>
</tr>
<tr>
<td>15</td>
<td>0.655</td>
<td>depress</td>
<td>interruption</td>
<td>40</td>
<td>0.43</td>
<td>reaction</td>
<td>interruption</td>
</tr>
<tr>
<td>16</td>
<td>0.62</td>
<td>anastrozol</td>
<td>interruption</td>
<td>41</td>
<td>0.43</td>
<td>osteopenia</td>
<td>taking</td>
</tr>
<tr>
<td>17</td>
<td>0.565</td>
<td>hair</td>
<td>taking</td>
<td>42</td>
<td>0.43</td>
<td>estrogen</td>
<td>interruption</td>
</tr>
<tr>
<td>18</td>
<td>0.56</td>
<td>onc</td>
<td>interruption</td>
<td>43</td>
<td>0.43</td>
<td>anti</td>
<td>interruption</td>
</tr>
<tr>
<td>19</td>
<td>0.55</td>
<td>fog</td>
<td>interruption</td>
<td>44</td>
<td>0.425</td>
<td>bilat</td>
<td>interruption</td>
</tr>
<tr>
<td>20</td>
<td>0.545</td>
<td>therapi</td>
<td>interruption</td>
<td>45</td>
<td>0.42</td>
<td>vit</td>
<td>taking</td>
</tr>
<tr>
<td>21</td>
<td>0.54</td>
<td>oncologist</td>
<td>interruption</td>
<td>46</td>
<td>0.42</td>
<td>period</td>
<td>taking</td>
</tr>
<tr>
<td>22</td>
<td>0.53</td>
<td>surgeri</td>
<td>interruption</td>
<td>47</td>
<td>0.415</td>
<td>ekg</td>
<td>interruption</td>
</tr>
<tr>
<td>23</td>
<td>0.53</td>
<td>inhibitor</td>
<td>interruption</td>
<td>48</td>
<td>0.41</td>
<td>ultrasound</td>
<td>taking</td>
</tr>
<tr>
<td>24</td>
<td>0.53</td>
<td>effect</td>
<td>taking</td>
<td>49</td>
<td>0.41</td>
<td>cortison</td>
<td>interruption</td>
</tr>
<tr>
<td>25</td>
<td>0.525</td>
<td>nausea</td>
<td>taking</td>
<td>50</td>
<td>0.405</td>
<td>methotrex</td>
<td>interruption</td>
</tr>
</tbody>
</table>

Table 3: 50 medical terms that are the most useful in distinguishing posts mentioning taking behavior from posts mentioning interruption behaviors. The Pearson biserial correlation between all of these terms is significant at a level of 0.01 ($p < 0.01$).

Second, common side effects of hormonal therapy medications are more likely to be mentioned with a taking behavior, such as hot flashes, hair (loss), nausea, sweating, and osteoporosis. Common drugs or supplements mentioned with a taking behavior include Tylenol, vitamin and calcium. By contrast, depression, pain, and fog are often likely to be mentioned with an interruption behavior. As a user complained:

**Example 2** “This spring I switched to Exemestane due to terrible depression, sleep issues, back/shoulder/neck pain.”

It is not surprising to see that Wellbutrin, an antidepressant, is also more likely be mentioned with an interruption behavior. Note that methotrexate, a medication to treat cancer and “usually given after other medications have been tried without successful treatment of symptoms” is also more likely to be mentioned with an interruption behavior. Heart-related terms such as carpel, cortison and ekg are also more likely to be mentioned with an interruption behavior.

Third, there are some types of terms that are mentioned with only one of the two types of behaviors. For instance, professionals (e.g., oncologist or onc) are more likely to be mentioned with an interruption behavior. This suggests that these users’ interruption behavior may be associated with their physicians and possibly with permission (or the suggestion of) other care professionals. As one user stated:

**Example 3** “In the 3 years I’ve been on Femara, I have negotiated 2 one-month “vacations” from Femara with my onc when the S/Es got too bad ...”

People are more likely to mention exercise when discussing their taking behavior, but may mention recurrence when talking about an interruption behavior. As one user said:

**Example 4** “I actually felt so bad on tamoxifen that I went off of it last year, then had recurrence 4 months later ...”
Finally, people with an interruption behavior tend to mention surgery, which may be a possible reason why they pause a medication. People with a taking behavior are more likely to mention lumpectomy, ultrasound, chemo, radiation, and period.

Discussion

Findings. In this study, we built a framework to learn about hormonal therapy adherence behaviors, with a focus on classifier design and content comparison between taking and interruption behaviors. There are two notable contributions in this work: 1) we find that features based on embeddings (e.g., word2vec) can assist in establishing efficient detectors of HTAB mentions. For instance, compared to an RFC model fit with single word features, an RFC model fit with word2vec features resulted in a ~21% improvement of recall and a ~12% improvement on precision. The performance of the model ensures high quality for the extracted content; 2) by focusing on medical terms, we gain insights into the different factors associated with taking and interrupting hormonal therapy treatment behavior. For example, we find that people with an interruption behavior tend to mention depression and related antidepressant. This is in alignment with other studies\textsuperscript{36,37} where depression was found to be significantly associated with non-adherence to hormonal therapy treatment. Our findings further suggest that certain common side effects (e.g., hot flashes, nausea and osteoporosis) may not be as likely to induce an interruption behavior severe as depression.

It is further interesting to note that people exhibiting an interruption behavior are likely to mention their care professionals. This suggests that their interruption behavior may be an artifact of, or even suggested by, their care professionals (e.g., take a break due to the following surgery). This is noteworthy because this interruption behavior might still be under their care professionals’ control.

Implication. By relying on natural language processing techniques, machine learning models and statistical inference tools, we built an automated framework to study treatment adherence through an online breast cancer community. Given that patients with long-term hormonal therapy tend to see an oncologist only twice per year, our framework provides a supplemental perspective (beyond routine clinical information) to learn about these patients’ treatment experiences. Our findings, which are based on a comparison of medical term predictive ability, demonstrate the potential power of this type of online data to conduct treatment adherence research. Moreover, this framework may be extendable to assist in the study of adherence for other chronic diseases (e.g., depression or diabetes) through patient-authored online data. However, we acknowledge that domain-specific knowledge would be needed to customize behavior patterns for each disease. Still, once such information has been generated, the framework proposed in this paper could easily be adopted.

Limitation and further work. There are certain limitations that we highlight, which can serve as guidance for future research. First, we did not tune the hyperparameters of the models. It will be useful to determine if a combination of the proposed features can boost model performance. Second, we applied our observed patterns to filter the taking and interruption behaviors. While this strategy leads to high precision, it will miss adherence behavior mentions that fail to follow these patterns. Thus, as a next step, we anticipate including posts that communicate both taking and interruption behaviors with algorithms that achieve higher fidelity. Finally, we extracted medical terms without consideration for the grammar and context in which they are situated. We believe that more efficient models may be develop to detect different types of behaviors and interpretable medical concepts.

Conclusion

In this paper, we proposed a framework to learn about hormonal therapy treatment adherence behaviors (HTABs) through an online breast cancer discussion forum. The framework consists of three core components: 1) data preparation, 2) classifier engineering and 3) comparison of HTABs. We analyzed a dataset consisting of over 130,000 posts across a 9-year period and demonstrated that features based on embeddings (e.g., word2vec) can help build a more efficient and effective classifier for HTAB mentions in online generated data. Furthermore, by comparing the predictive capability of medical terms used in describing taking and interruption behaviors, we discovered that people with an interruption behavior are more likely to mention depression and their care professionals, while people with a taking behavior are more likely to mention common side effects (e.g., hot flashes, nausea and osteoporosis), vitamins and exercise. This study demonstrates that an individual’s discussion of hormonal therapy in an online environment may
provide insight into treatment adherence behaviors. We further believe that this framework has the potential for extension to learn adherence for other chronic diseases treatment (e.g., diabetes and depression), provided domain-specific guidance is available.

Acknowledgements
This research was sponsored by grant IIS1418504 of the National Science Foundation.

References


[23] Weiss JB. Building an online community to support local cancer survivorship: combining informatics and participatory action research for collaborative design. Vanderbilt University; 2009.


Initial Usability Evaluation of a Knowledge-Based Population Health Information System: The Population Health Record (PopHR)

Mengru Yuan, MScPH1, Guido Powell, MSc1, Maxime Lavigne, MSc1, Anya Okhmatovskaia, PhD1, David L. Buckeridge, MD PhD1

1Department of Epidemiology, Biostatistics, and Occupational Health, McGill University, Montréal, Québec, Canada

Abstract

We report the baseline usability of a novel web-based application, the Population Health Record (PopHR), designed to facilitate the effective use of population health information by public health professionals and to support evidence-based decision-making. The usability test was conducted with ten potential users who each completed eight tasks using the PopHR system. Participant responses were recorded, including timestamps for each data entry. Overall, the task completion rate was 96% while the success rate was 88%. The average time-on-task was 3.11 minutes, with more time spent on tasks requiring a user to stratify data along multiple dimensions, such as age, sex, or geographical region. Usability scores indicated that the current version of PopHR has good usability. Potential improvements identified included adding supporting information, offering different visualizations, and enhancing system stability. These findings are examples of addressable usability problems encountered in developing a population health record system.

Introduction

Despite a broad consensus that evidence-based decision-making can improve the effectiveness of the public health system, decision-makers are currently unable to access information and combine it with evidence to make important choices about the health of their communities1. Information is scattered across ‘siloes’ or multiple systems1 and is difficult to extract2. Evidence regarding interventions is generally not integrated into information systems, and is available only through online repositories of systematic reviews3. Paradoxically, practitioners suffer from information overload4 while being unable to find the information they need2 and align it with evidence5.

These problems with using information and evidence in practice indicate that existing information systems do not effectively address important barriers to evidence-based public health6-8. To address these barriers, professional bodies9 and expert panels1 have called for the creation of a new breed of information systems, called Population Health Records (PopHR), which are designed using principles from cognitive science and biomedical informatics to improve decision-making in the public health system10. We designed a knowledge-based system to meet these requirements11 and developed a software application according to this design12. The first version of the system is now ready for use and we are implementing the application with our public health partners for initial use by early adopters. Here we report on the first formal usability evaluation of the PopHR software.

Usability testing is essential for user-centered design as it allows the designer to determine if their conceptual model of the application is effectively conveyed to users through the interface13. A well-accepted definition of usability is “the degree to which a product or system can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use”14. Thus, usability can be measured by the accuracy and completeness of users solving specified tasks, the effort required to achieve goals effectively, and the degree of satisfaction with using the system. Systems with poor usability may result in user errors, a waste of resources, and low adoption rates15, 16, therefore, assessing the usability early in the design process and enhancing the system based on user feedback is critical.

Although population-based health systems exist, such as CDC WONDER17, Infocentre18 and Snapshots19, these systems have been assessed through user satisfaction surveys only20, no usability evaluations were conducted. In the biomedical informatics literature, usability evaluations have been reported mainly for the clinical and hospital information systems21, with few examples of usability challenges and solutions for population health systems. The goal of an information system in clinical medicine is to assist clinicians in tasks such as diagnosis and management so that they can deliver optimal healthcare in a high volume and time critical working context. In contrast, the goal in population health is to support practitioners in monitoring the health status of a defined population so that they can deploy and manage, efficient, effective, and equitable programs to promote health and prevent diseases22. As the goals, tasks, and context in public health differ from clinical medicine, the absence of evidence about the usability of population health information systems is an important gap. Therefore, we conducted a user-centered usability evaluation of the current version of the PopHR system. The objective of this study was to measure the effectiveness,
efficiency and end-user satisfaction of the PopHR system, to identify usability issues, and to inform the future development of the PopHR and other population health information systems.

**Method**

*PopHR description*

PopHR is a semantic web platform, designed to organize, make interpretable, and disseminate information and evidence to support population-health decision makers. It centralizes multiple sources of heterogeneous data, estimates indicators of population health, and uses epidemiological knowledge to contextualize and facilitate the interpretation of health indicators. The integration and exploration of large-scale population health data in a conceptually coherent manner is made possible through PopHR’s explicit semantic model of population health\(^ {12, 21}\). Health indicators are organized according to a determinants of health framework\(^ {22}\), with each indicator linked to relevant concepts about diseases or risk factors. Consensus knowledge, expertise and scientific literature were used to establish the links between concepts\(^ {24}\). The implementation of PopHR used for this study integrates data for an open cohort sampled randomly (25%) from the general population of the Census Metropolitan Area (CMA) of Montreal, Canada (approximately 1 million individuals in the cohort). The provincial government health insurer, RAMQ, provides individual-level health services utilization data (physician claims, hospitalizations, drug dispensations), while partners at the province’s public health agency, INSPQ, provide validated case definitions and algorithms for calculating the health indicators from these data. Finally, data from Statistics Canada and other sources were used to generate indicators of area-level risk factors and upstream determinants of health.

*PopHR Interface*

The PopHR interface includes six distinct views to visualize information: causal graph, bar graph, table, map, time series, and dashboard views (Figure 1). The entry page of PopHR is the causal graph view, showing a directed acyclic graph centered on a selected health concept (i.e., a health determinant or disease) along with its determinants and downstream diseases known to have causal links to the concept. The user can also visualize Spearman’s rank correlations between default indicators of each concept computed ecologically between values for geographical regions, through overlaid color and thickness representing direction and strength, respectively. Upon selecting a concept, the user chooses an indicator (e.g. prevalence, incidence) from an ‘indicator box’, a visualization option (map, bar graph, data table, or time series), as well as stratification levels for geography (various levels of health regions), age, and sex. Direct standardization by age, sex, or both can also be selected to account for different demographic distributions in regions being compared.

![Figure 1. Six views in the PopHR application: causal graph, bar graph, table, map, time series, and dashboard view (left to right, top to bottom).](image)

*Participants and testing environment*

The usability test was approved by McGill University’s Research Ethics Board. Ten participants were recruited from graduate-level programs in the department of Epidemiology, Biostatistics, and Occupational Health at McGill
University and none of the participants had used PopHR before. All participants were present at a half-hour training session held by the research team. Afterwards, participants conducted the usability test independently on their personal computers, in their own work or study environments during the period of February 8-22, 2017.

Task development and testing procedures

In comparison to clinical information systems, the usability test methods we used here to evaluate a population-health system do not place an emphasis on the user environment. Although the environment in which the system is used is an important factor in assessing the usability of a system, its effect is more significant in clinical settings, such as in emergency departments or intensive care units. Therefore, we adopted a user-task-system evaluation framework for our study, omitting consideration of context. Eight usability tasks were developed and subsequently reviewed by the research team domain expert (DLB) to fit predefined use cases of the application. Tasks were variants of queries in PopHR accomplished with the purpose of identifying population health patterns. Successful completion of the eight tasks required the use of all the features in the application except those in the dashboard view.

In the training session, a member of the research team described the usability test procedure and introduced the PopHR application to the participants through a 3-minute video tutorial. After the session, each participant received an Excel file containing the application link, the test procedure, questionnaires, and usability tasks. After completing a baseline questionnaire about demographic characteristics and experience with population health information systems, participants performed eight tasks sequentially. For every task, participants read the task, clicked on a “Start” button in the file, switched to PopHR to find the answer, and then returned to the spreadsheet to write the answer in the corresponding cells. After every task, participants answered three questions from the after-scenario questionnaire (ASQ) and participants were strongly encouraged to provide feedback on the usability issues they encountered during the task and express their expectations. Following the completion of all tasks, participants completed the System Usability Scale (SUS) questionnaire. Two additional questions asked participants to rate the PopHR application in comparison to other population health information systems they had experience with and to rate the usefulness of the application. All participants were advised to interact with the application for several minutes before starting the first task and to take breaks only between tasks.

Data collection and analysis

Participants’ performance on tasks, usability scores and feedback were examined to evaluate the usability of the PopHR application. Performance on task was measured in terms of task completion rate, task success rate and time-on-task. The time-on-task was determined as the time elapsed between a participant clicking on the “Start” button and writing down the answer. The time was tracked using the NOW function in Microsoft Excel, which was hidden from the participants. ASQ and SUS scores were calculated to provide task-specific and overall usability assessment of the application, respectively. We conducted thematic analysis with the qualitative data collected from participants’ written feedback. The themes used to categorize the usability issues were modified from previous studies. Additionally, Google Analytics was used to track and analyze the site traffic attributable to the participants during the test period.

Results

Participants characteristics

In total, ten students were recruited for this study. Of the ten participants, seven were under 25 years old and three were between 25 to 29. Seven participants had 1-2 years of working experience in the public health field, while others had less than one year of experience. All participants made use of computers for at least 3 hours per day for their professional use, and rated themselves as skilled Internet users. In terms of experience in accessing chronic disease indicators, only one participant reported that he often searched for indicators, while others reported they did so sometimes (5/10) or rarely (4/10). Statistics websites and research papers were the main sources from which participants accessed indicators.

Task performance

The usability performance captured in this test will be treated as the baseline measurements of the PopHR system and will be used in comparison with the usability results of subsequent versions. The summary of the usability test is shown below (Table 1). The overall completion rate was 96.25%, with only 3 tasks (task 3, 4, 5) not completed by all participants. A system error and failure to find a view allowing multiple stratifications were the reasons for the uncompleted tasks. The overall success rate, defined as the percentage of tasks that participants completed correctly, was 87.5% ranging from 70% to 100% for each participant. The average time was 3.28 minutes for successful tasks and 3.11 minutes for all tasks, ranging from 1.61 to 4.72 minutes. Considering only completed tasks gives a cleaner
measure of efficiency, while using the time for all tasks could better reflect the overall participants’ experience in a real-world situation. There was generally little variation in the time-on-task across participants except for one participant who spent on average of 7.16 minutes compared to the average of 2.5 minutes for all other participants.

The score from the ASQ after each task showed that 77.5% (mean agreement rating =2.54) of the participants found it easy to complete the task, 83.75% (mean agreement rating =2.29) were satisfied with the time to complete the task, and 86.25% (mean agreement rating =2.32) agreed there was adequate supporting information within the application. Across tasks, the ASQ scores ranged from 2.12 to 3.85.

**Table 1.** Results of the usability test by task.

<table>
<thead>
<tr>
<th>Task</th>
<th>Success rate</th>
<th>Time on successful task (mins)</th>
<th>Time on all task (mins)</th>
<th>ASQ*</th>
</tr>
</thead>
<tbody>
<tr>
<td>1- Compare indicator values by region in a certain year</td>
<td>7/10</td>
<td>2.54</td>
<td>1.98</td>
<td>2.23</td>
</tr>
<tr>
<td>2- Compare the time trend of indicator values by sex</td>
<td>10/10</td>
<td>4.01</td>
<td>4.01</td>
<td>2.23</td>
</tr>
<tr>
<td>3- Compare indicators’ confidence intervals by sex</td>
<td>9/10</td>
<td>4.73</td>
<td>4.72</td>
<td>3.38</td>
</tr>
<tr>
<td>4- Compare indicator values by age and sex in a certain year</td>
<td>8/10</td>
<td>3.89</td>
<td>3.38</td>
<td>3.42</td>
</tr>
<tr>
<td>5- Find age-standardized indicator values in a certain year</td>
<td>9/10</td>
<td>1.61</td>
<td>1.61</td>
<td>2.12</td>
</tr>
<tr>
<td>6- Find a disease’s causal relationship with other diseases or health determinants</td>
<td>7/10</td>
<td>2.07</td>
<td>2.09</td>
<td>2.19</td>
</tr>
<tr>
<td>7- Find and download the time trend of a disease indicator by region and sex</td>
<td>10/10</td>
<td>3.16</td>
<td>3.16</td>
<td>2.27</td>
</tr>
<tr>
<td>8- Find and download the time trend of a health determinant indicator by region and sex</td>
<td>10/10</td>
<td>3.89</td>
<td>3.89</td>
<td>3.85</td>
</tr>
</tbody>
</table>

*Satisfaction score is the average rating across the 3 post-task questions: ease of completing the task, amount of time it took to complete the task, and the support information in the system. The satisfaction rating scale ranged from 1 (strongly agree) to 7 (strongly disagree).*

**SUS results**

The average SUS scores from the post-test questionnaire was 70.75 (SD =14.24) out of 100 points. This score is above the average score for web applications, indicating good usability29 of the PopHR system. Among all participants, four scored above 80, four between 60 to 80 and two participants reported “OK” usability around 50. Seven out of ten participants stated in the questionnaire that they would use the PopHR system frequently and eight agreed the application was useful for their work or study. Compared to the main statistics websites participants use to search chronic disease indicators, the majority of participants found PopHR performed better in terms of data visualization, analytics functionality, data quality and ease of navigation. However, only four out of ten agreed that PopHR outperformed the comparison websites regarding the amount of information provided.

**Google Analytics results**

During the usability test period, there were 1,864 page views (1,025 unique page views) and the average time on page was 46 seconds. All six views were explored for more than 200 times each except the dashboard view which was viewed 20 times in total. Although the causal graph is the entry page (372 page views), the bar graph was explored most often (607 page views) by participants. Excluding the dashboard view, the shortest average time on page was the bar graph (16 seconds), while the longest was in the causal graph (74 seconds).

**Descriptive findings from participants’ feedback**

In total, 64 comments were provided. Participants praised the causal graph as a meaningful way to organize related variables (e.g., upstream health determinants, and downstream health conditions) and provide domain knowledge.
Four participants found the application was easy to navigate and one stated the system was easy to learn. Features such as the time slider and time series were noted for their usefulness.

From participants’ feedback, usability issues were identified and a table below (Table 2) shows the themes, the frequency of the usability problems mentioned by participants, and the example of usability issues or user expectations. All usability issues were categorized into four themes: 1) function; 2) information; 3) visualization; and 4) system status. Information and visualization were further divided into other categories. Function problems were brought up most often and most issues were related to the lack of a function for filtering data, which has not been implemented in the current version of the PopHR system. This issue becomes more pronounced as a user selects more stratification categories. For instance, in task 8, all participants could find the time trend of the indicator values stratifying by geographical regions (15 categories) and sex (2 categories). However, with 30 trend lines displayed, eight out of ten participants found it hard to read and interpret the results. A filtering feature would allow users to remove information for regions that are not of interest, facilitating interpretation of the most relevant information.

Seven participants mentioned issues about labeling in the application, especially in relation to the causal graph view and the “data linked” function. Most participants were not able to easily understand the difference between the lines indicating a causal relationship (based on knowledge from the biomedical literature) and lines indicating a correlation between indicators calculated from the data within the application. More labels and tooltips explaining the graph were expected by the participants. Visualization issues including the layout, resolution, and graphics problems were stated 11 times. Four comments criticized the stability of the system, noting the webpage needed to be refreshed following a system error.

<table>
<thead>
<tr>
<th>Theme</th>
<th>Frequency</th>
<th>Example of the usability issues or user expectations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Functions</td>
<td>New functions 21</td>
<td>Too much information was displayed when a user selected multiple stratification categories and the user expected to be able to filter data; The search bar didn’t recognize acronyms for some terms and users expected a more flexible search function.</td>
</tr>
<tr>
<td>Information</td>
<td>Label 13</td>
<td>In the causal graph, the labels were unclear and the user needed more tooltips explaining how to interpret the graph and data linked function</td>
</tr>
<tr>
<td></td>
<td>User instruction 2</td>
<td>The user needs a more detailed user manual and a complete indicator list</td>
</tr>
<tr>
<td>Visualization</td>
<td>Layout 6</td>
<td>The legend sometimes would block the data points</td>
</tr>
<tr>
<td></td>
<td>Resolution 3</td>
<td>After resizing the browser window, the application failed to properly display all options and buttons</td>
</tr>
<tr>
<td></td>
<td>Graphics 2</td>
<td>The legend in the downloaded graph didn’t display the dashed line (representing males) properly</td>
</tr>
<tr>
<td>System status</td>
<td>System stability 4</td>
<td>The system was sometimes unstable and users needed to reload current page or return to the home page</td>
</tr>
</tbody>
</table>

**Discussion**

In this study, we assessed the usability of the first implemented version of a knowledge-based population health information system, the PopHR. Users achieved a completion rate of 96% for tasks meant to be representative of those in public health practice. However, the evaluation uncovered issues with how effectively and efficiently the application communicated information and evidence. Unlike previous evaluation studies on public health information systems that mainly focused on the assessment of data quality, our principal objective was assessing how well our system allows for evidence-based decision making, and thus how it communicates information and evidence. Findings indicative of effectiveness issues included a low success rate for two tasks (1 and 6). Our results showed that usability issues, such as inadequate supporting information and an unclear interface, can lead users to draw incorrect inference and may lead to poor decision making. For instance, forgetting to change the year and geographic resolution was the reason for users failing to successfully complete those two tasks, and this type of error can be easily avoided with better labeling. The efficiency with which information is conveyed should also be a critical attribute in evaluating public health information systems as this characteristic of the system is directly related to user experience and influences the productivity of public health professionals. In our study, for all tasks where users took longer than 3
minutes, the success rate was high (94%). This result indicates that even though users could find the right answers, they did not find the information immediately. When users took longer to complete a task, they tended to report a lower satisfaction level.

Among all the usability problems identified through the test, opportunities to improve the application were pinpointed and should be relatively straightforward to address, such as the information and visualization issues. The new functions characterized by the participants coincides with the research team’s design for the next version of the application. While more development effort is needed to implement a filter function and a more flexible search bar, these two features should help users to locate relevant information more efficiently and effectively.

Although the number of participants in this usability study was limited, their interactions with the application were extensive as seen in the Google Analytics results. Additionally, most usability problems were identified repeatedly by different participants indicating that the issues identified most likely represent the majority of the total usability problems. Another limitation of the study is that the participants were all graduate students. In terms of the work experience, age, and computer skills, the students are not fully representative of our target users, who are public health planners and analysts with a focus on non-communicable diseases. This limitation may result in overlooking some disease-specific or computer skill-related usability issues. However, our research team has recently completed an additional usability test with employees from provincial and city public health departments, and those results should support more generalizable conclusions. The third limitation lies in the evaluation method used in this study. The usability test was conducted without an instructor on site and time-on-task may have been tracked less accurately compared to a screen recording approach. However, the method used in this study can reflect the user experience with the system in real working circumstance and may eliminate an evaluator effect. Additionally, this method provides a much more efficient way to conduct a usability assessment and is valuable for development teams with a rapid iterative design cycle.

In this study, we assessed the baseline usability performance of the PopHR system, identified its usability issues and reported user expectations. Our results showed that the usability of the current version of PopHR is good, but can be improved through adding new functions, offering more supporting information, enabling new visualizations, and improving system stability. The usability issues identified will be addressed in subsequent releases of PopHR to ensure the effective use of the application. Meanwhile, as PopHR continues to evolve, more data from various sources and new functions (e.g. integrating public health intervention knowledge) will be added to the application. We anticipate that continuing to improve the usability of the system while adding new features will present additional challenges.

References

Contralateral Breast Cancer Event Detection
Using Nature Language Processing

Zexian Zeng1, Xiaoyu Li2, Sasa Espino3, Ankita Roy3, Kristen Kitsch3, Susan Clare3, Seema Khan3, Yuan Luo1*

1Department of Preventive Medicine, Northwestern University Feinberg School of Medicine, Chicago, IL, USA; 2Department of Social and Behavioral Sciences, Harvard T.H. Chan School of Public Health, Boston, MA, USA; 3Department of Surgery, Feinberg School of Medicine, Northwestern University, Chicago, IL, USA; *Corresponding author

Abstract

To facilitate the identification of contralateral breast cancer events for large cohort study, we proposed and implemented a new method based on features extracted from narrative text in progress notes and features from numbers of pathology reports for each side of breast cancer. Our method collects medical concepts and their combinations to detect contralateral events in progress notes. In addition, the numbers of pathology reports generated for either left or right side of breast cancer were derived as additional features. We experimented with support vector machine using the derived features to detect contralateral events. In the cross-validation and held-out tests, the area under curve score is 0.93 and 0.89 respectively. This method can be replicated due to the simplicity of feature generation.

Introduction

Contralateral breast cancer is defined as a solid tumor developed in the opposite breast after the detection of the first primary breast cancer. Woman with a first primary breast cancer has two to six folds of increased risk to develop a contralateral breast cancer compared to the normal population1. Understanding the etiology of contralateral breast cancer can not only help us understand the risks associated with breast cancer development, but also help monitor the effects of treatments2. Efforts have been devoted to study the shared risk factors between the first and second primary breast cancer, including family history3, environmental exposures4, and genetic mutations5. These studies require us to identify the group of patients with contralateral breast cancer accurately. The prevalence of Electronic Health Records (EHR) has enabled large cohort study for different clinical problems, including the contralateral breast cancer. The abundant available information in EHR makes deep phenotyping in large cohort studies more achievable. However, in most cases, identifying contralateral events are still based on manual chart review, which is time consuming and labor intensive.

Because contralateral event is a progressive event, a patient may have been associated with risks of developing such an event for an extended period along his/her life time. The amount of work to capture and maintain pathophysiologic data along the development of risk factors and to identify new events is not trivial. On the other hand, the patient’s progressive information and clinical status are well recorded in the progress notes during the course of a hospitalization or over the course of outpatient care. In addition, the progress notes are readily and prevalently available. Moreover, in most cases, every diagnostic procedure of breast cancer generates at least one pathology report. Usually, if a patient has bilateral breast cancer, the patient should have at least one pathology report generated for each side.

In this study, we proposed a new method for detecting contralateral breast cancer using the narrative text in progress notes and the numbers of pathology reports generated for each side of the breast. With such a model, users can identify the group of patients with contralateral breast cancer among a large cohort efficiently.

Related Work

Capturing contralateral breast cancer events is one of the major tasks for the tumor registries. However, many of the registries, including National Cancer Institute’s Surveillance, did not successfully capture the contralateral events6. Studies heavily relied on manual chart review, which is both time consuming and labor intensive, thus not feasible for large cohort study7. Automated methods have been proposed to extract breast contralateral and recurrence events8-10. However, these studies did not distinguish breast cancer recurrence with contralateral breast cancer events, which significantly limited further cohort studies. Strauss et al. used the morphology codes and anatomic sites to detect contralateral breast cancer events11. However, the work required that the pathology reports are well documented in standard formats, which in reality is not true and requires special care from Natural Language Processing systems to unify the cross-institutional variations in pathology reports12,13. In addition, defining rules to retrieve information from
the copious pathology reports can be labor intensive. Furthermore, if the report did not state which side of breast was examined, the rule based system will have difficulty in calling a contralateral event. Efforts have been also devoted to apply claims data for contralateral event detection\textsuperscript{14,15}. However, claims data are believed to have limited validity for inferring cancer recurrence events\textsuperscript{16}.

Motivated by the limitation from previous studies, we proposed a method to extract features from the common narrative text in progress notes, together with the numbers of pathology reports for each side of breast cancer, to detect contralateral breast cancer events. We experimented with Support Vector Machine (SVM) and quantitatively assessed the probability of a breast contralateral event.

**Study Cohort**

The Northwestern Medicine Enterprise Data Warehouse (NMEDW) is a joint initiative across the Northwestern University Feinberg School of Medicine and Northwestern Memorial HealthCare\textsuperscript{17}. The Lynn Sage database in NWEDW was searched for women who underwent breast conservation surgery for Ductal Carcinoma in Situ (DCIS) or primary invasive breast cancer. We identified 1063 women who underwent breast conservation surgery for a new diagnosis of stage $0$ to stage $3$ breast cancer. Three co-authors (SE, AR, KK) performed chart review for these patients, and identified 33 contralateral events among these 1063 women. Study procedures were approved by the hospital’s Institutional Review Board (IRB).

**Method**

We first randomly split the 1063 subjects into a training set and a held-out test set according to a $7:3$ ratio. In the training dataset, progress notes from 15 women with contralateral breast cancer were extracted and reviewed. The sentences or partial sentences indicating the occurrence of contralateral breast cancer and cancer diagnoses related events were retrieved and summarized in Table 1. These partial sentences were then tagged by MetaMap, which is a nature language processing (NLP) tool to map the biomedical text to the Unified Medical Language System (UMLS) Metathesaurus\textsuperscript{18}. The concept unique identifier (CUI) corresponding to each concept is obtained by parsing the MetaMap outputs. To reduce the noise, CUIs that are not related to breast cancer event is manually filtered and discarded, such as the CUIs of ‘with’, ‘has’, ‘seen’, and etc. were filtered. After filtering the CUIs, 42 CUIs were retained and these CUIs together represent the descriptions for contralateral breast cancer events. We refer to these 42 CUIs as a positive CUI dictionary. These 42 CUIs appear in Appendix A.

After obtaining the positive CUI dictionary, a number of pre-processing steps were performed on the progress notes. Such as removing duplicate copies, dividing the notes to sentences, and removing non-English symbols. Negation and uncertain sentences containing the words of ‘no’, ‘risk’, ‘concern’, ‘worry’, ‘unremarkable’, ‘rule out’, ‘deny’, ‘evaluation’, and their different inflections (e.g., tenses of verbs), were excluded. Following these pre-processing steps, the remaining sentences were tagged using MetaMap. Once we got the MetaMap output, the CUIs with negations were excluded. In addition, those CUIs that are not in the positive dictionary were excluded. The retained CUIs were used as features in our model. However, using single CUI as feature may not be informative enough for us to detect some of the contralateral events. For example, the sentence “Patient was first seen for right breast cancer who now has new left breast dcis.” If we look at each individual CUI, we won’t be able to conclude the contralateral event. We need the CUIs of ‘right breast cancer’ and ‘new left breast dcis’ to reach the conclusion. To this end, a complete combination of the CUIs in the same sentence would be able to help us discriminate contralateral events. Following this observation, additional features were generated by combining CUIs that were in the same sentence. In the above example, the feature combinations of (left, right), (new, left), (left, right, breast cancer) etc. were generated. Clearly, the feature of (left, right, breast cancer) offered clues for contralateral events. Using the sentence “Patient was first seen for right breast cancer who now has new left breast dcis” as example. CUIs ‘C0007124’ (Non-infiltrating Intraductal Carcinoma), ‘C0006142’ (Malignant neoplasm of breast), ‘C0444532’ (Right sided), ‘C0222601’ (Left breast), ‘C0205314’ (New) were generated. With a complete combination of these CUIs, we obtained 31 new features. For example, one of the new feature is {‘C0444532’ (Right sided); ‘C0205314’ (New); ‘C0222601’ (Left breast); ‘C0007124’ (Non-infiltrating Intraductal Carcinoma)} and we can use it to infer breast contralateral event.

Following the pre-processing of the progress notes, additional features were derived based on the number of pathology reports. For each sample, the numbers of pathology reports generated for left and right breast cancer were separately counted and used as two additional features. Intuitively, if a patient has contralateral breast cancer event, then the patient should have at least one pathology report for each side. One additional binary feature indicating whether the patient has pathology reports for both sides were derived. Ideally, every patient with contralateral event should have
pathology reports for both sides. In our experiment, checking whether the word ‘left’ or ‘right’ is contained in the report, we were able to derive such features.

These generated features were used to train a support vector machine (SVM) model for further contralateral event detection. We chose SVM because of its widely-acknowledged generalizability. To obtain a reasonable feature sample ratio and remove the redundant features, Chi-square test was applied to select features before training the model. Only top 50% features were retained for subsequent modeling.

**Table 1.** Sentences or partial sentences indicating the occurrence of contralateral events or cancer diagnoses related events

<table>
<thead>
<tr>
<th>Sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>New idc in r breast.</td>
</tr>
<tr>
<td>Newly diagnosed contralateral ilc.</td>
</tr>
<tr>
<td>Now with new primary on the right breast.</td>
</tr>
<tr>
<td>Recently with contralateral ilc.</td>
</tr>
<tr>
<td>Was first seen for right breast cancer who now has new left breast dcis.</td>
</tr>
<tr>
<td>Right breast infiltrating ductal carcinoma stage i and left breast ductal carcinoma in situ.</td>
</tr>
<tr>
<td>Presents for a new right breast cancer.</td>
</tr>
<tr>
<td>Newly diagnosed right breast carcinoma.</td>
</tr>
<tr>
<td>Right breast concerning for breast cancer.</td>
</tr>
<tr>
<td>History of bilateral breast cancer.</td>
</tr>
<tr>
<td>A second primary was diagnosed in the contralateral breast.</td>
</tr>
<tr>
<td>Bilateral breast cancer.</td>
</tr>
<tr>
<td>Bilateral breast cancer with infiltrating carcinoma of the left breast.</td>
</tr>
<tr>
<td>With bilat breast cancer.</td>
</tr>
<tr>
<td>The patient later presented with a contralateral breast cancer on the right side.</td>
</tr>
<tr>
<td>With contralateral breast cancer on the right side.</td>
</tr>
<tr>
<td>Developed a contralateral breast cancer on the right side.</td>
</tr>
<tr>
<td>With a history of bilateral dcis.</td>
</tr>
<tr>
<td>Lumpectomy with radiation on both sides.</td>
</tr>
<tr>
<td>Had contralateral bc.</td>
</tr>
<tr>
<td>Had contralateral dcis.</td>
</tr>
</tbody>
</table>

Five-fold cross-validation was applied on the training dataset to tune parameters for the model, which were then evaluated on the held-out test data. In our experiments, we trained four baseline classifiers on different feature types. Baseline 1 is the proposed model without the additional information from the number of pathology reports, referred to as combined MetaMap. Baseline 2 uses only the numbers of pathology reports, referred to as pathology report count. Baseline 3 uses only concepts in the positive dictionary without their combination, referred as Positive Dictionary without Combination. Baseline 4 uses bag of words as features, referred as Bag of Words. To generate the features for bag of words, TfidfVectorizer class in scikit-learn was used to convert the raw documents to a matrix of TF-IDF features.

An overview of the workflow employed in this study is shown in Figure 1.
Results

Among the 1063 subjects, the average numbers of pathology reports are 3.79 with 95% Confidence Interval of ± 0.19 for left side and 3.27 with 95% Confidence Interval of ± 0.19 for right side. If a subject has pathology reports for both sides, the new feature was labeled as 1. In the training data, among the 21 subjects with contralateral breast cancer, 21 (100%) have pathology reports on both sides. Among the 724 subjects without contralateral event, 259 (35.77%) subjects have pathology reports generated for both sides.

In total, 1282 features were generated in the baseline 1 of combined MetaMap, which used all information from progress notes but not from pathology reports. Three features were generated in baseline 2 by using the three features derived from the numbers of pathology reports. A total of 42 features were used in baseline 3 by using positive dictionary without combination. In baseline 4, we used bag of words and 55192 features were generated.

Table 2 shows the feature numbers and cross-validation area under curve (AUC) scores of our proposed model in comparison with the other four baselines. To account for performance variability due to different split of folds, the five-fold cross-validation was repeated 10 times and the standard deviation was obtained. It is clear that combined MetaMap outperforms the positive dictionary without combination and also the bag of words. We compared proposed model with combined MetaMap using Student’s t-test (α=0.05). The difference is statistically significant with P_value = 0.00027. We see improvements on AUC score in our proposed model compared to all other four baselines. In the cross-validation, the AUC score of our proposed model is 0.93 with standard deviation equals 0.02.

Table 3 shows the feature number and AUC scores on the model for prediction in comparison with the four other methods in held-out test. The AUC score is 0.89, which outperforms all four baseline methods by a large margin.

Using the trained model upon the training set, we obtained the coefficient for each feature. The top ranked six features appeared in Table 4.
Table 2. Cross-validation results using different methods. Standard deviation (SD) is included in the parenthesis.

<table>
<thead>
<tr>
<th>Model</th>
<th>Feature Number</th>
<th>AUC (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Combined MetaMap +Pathology Report Count</td>
<td>1285</td>
<td><strong>0.93 (0.02)</strong></td>
</tr>
<tr>
<td>Combined MetaMap</td>
<td>1282</td>
<td>0.82 (0.07)</td>
</tr>
<tr>
<td>Pathology Report Count</td>
<td>3</td>
<td>0.75 (0.07)</td>
</tr>
<tr>
<td>Positive Dictionary without Combination</td>
<td>42</td>
<td>0.46 (0.05)</td>
</tr>
<tr>
<td>Bag of Words</td>
<td>55192</td>
<td>0.66 (0.06)</td>
</tr>
</tbody>
</table>

Table 3. Held-out test results using different methods.

<table>
<thead>
<tr>
<th>Model</th>
<th>Feature Number</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Combined MetaMap +Pathology Report Count</td>
<td>1285</td>
<td>0.89</td>
</tr>
<tr>
<td>Combined MetaMap</td>
<td>1282</td>
<td>0.68</td>
</tr>
<tr>
<td>Pathology Report Count</td>
<td>3</td>
<td>0.67</td>
</tr>
<tr>
<td>Full MetaMap without Combination</td>
<td>42</td>
<td>0.30</td>
</tr>
<tr>
<td>Bag of Words</td>
<td>55192</td>
<td>0.70</td>
</tr>
</tbody>
</table>

Table 4. Top ranked features in a coefficient study. The descriptions in parenthesis right after CUIs are UMLS concept preferred names.

<table>
<thead>
<tr>
<th>Features</th>
<th>Coefficient</th>
<th>Feature descriptions</th>
</tr>
</thead>
<tbody>
<tr>
<td>{C0007097 (Carcinoma); C0449450 (Presentation)}</td>
<td>0.556</td>
<td>{Carcinoma; Presentation}</td>
</tr>
<tr>
<td>Pathology Report for Both Side Indicator</td>
<td>0.374</td>
<td>It is an indicator whether the patient has pathology reports generated for both sides</td>
</tr>
<tr>
<td>{C0205314 (New); C0222600 (Right breast)}</td>
<td>0.278</td>
<td>{New; Right breast}</td>
</tr>
<tr>
<td>{C0006141(Breast); C0007124 (Noninfiltrating Intraductal Carcinoma); C0007124 (Noninfiltrating Intraductal Carcinoma)}</td>
<td>0.256</td>
<td>{Breast; Noninfiltrating Intraductal Carcinoma; Noninfiltrating Intraductal Carcinoma}</td>
</tr>
<tr>
<td>{C0007124 (Noninfiltrating Intraductal Carcinoma); C0007124 (Noninfiltrating Intraductal Carcinoma); C1268990 (Entire breast)}</td>
<td>0.256</td>
<td>{Noninfiltrating Intraductal Carcinoma; Noninfiltrating Intraductal Carcinoma; Entire breast}</td>
</tr>
<tr>
<td>C0281267 (Bilateral breast cancer)</td>
<td>0.246</td>
<td>Bilateral breast cancer</td>
</tr>
</tbody>
</table>

Discussion

In this study of detecting contralateral breast cancer events from progress notes and counts of pathology reports, the AUC score for our proposed model is 0.93 (±0.02) in cross-validation and is 0.89 for held-out test. The model is able to retrieve contralateral breast cancer events by using the combination of narrative text in progress notes and the additional features derived from the numbers of pathology reports. The proposed model outperformed all four baseline methods, demonstrating that different features offer different levels of information. For the *combined MetaMap* feature
only, the AUC is low because some progress notes do not necessarily contain the concepts that exist in the positive CUI dictionary. On the contrary, all patients with contralateral event have pathology reports for both sides of breast cancer. In addition, many patients without contralateral breast cancer event also have pathology reports for both sides, indicating that this feature will help improve recall but may lower precision. Putting these two types of features together has the potential to address the limitations of each other and increase the chance of identifying contralateral events. It is acknowledged that we have seen some jargons in the progress notes. However, considering that one patient’s progress notes are often written by multiple clinicians, we still have a high chance to find well-formatted sentences in the progress notes. In the positive feature study, the derived variable ‘Pathology Report for Both Side Indicator’ ranked as second feature, indicating that the variable we have created is an efficient one. In addition, the top ranked features hinted us a story that if new carcinoma presents in right or left breast, or if Noninfiltrating Intraductal Carcinoma presents twice in one sentence together with breast or entire breast, the patient then have a high chance to have contralateral breast cancer. Obviously, ‘bilateral breast cancer’ is another indicator to be used to find the events. These semi-structured features may provide alternative perspectives that are useful in capturing contralateral recurrence. In the future, we plan to identify more semi-structured features to complement CUI-based features, where tensor modeling may provide useful tools for integrating different types of clinical features10.

In an error analysis, one of the patient with contralateral event was not identified because of concept positions in the power set. In the patient’s progress note, one sentence appears as: “this is a 61-year-old woman with right breast cancer newly diagnosed”. The power set derived was \{right side; breast cancer; newly diagnosed\}. However, in the model we have trained, we only have a position-sensitive feature of \{newly diagnosed; right side; breast cancer\}. The derived power set was not recognized and the event was not identified. In the future, instead of using position-sensitive power set, we plan to use graph based representation to capture the relations between medical concepts (CUIs) with more accuracy20,21. In another case, we saw: “including stage 2 l breast ca, dcis r breast” in the progress notes. However, the ‘left’ and ‘right’ are both abbreviated to ‘l’ and ‘r’. Our model is not yet complicated enough to recognize these abbreviations.

Conclusion

Using self-defined rule-based system, one can possibly identify the numbers of pathology reports for each side of breast cancer as supplementary features. We expect this study to generalize well across medical institutions. The easiness of replication can reduce the time-consuming manual effort to identify contralateral breast cancer events for cancer registries. Moreover, instead of binary classification, this model can provide the abstractors with the continuous probability score as confidence. This study can also be applied to retrieve other breast cancer events such as local recurrence, distant recurrence as long as the positive CUI dictionary is defined.

References

## Appendix A

Table A1: The CUIs identified in the positive dictionary

<table>
<thead>
<tr>
<th>CUIS</th>
<th>CUI Preferred Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>C0006041</td>
<td>Botswana</td>
</tr>
<tr>
<td>C0006141</td>
<td>Breast</td>
</tr>
<tr>
<td>C0006142</td>
<td>Malignant neoplasm of breast</td>
</tr>
<tr>
<td>C0006826</td>
<td>Malignant Neoplasms</td>
</tr>
<tr>
<td>C0007097</td>
<td>Carcinoma</td>
</tr>
<tr>
<td>C0007124</td>
<td>Noninfiltrating Intraductal Carcinoma</td>
</tr>
<tr>
<td>C0011900</td>
<td>Diagnosis</td>
</tr>
<tr>
<td>C0019665</td>
<td>Historical aspects qualifier</td>
</tr>
<tr>
<td>C0021367</td>
<td>Mammary Ductal Carcinoma</td>
</tr>
<tr>
<td>C0205090</td>
<td>Right</td>
</tr>
<tr>
<td>C0205091</td>
<td>Left</td>
</tr>
<tr>
<td>C0205225</td>
<td>Primary</td>
</tr>
<tr>
<td>C0222600</td>
<td>Right breast</td>
</tr>
<tr>
<td>C0222601</td>
<td>Left breast</td>
</tr>
<tr>
<td>C0238767</td>
<td>Bilateral</td>
</tr>
<tr>
<td>C0281267</td>
<td>bilateral breast cancer</td>
</tr>
<tr>
<td>C0439612</td>
<td>True primary (qualifier value)</td>
</tr>
<tr>
<td>C0439631</td>
<td>Primary operation</td>
</tr>
<tr>
<td>C0441988</td>
<td>Contralateral</td>
</tr>
<tr>
<td>C0443246</td>
<td>Left sided</td>
</tr>
<tr>
<td>C0205314</td>
<td>New</td>
</tr>
<tr>
<td>C0449450</td>
<td>Presentation</td>
</tr>
<tr>
<td>C0567470</td>
<td>Breast present</td>
</tr>
<tr>
<td>C0678222</td>
<td>Breast Carcinoma</td>
</tr>
<tr>
<td>C0684010</td>
<td>Rabbi</td>
</tr>
<tr>
<td>C0750546</td>
<td>Newly</td>
</tr>
<tr>
<td>C0853879</td>
<td>Invasive carcinoma of breast</td>
</tr>
<tr>
<td>C0998265</td>
<td>Cancer Genus</td>
</tr>
<tr>
<td>C1096616</td>
<td>Contralateral breast cancer</td>
</tr>
<tr>
<td>C1134719</td>
<td>Invasive Ductal Breast Carcinoma</td>
</tr>
<tr>
<td>C1268990</td>
<td>Entire breast</td>
</tr>
<tr>
<td>C1306459</td>
<td>Primary malignant neoplasm</td>
</tr>
<tr>
<td>C1366566</td>
<td>CCL27 gene</td>
</tr>
<tr>
<td>C1449563</td>
<td>Cardiomyopathy, Familial Idiopathic</td>
</tr>
<tr>
<td>C1527349</td>
<td>Ductal Breast Carcinoma</td>
</tr>
<tr>
<td>C1552822</td>
<td>Table Cell Horizontal Align - left</td>
</tr>
<tr>
<td>C1705078</td>
<td>CCL27 wt Allele</td>
</tr>
<tr>
<td>C1997028</td>
<td>History of malignant neoplasm of breast</td>
</tr>
<tr>
<td>C2603358</td>
<td>R prime</td>
</tr>
<tr>
<td>C2984916</td>
<td>Best Case Imputation Technique</td>
</tr>
<tr>
<td>C0444532</td>
<td>Right sided</td>
</tr>
<tr>
<td>C1387407</td>
<td>Personal history of primary malignant neoplasm of breast</td>
</tr>
</tbody>
</table>
Towards Supporting Patient Decision-making In Online Diabetes Communities

Jing Zhang MS, Rebecca Marmor MD, Jina Huh PhD
University of California San Diego, San Diego, CA

Abstract
As of 2014, 29.1 million people in the US have diabetes. Patients with diabetes have evolving information needs around complex lifestyle and medical decisions. As their conditions progress, patients need to sporadically make decisions by understanding alternatives and comparing options. These moments along the decision-making process present a valuable opportunity to support their information needs. An increasing number of patients visit online diabetes communities to fulfill their information needs. To understand how patients attempt to fulfill the information needs around decision-making in online communities, we reviewed 801 posts from an online diabetes community and included 79 posts for in-depth content analysis. The findings revealed motivations for posters’ inquiries related to decision-making including the changes in disease state, increased self-awareness, and conflict of information received. Medication and food were the among the most popular topics discussed as part of their decision-making inquiries. Additionally, We present insights for automatically identifying those decision-making inquiries to efficiently support information needs presented in online health communities.

Introduction
Diabetes is a chronic, typically incurable condition; patients must constantly manage and receive care throughout their lifetimes. The burden of self-management falls on the patients. People with diabetes make decisions every day that affect their health 1, such as food and exercise choices and consequently express high desires for gaining medical knowledge 2. Upon diagnosis, patients with diabetes must absorb information, change habits, and adopt new behaviors 3. As patients achieve control of their diabetes, new issues come up and they need to readjust their routines and behavior 4. Prior research has found that there is a lack of information provided to patients after their initial diagnosis 2,5–7. Patients are left with questions regarding lifestyle modifications and other required changes 6.

Online health communities (OHCs) are a critical source of information for patients with diabetes given the vast amount of anecdotal information available 8,9. Researchers are developing collective sense-making models 10, text classification methods 11,12, and visual decision-making techniques 13 to help patients make use of OHC content. The challenge here is that, as patients gather needed information and manage their diseases over time, the kinds of inquiries patients make also evolve over time.

According to the decision-making literature, the decision-making process includes establishing objectives, gathering information, developing and comparing alternatives, making the decision, implementing a plan of action, and reflecting on the decision 14,15. These inquiries manifest in OHCs as patient post questions pertaining to decision-making, specifically, as they compare alternatives. OHCs are filled with collective knowledge that has a great potential to support the evolving information needs of patients and their decision-making. Though current research is investigating ways to make use of OHC content, how to support those patients’ decision-making is understudied.

In this paper, we investigate patients’ information needs, specifically the decision-making inquiries in OHCs. We performed content analysis of the inquiries that the comparing stage of the patient decision-making process. Our study presents insights for automatically identifying these decision-making inquiries and opportunities to better support decision-making in online health communities.

Background
Decision-making challenges in diabetes management
In the US alone, in 2014, 29.1 million people were diagnosed with diabetes 16, a 12% increase from 26 million in 2010. Out of the total population with diabetes, 28% do not realize they have diabetes. Additionally, 86 million Americans have pre-diabetes—this data translates to the fact that more than 1 in 3 Americans have at least pre-diabetes 17. Diabetes management relies heavily on self-care 1 and has a substantial and increasing impact on the quality of life of Americans. The majority of people who suffer from chronic illnesses such as diabetes fail to receive appropriate or effective chronic illness management from primary care providers 18. To bridge this gap, patients with chronic illnesses need to become educated partners in their care 19.

According to Corbin and Strauss’s chronic illness trajectory framework, eight phases along the illness trajectory were proposed to explain the evolving nature of patients’ and families’ needs, with “trajectory” referring to the course of a
chronic disease in its different stages and phases. As such, chronic illness changes over time, with the person moving from one phase of illness to the next when there is a transition in the disease state. As patients transition out of stable phases along the chronic illness trajectory, they experience increased self-awareness and the seeking of empowerment, through the process of learning, making choices, and identifying changes needed.

During this transitional stage, patients will have intensified information needs and face complex decision-making challenges surrounding lifestyle and medical choices. For example, as patients with diabetes’ conditions develop, they require support through the progress from lifestyle changes to the use of medication and finally onto insulin. As they transition to each phase, they inquire and compare their options to decide on the next course of action. Consequently, during this time, it is especially important to provide informational support to aid informed decision-making, a process that involves various resources and understanding pros and cons of different options.

**Patient decision-making**

The decision-making process is a complex construct, where researchers continue to debate the level of rationality involved in the process. For instance, normative theories of decision-making (e.g., subjective expected utility theory) assumes patients consider decisions rationally and evaluate the risks and benefits of all available interventions. On the other hand, descriptive theories of decision-making (e.g., prospect theory) describes how cognitive biases can cause people to deviate from the rational ideal. The Dual-Process Theory of decision-making states people make decisions either intuitively or rationally where the latter results in less cognitive bias.

While decision-making theories highlight the various factors that influence people’s decision-making behavior, the decision-making process models share common steps leading to making a decision. These common steps include: establishing objectives, gathering information, and comparing alternatives. We are specially interested in the comparison process because it is indicative of decision-making and can be expressed explicitly through comparative terms in the natural language. Thus decision-making inquiry in our study will be characterized by comparison making. This type of inquiry provides a critical window of opportunity to support patient decision-making by informing them of options and about the risks, benefits, and consequences of the options. Decision-making inquiries help the quality of patients’ decisions to be improved. Current studies on patient decision-making, however, focused largely on supporting shared-decision-making between the patient and the clinician during the clinical encounter. Research on understanding patients with diabetes’ decision-making about self-management in everyday settings should be further investigated.

**Informational and emotional support in OHCs**

OHCs are a critical source of information for patients with chronic diseases, and extensive research has been performed on the two main types of social support OHCs offer—emotional and informational support. Researchers identified the critical role of emotional support over informational support in OHCs with regards to community commitment and member satisfaction. Biyani et al. used machine learning methods to identify the emotional and informational support in OHCs and found that influential members provide more emotional support as compared to regular members. Similarly, ‘Caretakers’ personas in OHCs who are long time posters provide emotional support to others. Introne et al. also found that the small, densely connected core members of OHC generate the majority of support for others. Vlahovic et al. looked at support matching and satisfaction, and their findings indicated receiving emotional or informational support predicted increased satisfaction of the support receivers.

Although emotional support is being highlighted as a critical role in OHCs, we know little about how patients with diabetes attempt to make decisions using information from OHCs. Mamykina et al. showed much of diabetes self-management involves the need to make quick decisions under uncertain conditions, such as making nutritional choices, deciding on the best time to exercise, and deciding whether or not to take over-the-counter medication. As a consequence, people come to make sense of these uncertain conditions through discussions with others in OHCs. And the collective sensemaking is achieved by constructing shared meaning through deep discussions, back and forth negotiation of perspectives and conflict resolution in opinions.

The importance of informational support is further highlighted as Budak and Agrawal found informational support more frequently than emotional support in educational Twitter chats. OHC users expressed less satisfaction when they sought informational support but received emotional support, and they expressed more satisfaction when they sought and received the informational support they needed. To improve support elicitation and receiving, Wang et al. presented the conceptual model of social support elicitation and provision, in which they concluded that question asking was effective in eliciting informational support. Moreover, researchers are developing ways to help patients make use of OHC content, such as text classification methods and visual decision-making techniques. However,
the challenge here is that, as patients manage their diseases over time, their information needs in OHCs are complex and change over time as well [40].

To better support patients’ information needs in OHCs, we need to understand how their needs evolve overtime. Massimi et al. 41 surveyed across multiple communities across patients’ health trajectory and concluded that the success of OHCs should not be restricted to growth and active participation and that leaving OHCs should be considered a natural part of the community’s lifecycle. One critical perspective in this finding is the importance of understanding the community’s lifecycle together with individual members’ evolving needs. Zhang, et al. found that the sentiment of posts significant increase over time 42. Wang, et al.’s 33 work further demonstrated that relationships exist between the type of inquiries and the timing of OHC participation in that receiving emotional support strengthens community commitment. On the other hand, informational support satisfied members’ short-term information needs.

Thus, prior research suggests information support benefits members’ changing and specific needs. However, we have limited knowledge of the context surround these needs. Our hypothesis is that OHCs present patient information needs regarding decision making. In this paper, we plan to investigate these needs through decision-making inquiries, and characterize those decision-making inquiries so we can better support information needs in OHCs.

**Methods**

We chose an active, publicly available online diabetes community (anonymized for privacy of the community members) as a data source to examine our research questions. The choice was based on the overall data size (are there enough activities observed in the community) and the public nature of the data. All messages posted to the community were open to public, where all posts can be read publicly and anyone can post using anonymous IDs.

We used a Java program to extract the community’s posts, which included: title, author name, body content, whether the post is a thread initiating post or a reply, and posting date of each post. The extracted posts were saved to a local database program. Since we were interested in the members’ inquiries (e.g., questions) we excluded replies. Thus, in our dataset, we had 2,359 unique posters’ 4,055 thread initiating posts posted between 2009 and 2013. The Institutional Review Board at our university’s Human Research Protections Program determined this project to be exempt.

To extract posts containing decision-making inquiries, we first employed LIWC (Linguistic Inquiry and Word Count) 43 to automatically detect sentiments that represent decision-making. LIWC is psycholinguistic tool that calculates the frequency with which words in a given text match their dictionary of sentiments. These sentiments represent linguistic and psychological constructs such as pronouns, tense, positive emotion, confidence, and fear 43. LIWC is widely validated and tested by researchers from psychology 44, social sciences 45, social computing 46,47, and medicine 48,49 in predicting sentiments for automated analysis of large text. According to our definition, decision-making inquiries are characterized by comparison making, which will be expressed through language with comparative terms. Building on this, we applied LIWC to all thread initiating posts and ranked them based on the “compare” sentiment.

Annotator 1 (First author--A1) first independently analyzed the top 50 posts from the ranked order based on the LIWC score. Out of the 50 posts, 14 posts contained what A1 considered as comparison of alternatives as identified in the decision-making literature 14,15. A1 shared the results with A2 (last author) to iteratively refine the codebook on identifying decision-making inquiries. As a result, for the posts to be included in the content analysis, the posts would have to meet the following criteria, building on the common steps leading to decision-making, which include gathering information and developing and evaluating alternatives 14,15:

1. **Relevance**: Posts have to be about diabetes care (e.g., self-care as well as care for family and friends) For example: “I tried some oatmeal this morning and it was like eating warm, soft cardboard. YUCK! Can u all make some suggestions as to how to make it more palatable??” This post is an example of patients trying out food that works for their taste. It is coded as relevant because diet and food is a main part of diabetes management.

2. **Comparison of alternatives**: The posts have to explicitly compare specific entities in the case of relative comparison (i.e. What is better? A or B?), or inquire superlative comparison (e.g. What is the best?). Alternatively, posts that express comparison in the form of similarities, differences, or choices making (A or B, A instead of B) are also included.

3. **Decision-making intent**: The posts have to reflect decision-making intent. An example of not having a decision-making intent is as follows: “What do you all think which one is harder/worse to have type 1 or 2 and why? Please provide as much detail as possible. And which type has more complications than the other or are they both pretty equal?” This post solicits member discussion comparing experiences between type 1 and type 2 diabetes but do not entail decision-making, thus this post is excluded for analysis.
After developing the codebook, A1 continued to code a total of 801 posts that started a thread, which accounts for approximately 30% of all posts with the “compare” sentiment score larger than “0” (2803 posts).

Results
Out of the 801 posts reviewed, 79 (~10%) met our criteria pertaining to decision-making for diabetes care. Note that we employed a strict, conservative approach to weeding out ambiguous posts that hinder us to establish streamlined process of automatically identifying relevant inquiries. In the excluded posts, 34 (4.70%) were irrelevant to clinical aspects of diabetes care, 13 (1.80%) did not infer further actionable strategies, and 675 (93.49%) did not contain explicit comparison of alternatives. This finding shows that the future step in our research is to further identify aspects of diabetes care, 13 (1.80%) did not infer further actionable strategies, and 675 (93.49%) did not contain OHCs that can inform clean rules for automatic detection of decision-making inquiries. We will next elaborate on our findings on the topic of decision-making, how posters express such inquiries, the triggers and motivations behind those inquiries. In describing these inquiries, we aim to better understand patients with diabetes’ information needs pertaining decision-making.

The topics of decision-making inquiries
Our content analysis informed the subject matters OHC posters needed to gather information and make decisions upon. We found that the inquiries encompassed a range of topics specific to diabetes care, from oral medication, to insulin pumps, to sweeteners. The complete list and the breakdown from the analysis is shown in Figure 1.

![Figure 1. The common topics of decision-making inquiries](image)

Medication was the most common topic of comparison. One in three inquiries talked about medication. Medication included oral medication and insulin treatment. Posters were concerned about medication efficacy, side effects, costs, and when is the best time to take the medications. Questions posters asked surrounding medication were as general as asking if one oral medication is better than another (example post #1). Questions regarding medications could also be specific, where they went into a detailed narrative of personal history and responses to a medication and then asked for advice from peers with knowledge or in similar situation (example post #2).

Example post #1: “Do you think one oral medication is better than another?” (P8)

Example post #2: “I have been prescribed using glipizide... I have been noticing that my blood sugar went up instead of it going down as my healthcare workers told me it should... Does anyone have similar experience with glipizide? Are the other good alternative medications in place of glipizide? Thanks”(P52)

Food followed medication as the second most common topic in decision-making inquiries. As an important part of diabetes weight management, posters expressed various needs to learn about their food options. Newly diagnosed patients were eager to learn what kinds of foods to avoid, as illustrated in this example: “... I have been diagnosed with diabetes. I need to begin a Diabetic Diet right away and I am unsure as to what food would be best for me to eat? What foods would be best for me to purchase from the grocery store... ”(P54) More seasoned patients tried to
optimize their self-care routine by finding out the best foods for blood sugar control “I want to know what food group will help best to keep my sugar level down.” (P31)

Following medication and food, posters inquired about tools, devices used to aid with diabetes self-care, such as logbook and test kit. They were also interested in action plans as to how to proceed when problems came up. Additionally, drinks and sweeteners (sugar/sugar substitute) have been discussed moderately since they are so embedded in one’s daily life and can have an impact on the management of their blood sugar.

How posters express decision-making
We were interested in how posters expressed decision-making (Figure 2). We first divided the inquiries to the categories that might have differing expressions: “soliciting opinions” and “soliciting experiences.” The post would be considered “soliciting opinions” if it asked for opinions, views, or advice. Opinion-soliciting posts comprised the majority of the posts we identified as decision-making (68 out of 79, 86%). These posts tended to not be geared towards a specific group of people. For example, one poster asked: “Which one is the milder oral medication for a diabetic. Glucophage or Actos? Which will be the first one to try?” (P4)

![INQUIRY TYPE](image1)

![COMPARISON TYPE](image2)

**Figure 2.** Inquiry type and comparison type

In the remaining 14% (11 out of 79) of the posts, the posters explicitly asked if others have had similar experience. These posts typically involved a narrative of a specific context of the poster and ended with phrases such as “has anyone used...” “has anyone had experience...”. These experience-soliciting posts tended to be more specific in asking questions, and the posts were targeted towards the portion of posters with similar experience. This pattern is unsurprising since, without concrete contextual information from the poster, it is hard for other posters to self-identify and match their experiences to the poster’s.

After we identified the inquiry categories in these posts, we also uncovered how the idea of comparing alternatives was expressed through the use of language, as illustrated in what we term “comparison types.” Comparison was most frequently expressed with “comparative” terms such as “better”, “more ...” or “as ... as.” An example post would be: “Which one is the milder oral medication for a diabetic. Glucophage or Actos?...” (P4) Posts that encompassed the idea of comparing and contrasting differences and similarities also fell into this category. Subjects of comparison were explicitly stated in these posts as seen in the example of Glucophage vs. Actos.

The subjects of comparison were not as explicitly stated in the “superlative” comparison type. The superlative comparison is characterized by the posts that inquire about “the best” option. An example is as follows; “I have high fasting bg, typically around 115. All other readings are normal. Which medicine would be most suitable for this?” (P53) In this post, no specific medicine was discussed. It invites posters to bring forth suggestions. Due to the lack of explicit comparison topics, this type of posts tended to provide more contextual information, such as the condition of the poster, or the goal the poster wants to achieve.

**Triggers of decision-making**
Patients with diabetes reconstruct life with illness through the process of learning, making choices, and identifying changes needed. This process was reflected through their decision-making inquiries in this OHC, triggered by various conditions, events, or changes along their illness trajectories. These triggers were important to identify because they indicated when in the illness trajectory patients needed information to support decision-making. The triggers also contextualized the motivations behind making the inquiries. We found that the changes in disease states, increased
self-awareness, contradiction in information, and forming a goal or an outcome were the triggers identified from the decision-making posts.

(1) Changes in disease states
Our analysis revealed that posters posted decision-making inquiries when changes in disease states happened or when posters anticipated changes will happen. Examples included when posters learned that they were pre-diabetic with high fasting blood sugar, when posters were recently diagnosed with diabetes and wanted to learn about their required dietary changes, or when posters anticipated beginning insulin treatment. These changes in disease states might necessitate posters to look into their treatment options, or alternatives to their current treatment plans.

(2) Increase in self-awareness
Patients experience an increase in self-awareness along with the changes in disease state 22. However, sometimes changes in the disease states do not need to be present for the self-awareness to emerge. An example of the increase in self-awareness as a trigger in our data was when a poster learned of his family history of diabetes and asked about preventative measures. The increase in self-awareness led the posters to seek information to educate themselves: “I am trying to eat right…I like one kind of bread that is 10 grams of fiber for two slices and same calories as one slice of regular bread. But some contain more whole grain and less fiber. Which is better?”(P9) Being aware that there might be best practices for managing different aspects of diabetes led posters to make decision-making inquiries. Examples included learning how to use a new device, how to manage diabetes with insulin, or even how to obtain the best information. Because few things only have one correct way, learning the ‘how’ involved comparing different methods to find which one was the ‘best’ for the poster.

(3) Contradiction in information
Patients today receive information from various channels. Information can come from physicians, diabetes educators, peer patients, or internet searches. When faced with various sources of information, contradiction is likely to occur. Posters in this OHC posed decision-making inquiries in the midst of information contradiction they faced during this health information seeking process. An example was when the posters were attempting to understand the controversial issue of fruit intake; They inquired to decide whether or not they could eat fruit, and if so, what kinds of fruits. Contradiction also happened when posters’ expectations and experiences did not match. For example, a poster who took ‘glipizide’ for his blood sugar expected his blood sugar to go down with the medication, but instead it went up. This experience led to frustration and the poster’s desire for getting the answers to why the discrepancy happened and what the best practice going forward should be.

(4) Forming specific goal or outcome
When patients do not receive adequate or useful information through conventional channels, they fill their information gap with information culled from OHCs 36. An example showed a poster becoming frustrated with the inability of physicians to prescribe pain medications that worked: “I felt it was going to be a waste of time (with the doctors)...”(P12) As a consequence, the poster turned to OHCs to actively inquire options from other members to manage pain: “Is pot better than the near worthless and addictive Vicoden I have?”(P12)

We also found that posters made decision-making inquiries when they had a specific goal in mind, such as inquiring about options with the most or most desired properties, such as accuracy and reputability, when it comes to devices, or efficacy and safety, when it comes to medications. For instance, a poster is well aware of the current medication’s side effect and kept him/herself educated on the alternatives and formulated an inquiry that demonstrated knowledge and objective: “I read about using Alpha Lipoic to treat neuropathy...I would like to use it instead of Neurontin since I don’t like Neurontin’s side effects. Can I just use the generic kind you find in Walmart, or do I need to get it at a specialty source (like GNC)?”(P38)

Discussion
Our findings filled a gap in understanding the context that gives rise to posters’ decision-making inquiries in online diabetes communities. This context includes the topics discussed for decision-making and posters’ values and preferences for their options. As Sepucha et al, pointed out, the quality of the decision-making process can be improved by helping patients feel informed about the options, as well as the risks, benefits, and consequences of the options 24. Guided by this measurement construct for patient decision aids, we can build a patient-driven, aggregated knowledge base of diabetes management. Recent research looked into developing a problem-solving knowledge base for diabetes self-management 50. We should utilize existing and on-going peer patients’ contributions on OHCs that our posters attempted to gain.
The knowledge of how OHC posters express inquiries pertaining to decision-making in OHCs provided valuable insight for automatically identifying decision-making needs. Identification of such moments has practical implications for supporting decision-making, which will be detailed in the next section.

Technical Implications
Based on our findings we suggest several technical implications, including methods for detecting decision-making inquiries and providing informational support to decision-making inquiries.

(1) Detecting decision-making inquiries
In this study, the decision-making inquiries were identified through a two-step process including ranking by filtering through LIWC features with “compare” sentiment and manual coding. Manual coding distinguished posts relevant to diabetes care, containing comparison and decision-making intent. Our findings informed the following indicators as potential predictors for decision-making inquiries: changes in disease states, increase in self-awareness, information prone to contradiction, and goal- or outcome-oriented languages. However the qualitative indicators need to be fleshed out further in order to transform them into a measurable indicator. This presents a research opportunity for NLP, social computing, and machine learning researchers. Our next step would be to develop these indicators into measurable outcomes and test the predictability on a different diabetes community and to different disease type communities.

Once the identification happens, we can funnel these OHC members to aggregated solutions and knowledge uniquely available in OHCs as discussed below.

(2) Providing informational support to decision-making
We learned that changes in disease states and in self-awareness were some of the triggers for decision-making. We can utilize this knowledge to construct an illness information roadmap along the diabetes illness trajectory. This information road map can help patients situate their conditions and information needs in context. Disease progression and the subjects relevant to a specific disease stage will be identified and included in the road map.

For instance, sugar substitute helps patients with diabetes limit their sugar intake 51 and our findings showed choosing a suitable sugar substitute is among the top five topics of inquiry. The topics can be matched to patients and provide tailored information gathering mechanism. The information on sugar or sugar substitute should then be presented in ways that the patients can easily perform their comparison process on. For instance, when patients come into OHCs trying to decide on sugar substitute, they could be presented with aggregated information on the options people have discussed, as well as their pros and cons. Since the mechanism of comparison is central in human decision-making 52, we can externalize this process by presenting options from previously generated content in an easy-to-compare format.

With a roadmap, patients have access to transparent illness progression information and can anticipate future events and topics. As a result, the uncertainty and anxiety facing patients will be reduced, patients would feel more in control and confident in their knowledge of the condition, and feel more resourceful in managing it.

One limitation of this work is that we applied strict criteria to capture the decision-making inquiries limiting to posts expressing comparison making. Future studies can broaden the scope to include broader instances of the decision-making inquiries. However, the findings contributed to establishing rules we can build on in understanding decision-making takes place in online health communities. Another limitation is that we studied one OHC. Since different OHCs will have different cultures, including patient inquiries, disease characteristics and interaction dynamics, future work could benefit from including a diverse sample of OHCs covering conditions beyond diabetes. Since different forms of social media platforms are emerging with new capabilities for patients to connect with each other and exchange questions and answers, future study should also compare and validate our findings with different platforms in which patient information exchange occurs. Furthermore, our data was limited to what the patients reported on an OHC— we cannot assess the results of their decision-making process in their offline practice. Such limitation is a common one, however, for studying people’s behavior in online communities. Future work should involve triangulating sources not only from OHCs but also their offline behaviors through sensors, surveys, and interviews.

Conclusion
In this paper, we discussed the findings from a diabetes online community’s posts to help us understand patient information needs surrounding everyday decision-making. The knowledge generated through this study shed light on patients’ experience and challenges in the self-management of diabetes. Understanding the implications of the findings will guide us in providing better support for patient decision-making. Our study contributes to understanding decision-making inquiries patients make in OHCs and inform how clinicians, administrators, healthcare institutions, and researchers should approach supporting patients’ everyday decision-making process.
References
1. Association of the british pharmaceutical industry. The Diabetes Information Jigsaw Seek knowledge for informed choice and better health. 2006.


22. Emery C. Women living with chronic illness experienced transition that involved stages of distress and a quest forordinariness. Evid Based Nurs [Internet]. 2003;6(2):63–63. Available from: http://ebn.bmj.com/cgi/doi/10.1136/ebn.6.2.63


Can SNOMED CT Changes Be Used as a Surrogate Standard for Evaluating the Performance of Its Auditing Methods?

Guo-Qiang Zhang, PhD1,2, Yan Huang,1,2 Licong Cui, PhD2
1 Institute for Biomedical Informatics, University of Kentucky, Lexington, KY
2 Department of Computer Science, University of Kentucky, Lexington, KY

Abstract. We introduce RGT, Retrospective Ground-Truthing, as a surrogate reference standard for evaluating the performance of automated Ontology Quality Assurance (OQA) methods. The key idea of RGT is to use cumulative SNOMED CT changes derived from its regular longitudinal distributions by the official SNOMED CT editorial board as a partial, surrogate reference standard. The contributions of this paper are twofold: (1) to construct an RGT reference set for SNOMED CT relational changes; and (2) to perform a comparative evaluation of the performances of lattice, non-lattice, and randomized relational error detection methods using the standard precision, recall, and geometric measures. An RGT relational-change reference set of 32, 241 IS-A changes were constructed from 5 U.S. editions of SNOMED CT from September 2014 to September 2016, with reversals and changes due to deletion or addition of new concepts excluded. 68, 849 independent non-lattice fragments, 118, 587 independent lattice fragments, and 446, 603 relations were extracted from the SNOMED CT March 2014 distribution. Comparative performance analysis of smaller (less than 15) lattice vs. non-lattice fragments was also given to approach the more realistic setting in which such methods may be applied. Among the 32, 241 IS-A changes, independent non-lattice fragments covered 52.8% changes with 26.4% precision with a G-score of 0.373. Even though this G-score is significantly lower in comparison to those in information retrieval, it breaks new ground in that such evaluations have never performed before in the highly discovery-oriented setting of OQA.

Introduction

Large, comprehensive terminological systems such as SNOMED CT [1] continue to evolve over time, with ontology quality assurance (OQA) an indispensable part of the terminology lifecycle [2]. OQA approaches typically involve the implementation of computational tools that translate ontological principles into specific rules and patterns [3]. Ontological systems are then audited using such tools to infer and flag out substructures violating such rules and patterns, pointing to potential errors to be corrected in the next release after validation and review.

Finding errors in existing ontologies is a creative discovery process. Because of the highly discovery-oriented nature of OQA, the performance measure of precision (i.e., the percentage of true errors among the candidates that have been examined) for auditing methods is neither an initial consideration nor as glorious to quantify, unlike the setting for information retrieval [4]. Similarly, recall, the percentage of errors discovered among all true errors (“the ground truth”), is impossible to measure because of the lack of ground truth: a complete, validated error list is impossible to construct, again because of the highly discovery-oriented nature of the task.

Nevertheless, reference standards, or benchmarking data sets with validated results, have played critical roles for advancing disciplines such as image analysis (e.g., Face Recognition Technology [5] and Wilt Dataset [6]) and information retrieval (e.g., The Text Retrieval Conference series (http://trec.nist.gov)), and would be an important resource for OQA. Despite the fact that it is impossible to obtain precision and recall measures for OQA methods in absolute terms, it may still be meaningful to investigate such measures relative to some “partial ground truth.”

The goals of this paper are twofold: to develop reference sets for evaluating the performance of OQA methods for SNOMED CT, and to demonstrate how such reference sets may be applied to evaluate the performance of lattice vs. non-lattice-based methods with randomized review as a background benchmark. We propose RGT, Retrospective Ground-Truthing, as a surrogate reference standard for evaluating the performance of automated OQA methods. The key idea of RGT is to leverage the cumulative SNOMED CT changes derived from its regular longitudinal distributions by the official SNOMED CT editorial board as a partial, surrogate reference standard. Three performance measures are proposed: RGT recall, RGT precision and RGT geometric mean (G-measure), formulated by adapting the standard measures using RGT relational changes derived from SNOMED CT U.S. distributions as the reference set.

We demonstrate the feasibility of this approach by constructing an RGT reference set for SNOMED CT relational changes and performing a comparative evaluation of the performances of lattice, non-lattice, and randomized relational error detection methods using the proposed RGT measures. A RGT relational-change reference of 32, 241 IS-A changes were constructed from 5 versions of SNOMED CT from September 2014 to September 2016, with reversals and changes due to deletion or addition of new concepts excluded. 68, 849 independent non-lattice fragments, 118, 587 independent lattice fragments, and 446, 603 relations were extracted from the U.S. version 20140301. Com-
parative performance analysis of smaller (less than 15) lattice vs. non-lattice fragments were also given to approach the more realistic setting in which such methods may be applied. Among the 32,241 IS-A changes, independent non-lattice fragments covered 52.8% changes with 26.4% precision with a G-score of 0.373. Our results show that the non-lattice auditing method had significantly better overall performance for detecting incorrect and missing IS-A relations. With independent further work in the accurate identification of specific relational errors contained in individual fragments [7], non-lattice auditing method could prove to be a powerful, versatile approach to OQA.

1 Background

The Evolution of SNOMED CT. SNOMED CT [8] is the most comprehensive, multilingual clinical healthcare terminology in the world, developed by SNOMED International, the trading name of the International Health Terminology Standard Development Organization (IHTSDO). It provides a consistent representation of clinical content in electronic health records and has been used in more than fifty countries. It contains over 300,000 active concepts with unique numeric identifiers (SNOMED ID, or ID in short), organized into 19 hierarchies including Clinical Finding and Body Structure. To connect such concepts, over 1,500,000 relations, including IS-A (subtype) relationships and attribute relationships (e.g. associated morphology and finding site) [9], are constructed and maintained. Release Format 2 (RF2) [10] is used to support reference set creation.

For each version of SNOMED CT, RF2 provides a snapshot folder containing such items as: all active content (concepts, description, and relationships); a “delta” folder which includes identified new and changed content from previous snapshot version; and a “full” folder with the history of all content. Each relationship consists of numeric attributes, such as effectiveTime (date of change), active (“0” for not active, “1” for active), sourceId (the source concept of a relation), destinationId (the target concept of a relation), relationshipGroup, and typeId (relationship type).

Ochs et al. [11] captured editing operations and created an easily understandable SNOMED CT “delta” set that included added relations and removed relations. An inactive relation in the “delta” with type IS-A can be treated as an IS-A relation deletion or an IS-A relation error in the prior version. Similarly, an active relation in the “delta” with type IS-A can be treated as an IS-A relation insertion or a missing IS-A relation in the prior version.

To illustrate, Fig. 1 displays a fragment in the SNOMED CT version 20140301 (left), which was changed in version 20160901 (right) with an IS-A deletion (dashed red line) and an IS-A addition (solid red line).

![Figure 1: An example of SNOMED CT IS-A changes. A fragment in the SNOMED CT 20140301 release (left) was changed in the 20160901 release (right), with an IS-A deletion (dotted red) and an IS-A addition (solid red).](image)

Non-lattice Auditing. We review several related notions for non-lattice auditing (see more details in [12, 13]). The concepts and the IS-A relationship between concepts in an ontology such as SNOMED CT can be viewed as a partially ordered set (poset) \( L \) with a reflexive, transitive relationship \( \leq \). That is, \( L \) is the set of concepts, and \( \leq \) is the IS-A relationship. A concept (or element) \( u \) is called an upper bound of a subset \( X \subseteq L \) if for each \( x \in X \) we have \( x \leq u \). A concept \( m \) is called a minimal upper bound of a subset \( X \subseteq L \), if \( m \) is an upper bound of \( X \), and for any \( n \leq m \) such that \( x \leq n \) for each \( x \in X \), we have \( m = n \). We use \( \text{mub}(X) \) to denote the set of minimal upper bounds of \( X \).

Dually, an element \( l \in L \) is called a lower bound of a subset \( X \subseteq L \), if for each \( x \in X \) we have \( x \geq l \). An element \( m \) is called a maximal lower bound of a subset \( X \subseteq L \), if \( m \) is a lower bound of \( X \), and for any \( n \geq m \) such that \( x \geq n \) for each \( x \in X \), we have \( m = n \). We use \( \text{mlb}(X) \) to denote the set of maximal lower bounds of \( X \).

A lattice pair is a pair of concepts that have a unique maximal lower bound (when the pair has any shared lower bound). For example, on the left of Fig. 2 (an induced subgraph of SNOMED CT, in the sense that all corresponding relations
in SNOMED CT must be included in the subgraph when the nodes are included), the two concepts (in green) “Brain stem infarction” and “Cerebellar infarction” form a lattice pair: this pair has a unique minimal upper bound “Posterior cerebral circulation infarction” (in pink) and unique maximal lower bound “Claude’s syndrome” (in crimson). Thus this concept pair is a lattice pair. This fragment is called a lattice fragment.

A non-lattice pair \((a, b)\) is defined as a pair of concepts that have at least two maximal lower bounds, that is, the size of \(\text{mlb}\{a, b\}\) is greater than 1. For example, on the right of Fig. 2 (another induced subgraph of SNOMED CT), the concept pair “Acute nervous system disorder” and “Cerebral infarction” has two (hence not unique) maximal lower bounds: ‘Vascular dementia of acute onset’ and “Acute lacunar infarction.” Thus this concept pair is a non-lattice pair. This fragment (including the maximal lower bounds) is called a non-lattice fragment.

The Lattice-based Structural Auditing (LaSA) principle (a.k.a. non-lattice auditing [14]) provides a mathematically grounded, error-agnostic method for auditing biomedical ontologies. LaSA focuses on the order structure induced by the hierarchical relationship (IS-A) and requires that such a structure forms a lattice: every concept pair should not have more than one maximal lower bound. It extracts non-lattice pairs and generates non-lattice fragments, consisting of concepts in-between a maximal shared descendant and a member of the non-lattice pair. Non-lattice fragments are in conflict with the Fundamental Theorem of Formal Concept Analysis [15], which states that concept hierarchies derived from the duality of intension and extension always have their order structure being a (complete) lattice. In recent work, Cui et al. [7] provide a new method for mining non-lattice lexical patterns for detecting missing concepts and hierarchical relations in SNOMED CT. Evaluation using a random selected 100 fragments by experts, showed that non-lattice fragments have a high frequency of containing missing IS-A relationships.

2 Methods

We first construct a partial reference set focusing on two types of relational errors derived from the “delta” of SNOMED CT releases: incorrect IS-A relations represented by IS-A deletions, and missing IS-A relations as captured by IS-A insertions. We then perform a comparative evaluation of lattice, non-lattice, and randomized relational auditing methods using the standard precision, recall, and geometric measures.

2.1 Constructing RGT

We use 5 versions (U.S. version 20140301 - U.S. version 20160901) of SNOMED CT for capturing relational changes (change numbers are shown in the second column in Table 1). We focus on erroneous and missing relations on shared concepts, so newly added relations that involve newly added concepts are ignored. An IS-A relational change (or IS-A change in short) with the same source and target concepts may have different “modualId,” which may cause a repeated count. Moreover, IS-A relations may be reversed back to a prior version as a part of the changes in later versions [16]. For example, relation “Streptococcal tonsillitis (ID 41582007) IS-A Tonsillitis (ID 90176007)” is deleted in the July 2014 version and added in the January 2016 version. To construct a robust reference set of relational changes, we perform three preprocessing steps:

- Extracting only IS-A changes;
- Removing duplicated counts and reversed changes; and
2.2 Extracting Independent Lattice and Non-lattice Fragments

We first extract all lattice and non-lattice fragments for the U.S. version 20140301 of SNOMED CT using our MapReduce pipeline [12], which consists of two MapReduce phases to extract non-lattice fragments from large partially ordered ontological structures. The resulting fragments are not “independent” in the sense that one fragment may be contained in another, making further error detection intertwined, as well as violating the independent sampling assumption for statistical analysis. However, given a large collection of fragments, obtaining a reduced “independent” collection with exhaustive pairwise comparison is computationally prohibitive. To address this issue, we formulate the notion of independence as follows. For (induced and connected) fragments \( f_1 = (C_1, R_1) \) and \( f_2 = (C_2, R_2) \) where \( C_1 \) and \( C_2 \) are sets of concepts and \( R_1 \) and \( R_2 \) are sets of relations,

- we say that \( f_2 \) is a subfragment of \( f_1 \) if \( R_2 \subseteq R_1 \); and
- we say that \( f_1 \) and \( f_2 \) are independent \( f \) neither of them is a subfragment of the other.

For example, the non-lattice fragment in Fig. 3 shows the non-lattice fragment generated by the non-lattice pair “Pelvic injury” (ID 282771003) and “Closed fracture of neck of femur” (ID 35982003) has a non-lattice subfragment generated by the non-lattice pair “Pelvic injury” (ID 282771003) and “Closed fracture of neck of femur” (ID 35982003).

In general, a collection of fragments is called independent if each pair of fragments from the collection is independent.

Formally, for a non-lattice pair \((a, b)\) and their maximal lower bounds \(mlb\{a, b\}\), the non-lattice fragment determined by the pair \((a, b)\) is defined as a subgraph containing the concepts between the pair \((a, b)\) and any concept in \(mlb\{a, b\}\). For a lattice pair \((c, d)\), their only maximal lower bound is the unique element in \(mlb\{c, d\}\), and their only maximal upper bound is the unique element in \(mub\{c, d\}\), the lattice fragment determined by the pair \((c, d)\) is defined as a subgraph containing the concepts between the pair \((c, d)\) and the concept \(mub\{c, d\}\), combining the concepts between the pair \((a, b)\) and the concept \(mub\{a, b\}\). The MapReduce pipeline in [12] can be used to exhaustively detect non-lattice and lattice pairs. We further compute independent fragments using the generating pairs.

By definition, comparing all relations between every pair of fragments is required for constructing a dependent collection of non-lattice fragments. To reduce the computational cost involved, we propose an algorithm to detect all possible non-lattice subfragments for each given non-lattice fragment. Lattice fragment dependency can be computed in a similar way.

Let \(P\) be the set of all non-lattice pairs, and \(F\) be the set of all non-lattice fragments for a given SNOMED CT version. Every non-lattice fragment \(f_{(a,b)} \in F\) is generated by some non-lattice pair \((a, b) \in P\). Then \(f_{(x,y)}\) is subfragment of \(f_{(a,b)}\) if

- \((x, y) \neq (a, b), (x, y) \in P, \{x, y\} \subseteq f_{(a,b)}\); and
- for any relation \((u, v) \in f_{(x,y)}, we have \((u, v) \in f_{(a,b)}\).

The second condition is required because a non-lattice pair may generate a non-lattice fragment without itself being a part of the fragment. For instance, the non-lattice fragment in Fig. 4 shows an independent non-lattice fragment generated by the non-lattice pair “Nonvenomous insect bite with infection” (ID 10461000) and “Infected insect bite of upper limb” (ID 283347003). This fragment contains another non-lattice pair “Nonvenomous insect bite of trunk with infection” (ID 19108007) and “Infected insect bite of upper limb” (ID 283347003), although this pair does not generate a subfragment of the displayed fragment.

2.3 Performance Measures

To the best of our knowledge, there are no existing measures for comparing distinct OQA methods. We introduce RGT recall, RGT precision, and RGT geometric mean to measure the performance of OQA methods, motivated by the precision and recall measures commonly used in information retrieval.
We consider an OQA method $M$ as a group of fragments (in general terms as induced subgraphs). Each fragment may potentially capture some ontological errors that involve concepts as nodes and IS-A relations as edges. A fragment $f$ is a graph consisting of a set of IS-A relations. The size of $f$ is defined as the number of concepts involved in it.

We can view a benchmark of validated changes (i.e., ground truth) as a tuple $E = (E_0, E_1)$, where $E_0$ is the set of validated relational errors (where a relation deletion is indicated), and $E_1$ is the set of validated missing relations (where a relation insertion is indicated).

To evaluate the performance of $M$ against $E$, we can break the RGT measure of precision and recall into two categories: measures with respect to deletion, and measures with respect to insertion:

- The deletion recall is defined as the ratio
  $$\frac{|\{r \in E_0 \mid \exists f \in M, r \in f\}|}{|E_0|};$$

- The deletion precision is defined as the ratio
  $$\frac{|\{f \in M \mid \exists r \in f, r \in E_0\}|}{|M|};$$

- The insertion recall is defined as the ratio
  $$\frac{|\{r \in E_1 \mid \exists f \in M, r \in f\}|}{|E_1|};$$

- The insertion precision is defined as the ratio
  $$\frac{|\{f \in M \mid \exists r \in f, r \in E_1\}|}{|M|}.$$

Note that while RGT recall is defined as a ratio of the number of relations, RGT precision is defined as a ratio of the numbers of fragments. One can combine precision and recall to obtain the geometric mean measure (G-measure) $\sqrt{\text{recall} \times \text{precision}}$. The G-measure shows a combined performance of precision and recall. The higher the G-measure, the greater the agreement between an OQA method and the SNOMED CT changes.
3 Results

Relational changes in SNOMED CT U.S. versions ranged from 19,753 in the 20160901 release to 64,676 in the 20160301 release (Table 1). All the changes were calculated based on the released “delta” set for each version. There has been a total of cumulative 263,994 relational changes from the 20140301 release to the 20160901 release, after removing relation reversals and duplicates. Therefore, the cumulative change is not a simple summation of the numbers in all prior changes.

<table>
<thead>
<tr>
<th>SNOMED CT version</th>
<th>Relational changes</th>
<th>IS-A deletion</th>
<th>IS-A insertion</th>
</tr>
</thead>
<tbody>
<tr>
<td>20140901</td>
<td>57,879</td>
<td>4,799</td>
<td>2,723</td>
</tr>
<tr>
<td>20150301</td>
<td>60,433</td>
<td>4,426</td>
<td>3,807</td>
</tr>
<tr>
<td>20150901</td>
<td>61,253</td>
<td>5,866</td>
<td>3,350</td>
</tr>
<tr>
<td>20160301</td>
<td>64,676</td>
<td>7,888</td>
<td>3,350</td>
</tr>
<tr>
<td>20160901</td>
<td>19,753</td>
<td>1,212</td>
<td>835</td>
</tr>
<tr>
<td>Cumulative</td>
<td>263,994</td>
<td>20,744</td>
<td>11,497</td>
</tr>
</tbody>
</table>

Table 1: SNOMED CT change statistics from U.S. version 20140901 to U.S. version 20160901.

We compared five auditing methods according to our formulation of each method as a collection of fragments. Non-lattice auditing consists of all non-lattice fragments and lattice auditing consists of all lattice fragments. Similarly, independent non-lattice auditing consists of all independent non-lattice fragments and independent lattice auditing consists of all independent lattice fragments. Additionally, single-edge auditing consists of all fragments made of a single relation (edge). This corresponds to randomized edge examination.

We used 20140301 release of SNOMED CT to generate non-lattice fragments, lattice fragments, and single-edge fragments. We also extracted all independent fragments from lattice and non-lattice fragments. To calculate the deletion recall using the RGT formulation, we collected the number of SNOMED CT IS-A deletions discovered through methods (numerator) and the number of 5 versions SNOMED CT IS-A deletion cumulation (denominator). The result is shown in Fig. 5 where the y axis is the numerator and the x axis is the denominator. The 5 numbers of cumulative IS-A deletion $E_0\text{ from 20140901 release to 20160901 release (4,799; 8,965; 13,844; 19,968; and 20,744) were computed in the same way as the last row in Table 1. For each of the 5 methods, the slope is the deletion recall, and the larger the slope, the better deletion recall the method has. The line of best fitting for each method is linear, which indicates that the deletion recall is stable with different numbers of SNOMED CT IS-A deletion for all methods. In other words, deletion recall for each method is almost a constant, e.g., the slope for single-edge is always 1 because
its deletion recall is 100%. Such a feature indicates that deletion recall is a robust measure for OQA method evaluation. With such a feature, we may predict that the IS-A relations deletion recall relative to the ground truth for a method may be close to its deletion recall relative to SNOMED CT IS-A deletion.

We also divide lattice and non-lattice fragments into large sized groups and small sized groups for detailed comparison. We consider larger fragments those fragments with sizes larger or equal to 15, and small fragments those fragments with sizes smaller than 15. The second column in Table 2 presents the numbers of non-lattice and lattice fragments in each group obtained in SNOMED CT U.S. version 20140301, which is the denominator in the formulation of precision. The third column is the numbers of fragment hits (enumerator), the number of changes detected by the methods using cumulative SNOMED CT changes.

<table>
<thead>
<tr>
<th>Method</th>
<th>Number of fragments</th>
<th>Number of hits</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-lattice fragments</td>
<td>595,960</td>
<td>70,182</td>
</tr>
<tr>
<td>Independent non-lattice fragments</td>
<td>68,849</td>
<td>18,183</td>
</tr>
<tr>
<td>Independent non-lattice fragments with size ≤ 15</td>
<td>62,001</td>
<td>15,070</td>
</tr>
<tr>
<td>Independent non-lattice fragments with size &gt; 15</td>
<td>6,848</td>
<td>3,113</td>
</tr>
<tr>
<td>Lattice fragments</td>
<td>835,015</td>
<td>82,801</td>
</tr>
<tr>
<td>Independent lattice fragments</td>
<td>118,587</td>
<td>20,552</td>
</tr>
<tr>
<td>Independent lattice fragments with size ≤ 15</td>
<td>113,248</td>
<td>18,641</td>
</tr>
<tr>
<td>Independent lattice fragments with size &gt; 15</td>
<td>5,339</td>
<td>1,911</td>
</tr>
</tbody>
</table>

Table 2: List of non-lattice and lattice fragments number and hits obtained in SNOMED CT version March 2014.

Fig. 6 displays the results for deletion recall and deletion precision. Fig. 7 displays the results for insertion recall and insertion precision. Table 3 displays the results for the G-measure for IS-A deletion, IS-A insertion, and the changes with respect to each method.

The first three rows of Table 3 show that non-lattice auditing performs better than lattice auditing, which is much better than single-edge using G-measure. From Fig. 6 and Fig. 7, we can find the reason that non-lattice has a higher G-measure is non-lattice has higher deletion recall and insertion recall (79.9% and 53.4%) than lattice while they perform similarly in deletion precision and insertion precision, this means non-lattice auditing has better coverage on incorrect IS-A relational errors and missing IS-A relational errors.

Among all methods, independent non-lattice auditing performed the best on discovering incorrect IS-A relations with a 68.4% deletion recall and a 26.2% deletion precision. Only 68,849 independent non-lattice fragments (about 11% of all non-lattice fragments) exist in SNOMED CT version 20140301, and 62,001 (90%) of them are small sizes (≤15) so that they are amendable for human inspection. Our result indicates that independent non-lattice fragments could play significant role for detecting incorrect IS-A relations for SNOMED CT. However, independent non-lattice fragments failed to perform well on discovering missing IS-A relations, while non-lattice auditing returns the best G-measure in
Figure 7: Insertion recall (left) and insertion precision (right) for each method. The methods are non-lattice (NL), lattice (L), single-edge (SE), independent non-lattice (INL), independent lattice (IL), independent non-lattice with fragments size $\leq 15$ (INLS), independent non-lattice with fragments size $>15$ (INLL), independent lattice with fragments size $\leq 15$ (ILS), independent lattice with fragments size $>15$ (ILL).

missing IS-A relation detection. We noticed that a missing relations is often needed to connect a concept inside an independent non-lattice fragment and another concept outside the independent non-lattice fragment. By this reason, larger non-lattice fragments with sub-fragments are able to catch such missing relationships. The scenario can also lead to undesirable insertion discovering result for independent non-lattice because missing relationships are only detected between concepts inside a fragment.

<table>
<thead>
<tr>
<th>Method</th>
<th>Deletion</th>
<th>Insertion</th>
<th>Both</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-lattice</td>
<td>0.291</td>
<td>0.160</td>
<td>0.288</td>
</tr>
<tr>
<td>Lattice</td>
<td>0.215</td>
<td>0.120</td>
<td>0.215</td>
</tr>
<tr>
<td>single-edge</td>
<td>0.216</td>
<td>0</td>
<td>0.173</td>
</tr>
<tr>
<td>Independent non-lattice</td>
<td>0.423</td>
<td>0.132</td>
<td>0.373</td>
</tr>
<tr>
<td>Independent lattice</td>
<td>0.255</td>
<td>0.089</td>
<td>0.226</td>
</tr>
<tr>
<td>Independent non-lattice small (size $\leq 15$)</td>
<td>0.387</td>
<td>0.079</td>
<td>0.298</td>
</tr>
<tr>
<td>Independent non-lattice large (size $&gt;15$)</td>
<td>0.348</td>
<td>0.107</td>
<td>0.273</td>
</tr>
<tr>
<td>Independent lattice small (size $\leq 15$)</td>
<td>0.244</td>
<td>0.062</td>
<td>0.190</td>
</tr>
<tr>
<td>Independent lattice large (size $&gt;15$)</td>
<td>0.076</td>
<td>0.026</td>
<td>0.061</td>
</tr>
</tbody>
</table>

Table 3: List of G-measures for various methods that detect SNOMED CT changes.

4 Discussion

Small G-scores. Compared with traditional information retrieval methods, the G-scores calculated in this paper are relatively low, even for the higher-performing non-lattice approaches. This should not be a concern for two reasons. One is that SNOMED CT change is continuing with each new release, and we are always dealing with parital ground truth. In lack of the existence of available complete ground truth set, the performances for any OQA methods would suffer. The second reason is that detecting ontology errors or defects is a highly discovery-oriented, and sometimes not even well-defined process. For the first time, this paper introduced a framework to provide the feasibility to calculate G-scores to enable the comparison of distinct OQA methods (but the comparative evaluation of some existing OQA methods may remain infeasible).

Comparison of methods between fragment sizes. By comparing methods with larger fragments sizes and smaller fragments sizes, methods with larger fragments always show better RGT precision than the methods with smaller fragments. It is intuitive that larger fragments have more concepts and relationships so that they have more possibility to contain errors. On the contrary, methods with smaller size fragments present better scores on RGT recall relative to deletion and insertion, which means SNOMED CT IS-A relationship changes are aggregated on small sizes of fragments.

Comparison of populations between methods. Independent non-lattice auditing performed best among the three evaluated methods, with merely 68,849 fragments. To demonstrate the statistical significance, we computed test
statistic Z-scores on precision between non-lattice and lattice (35.5); independent non-lattice and independent lattice (46.8); independent non-lattice and single-edge (201.2); and independent lattice and single-edge (149.3). Independent non-lattice auditing departs the most in outcome from larger population sets consisting of individual IS-A relations.

**Balance of RGT precision and recall.** Our setup for ontology auditing methods to capture quality issues is analogous to capturing fishes using fishing nets. A single method provides a collection of fishing nets. The fishes are the cumulative changes approved by the SNOMED CT editorial panel that have been officially reflected the releases. The nets are the SNOMED CT subsets or fragments determined by a particular method. To illustrate the possibility in the extremes, a method may provide a single net, which is the entire graph of SNOMED CT (100% recall, 100% precision against RGT reference set). But this is useless since no progress is made in making “the haystack” smaller. Another method may provide the finest possible net, which consists of single IS-A relations. This is also useless since it is equivalent to random examination, because of the largest possible numbers of “fishing nets” used. A more useful method should balance the number of “nets” as well as the sizes of the “nets.” We believe this is achieved using independent non-lattice fragments with sizes not exceeding 15.

**SNOMED CT changes vs. the ground truth of relational errors.** The entirety of the ground truth of ontology errors is unknowable because the lack of a complete and validated error list is the inherent nature of the task for OQA. SNOMED CT changes is the most trustworthy error list since it has been generated by the world’s most authoritative group - the SNOMED International Editorial Panel. However, there exist change reversals [16] that need to be accounted for, as is the case in this study.

**Limitations of non-lattice auditing.** Non-lattice auditing [14] is founded on the theory of formal concept analysis. This paper demonstrates a significant difference between the G-scores of lattice vs non-lattice based methods in their ability in capturing official SNOMED CT relational changes. However, knowing a non-lattice fragment contains a relational error and identifying this specific error remains two different matters, as each fragment still contains multiple relations. To address this issue, we developed data mining techniques [7] leveraging structural and lexical information for automated detection of relational errors in non-lattice fragments. Further development of data mining techniques combining a rich variety of information sources represent provide additional research opportunities in this area.

**Changes in the number of non-lattice fragments between versions.** The number of non-lattice fragments can be a significant measure to evaluate the quality of SNOMED CT. We computed the numbers of non-lattice fragments for 4 pairs of versions using shared common concepts between versions: 581,327 (version 20140301) and 575,927 (version 20140901), 562,819 (version 20140901) and 574,028 (version 20150301), 596,015 (version 20150301) and 587,263 (version 20150901), 610,785 (version 20150901) and 604,221 (version 20160301). We did not include the latest version in the comparison because of the sudden drop in the number of relational changes. 3 of 4 pairs showed a decrease of the number of non-lattice fragments. This may suggest that quality assurance work between these versions actually reduced the number of non-lattice fragments. However, the reasons for the unexpected increase in the 20150301 release may need further investigation.

**SNOMED CT versions used for RGT.** We used the latest five versions of SNOMED CT (the March 2017 version was released too late to be included in this study) for feasibility demonstration of our method. This choice was purely a matter of convenience and did not result from methodological or computational limitations of our approach. In fact, it would be desirable to use all the available SNOMED CT versions (in Release Format 1 as well as Release Format 2), with an appropriate starting release as the time-point for computing the fragments. One might only want to go so far to reach a point of relatively stable version (with the sizes of delta sets within a reasonable range). All the performance measures would only improve, as the RGT reference set will become larger. We also noted an abnormality in the sudden drop of relational changes in the 20160901 release, which needs further investigation.

**5 Conclusion**

This paper introduced an innovative notion of RGT reference set for SNOMED CT relational changes and performed a comparative evaluation of the performances of lattice, non-lattice, and randomized relational error detection methods using the newly introduced precision, recall, and geometric measures. An RGT relational-change reference set of 32,241 IS-A changes were constructed from 5 versions of SNOMED CT from September 2014 to September 2016, with reversals and changes due to deletion or addition of new concepts excluded. 68,849 independent non-lattice fragments, 118,587 independent lattice fragments were extracted from the SNOMED CT March 2014 distribution. Comparative performance analysis of smaller (less than 15) lattice vs. non-lattice fragments were also given to reflect the more realistic setting in which such methods may be applied. Among the 32,241 IS-A changes, independent non-lattice fragments covered 52.8% changes with 26.4% precision and a G-score of 0.373, showing non-lattice auditing...
as a superior approach than lattice auditing, confirming a theoretical predication implicitly given in [7, 14].

Acknowledgment

We thank Shiqiang Tao and Wei Zhu for their contribution to the development of tools used in this paper for non-lattice visualization. This research was supported in part by University of Kentucky Center for Clinical and Translational Science (Clinical and Translational Science Award UL1TR001998) and National Science Foundation under MRI award No.1626364.

References

Barriers, Facilitators, and Solutions to Optimal Patient Portal and Personal Health Record Use: A Systematic Review of the Literature

Jane Y. Zhao, MD¹, Buer Song, MD, PhD¹, Edwin Anand, MD¹, Diane Schwartz, MLS¹, Mandip Panesar, MD¹, Gretchen P. Jackson, MD, PhD², Peter L. Elkin, MD¹

¹Department of Biomedical Informatics, Jacobs School of Medicine and Biomedical Sciences, University at Buffalo, State University of New York, Buffalo, New York, USA; ²Departments of Surgery, Pediatrics, and Biomedical Informatics, Vanderbilt University Medical Center, Nashville, Tennessee, USA

Abstract

Patient portal and personal health record adoption and usage rates have been suboptimal. A systematic review of the literature was performed to capture all published studies that specifically addressed barriers, facilitators, and solutions to optimal patient portal and personal health record enrollment and use. Consistent themes emerged from the review. Patient attitudes were critical as either barrier or facilitator. Institutional buy-in, information technology support, and aggressive tailored marketing were important facilitators. Interface redesign was a popular solution. Quantitative studies identified many barriers to optimal patient portal and personal health record enrollment and use, and qualitative and mixed methods research revealed thoughtful explanations for why they existed. Our study demonstrated the value of qualitative and mixed research methodologies in understanding the adoption of consumer health technologies. Results from the systematic review should be used to guide the design and implementation of future patient portals and personal health records, and ultimately, close the digital divide.

Introduction

The successful treatment of many chronic diseases, like diabetes mellitus, has long been facilitated by strong patient investiture in his or her own healthcare. Empowered and well-informed patients who have access to their personal health data—such as their HbA1c trend over time—may be persuaded to make sound behavioral modifications that correlate with their disease progression. Patient portals and personal health records (PHRs) typically provide patients with access to their laboratory and imaging results, alternative modes of communication with their providers, and many other benefits.

Despite the availability of patient portals and PHRs for nearly a decade, the adoption rate and use of both consumer health technologies among patients have been low. Researchers are becoming increasingly attentive to the barriers that prevent certain members of the population from using patient portals and PHRs. Factors related to socioeconomic status, race, age, or health condition should not influence the penetrance of patient portals and PHRs into society to the extent that they currently do. These technologies need to be embraced across the population to be optimally effective, otherwise disparities in the quality of and access to care will increase.

Given the growing body of literature on the topics of patient portal and PHR adoption and use, in addition to regulatory pressures on healthcare organizations to increase patient enrollment, the need has arisen for a critical analysis of the current literature to best understand how to pursue subsequent research and policy endeavors in a meaningful way. In this study, we aim to systematically review the literature to identify publications that specifically address barriers, facilitators, and solutions to successful enrollment and use of patient portals and PHRs.

Methods

Overview & Scope

Patient portals are secure online tools that can stand alone or be tethered to a healthcare organization’s health record, through which patients can access their personal health information from anywhere with an internet connection. PHRs are online applications that are owned and managed by patients or their proxies and allow for patient input of information for greater control of patients’ own health information management. While the two consumer health technologies serve different functions for patients in terms of ownership, the factors that hinder or facilitate their enrollment and use are fairly equivalent. Moreover, patient portals and PHRs at times are used interchangeably in
the literature. Therefore, we conducted our systematic review to address barriers and facilitators of adoption and use of both patient portals and PHRs, rather than try to determine whether authors meant one versus the other based on terminology and context. In this study, any reference to patient portals is meant to be synonymous with a reference to PHRs, unless explicitly indicated.

Our review included studies that identified barriers, facilitators, and/or solutions. The reason for delving into all three rather than focus on one was to gain an in-depth understanding of what exactly has been studied in the literature and what else remains to be studied. For the scope of this review, the following traditional definitions were used:

- **Barrier**: a law, rule, problem, etc., that makes something difficult or impossible. 5
- **Facilitator**: one that helps to bring about an outcome by providing indirect or unobtrusive assistance, guidance, or supervision. 6
- **Solution**: an answer to a problem. 7

For the purposes of this study, the term “solution” included proposed or implemented methods to address enrollment or usability barriers, even if further study was still required. As such, “solutions” should be distinguished from “facilitators,” which have been studied and proven via evidence to help bring about improvements in patient portal enrollment and use.

Selections for analysis were made from quantitative, qualitative, or mixed methods studies published in scientific journals, peer-reviewed conference proceedings, and reputable sources identified by domain experts. Purely descriptive studies of patient portals or of individual characteristics (e.g., age, gender, race) were excluded from analysis.

**Literature Search**

The search strategy, including the Medical Subject Heading (MeSH) terms and keywords, was developed and performed under the guidance of a health science librarian (DS). The search was performed using MEDLINE® via Ovid as the database. This particular database was chosen for the level of granularity authors could employ in the search strategy compared to other databases. The search query used was [“Patient access to records” (MeSH) OR “patient participation” (MeSH) OR “health records, personal” (MeSH) OR “patient portal” (keyword) OR “web portal” (keyword) OR “patient health record” (keyword)] AND [“health services accessibility” (MeSH) OR “health knowledge, attitudes, practice” (MeSH) OR “patient acceptance of health care” (MeSH) OR “user-computer interface” (MeSH) OR “implementation barrier” (keyword) OR “usability” (keyword) OR “human factors” (keyword) OR “barriers to access” (keyword) OR “facilitators to access” (keyword) OR “provider access” (keyword) OR “caregiver access” (keyword)]. All searches were limited to human studies and published in English between the years 2000 and 2017 with full text availability. Reference mining was performed on key articles and from content experts. Prior systematic reviews were used to identify original studies.

**Screening & Data Extraction**

Three authors (JZ, EA, & BS) screened seven hundred twenty titles and abstracts for eligibility. Selected full-text articles were read in their entirety to determine final inclusion or exclusion (Figure 1). Initial screening and data extraction were performed together to facilitate inter-rater reliability. Subsequent screening and data extraction occurred independently. The workload was distributed evenly among authors at each screening iteration. Any discrepancies that occurred were discussed and resolved through consensus-building.

**Analysis**

At least two authors reviewed each article. Authors independently took notes on barriers, facilitators, and solutions mentioned in the selected full-text articles and recorded their findings in a shared data shell. Recurrent themes were identified and used to code the extracted data (Table 1).

**Results**

Seven hundred-twenty citations from MEDLINE® via Ovid were identified by the literature search. Forty-one were identified as potentially eligible after screening through titles and abstracts. Fifteen articles met eligibility after full-text review. Seventeen additional articles that met eligibility criteria but were not found via the initial literature
search were identified by reference mining. Thirty-two articles were ultimately included in the review (Table 2). Of the included articles, eight were quantitative, nineteen were qualitative, and five were mixed methods studies. Several of the captured themes were represented across studies (Figure 2).

**Barriers**

Thirteen studies identified attitude or culture as a barrier to effective enrollment and use of patient portals. Patients either felt overtly negative toward the patient portal (e.g., self-tracking was explicitly perceived as extra work, distrust of the healthcare system as a whole was a cultural norm among certain populations and patient portals were viewed as unwelcome extensions of healthcare) or were satisfied with the status quo (e.g., concern that embracing patient portals would erode the personal patient-physician relationship).

Six articles acknowledged mixed or negative attitudes among providers as barriers. Providers felt conflicted over whose responsibility it was to promote the patient portal and saw the patient portal as extra work added to an already long list of burdensome clinical responsibilities.

Thirteen studies mentioned interface challenges. Problems ranged from unintuitive design elements that created navigation difficulties to the use of text or language at a reading comprehension level too high for most users. In fact, most non-users had not completed high school, which exacerbated their ability to fully comprehend the complexities of their various comorbidities. Nine studies specifically made a point to underscore that no amount of patient portal training could help patients who were functionally illiterate at baseline. Many non-users also had poor digital literacy or lacked computer or internet access. Among these were patients who felt uncomfortable using public resources to access private health information online, e.g., at a public library.

In six of the studies, patients either had not been fully educated about the extent of functionalities offered by the patient portal and felt that the features they knew about were not worth the hassle of continued use, or had not even been informed about the patient portal’s existence prior to the study. Technical or logistical difficulties with the enrollment process (e.g., difficult navigation, lack of information technology support) prevented interested patients from completing registration in fifteen studies.

Privacy concerns were captured as barriers in five studies. People did not have adequate assurance that their personal data were secure on the internet. Others desired extra security after logging into the patient portal or PHR to hide the display of sensitive information. This finding was notable among senior patients who often required the assistance of family or friends to help them navigate the patient portal or PHR or to use it on their behalf via proxy access.

Cost concerns were addressed in two of the studies. Certain users chose to not explore different patient portal features out of concern they might be charged. These same users expressed willingness to pay for extra features in the portal (e.g., after-hours secure messaging with their physician) so long as prices were made transparent.

**Facilitators**

Ten studies identified successful registration as a facilitator to continued patient portal use. Institutions with support staff—i.e., informational technology personnel, nurses, and even doctors—who dedicated time to walk patients through the patient portal and show them the various capabilities on one or more trial runs were successful at committing patients to enroll. Patients who had family members help them enroll were also successful. Twelve studies described the benefits of training. Many patients were interested but felt uncomfortable navigating the patient portal. Training engendered familiarity with the interface and enabled ongoing use of the patient portal.

Patients and providers were likely to use the patient portal if they perceived the patient portal to be more beneficial than existing options (e.g., secure messaging versus calling). This was largely dependent on prior use of the patient portal as well as provider buy-in. To convince patients to enroll, providers had to promote the patient portal in the legitimate context of a clinical visit and provide assurance that the patient portal would supplement, not replace, the existing patient-physician relationship. Providers were particularly effective advocates of the patient portal or personal health record among senior and minority patients. Five studies described how
Figure 1. Flow diagram of article selection and elimination.

Table 1. Two examples of how similarities between extracted data about barriers, facilitators, and solutions (notes made by JZ, EA, BS) were codified by theme (rightmost column).

<table>
<thead>
<tr>
<th>Citation</th>
<th>Barriers</th>
<th>Facilitators</th>
<th>Solutions</th>
<th>Themes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Emani et al. 2012</td>
<td>(B4, B8) Baseline lower personal innovativeness in information technology and less access and use of other technologies; (B5) Lack of interoperability; (B7) Lower education levels and socioeconomic status</td>
<td>(F1) Multiple opportunities to observe and try out personal health record; (F3) Demonstrate perception of relative advantage over existing practices; (F3) Older age has a small but perceived impact on the perceived value of the personal head record; (F4) Providers who play an active role in the uptake of personal health records and their subsequent use; (F5) Easy to learn and use</td>
<td>(S2) Increase communication channels to teach about a personal health record</td>
<td>(B4) Lack of technology access or support (B5) Interface problems (B7) Poor health literacy (B8) Poor digital literacy (F1) Training (F3) Perception of benefit (F4) Provider buy-in (F5) Interface design improvement (S2) Provide training and support</td>
</tr>
<tr>
<td>Mishuris et al. 2014</td>
<td>(B1) Satisfaction with status quo communication with providers; (B3) Limited prior knowledge of patient portal; (B4) Limited computer and internet access</td>
<td>(F2) Surrogates acting as intermediaries between Veterans and their medical care; (F3) Desire to learn more about patient portal</td>
<td>(S1) Enroll Veterans at their point of care, rather than when they physically visit the VA; (S3) Provide written and verbal information about portal features and enrollment procedures; (S4) Provide surrogates with a unique access code to enable them to view authorized portions of health record (to protect Veteran’s privacy over other portions) and complete transactions on the Veteran’s behalf</td>
<td>(B1) Negative attitudes by patient (B3) Lack of awareness (B4) Lack of technology (F2) Institutional or family support (F3) Perception of benefit (S1) Policy changes (S3) Better marketing and outreach (S4) Better interface design</td>
</tr>
<tr>
<td>Author</td>
<td>Study Name</td>
<td>Year</td>
<td>Study Type</td>
<td>Total Subjects</td>
</tr>
<tr>
<td>--------</td>
<td>------------</td>
<td>------</td>
<td>------------</td>
<td>----------------</td>
</tr>
<tr>
<td>Ancker et al.</td>
<td>Use of an electronic patient portal among disadvantaged populations</td>
<td>2011</td>
<td>Quantitative</td>
<td>Patients n=11903</td>
</tr>
<tr>
<td>Ancker et al.</td>
<td>“You get reminded you’re a sick person”: personal data tracking and patients with multiple chronic conditions</td>
<td>2015</td>
<td>Qualitative</td>
<td>Patients n=22; health care providers n=7</td>
</tr>
<tr>
<td>Black et al.</td>
<td>True “meaningful use”: technology meets both patient and provider needs</td>
<td>2015</td>
<td>Qualitative</td>
<td>Patients n=31; providers n=13</td>
</tr>
<tr>
<td>Day et al.</td>
<td>Influencing factors for adopting personal health record (PHR)</td>
<td>2012</td>
<td>Qualitative</td>
<td>Patients n=10</td>
</tr>
<tr>
<td>Emani et al.</td>
<td>Patient perceptions of a personal health record: a test of the diffusion of innovation model</td>
<td>2012</td>
<td>Mixed Methods</td>
<td>Patients n=760</td>
</tr>
<tr>
<td>Ennis et al.</td>
<td>Collaborative development of an electronic personal health record for people with severe and enduring mental health problems</td>
<td>2014</td>
<td>Quantitative</td>
<td>Needs analysis n=133; preliminary testing n=13; preliminary implementation n=26</td>
</tr>
<tr>
<td>Fix et al.</td>
<td>Encouraging patient portal use in the patient-centered medical home: three stakeholder perspectives</td>
<td>2016</td>
<td>Qualitative</td>
<td>Focus groups n=4; providers n=3; program leaders n=10</td>
</tr>
<tr>
<td>Goel, Brown, Williams, Cooper, et al.</td>
<td>Patient reported barriers to enrolling in a patient portal</td>
<td>2011</td>
<td>Qualitative</td>
<td>Patients n=159</td>
</tr>
<tr>
<td>Gordon et al.</td>
<td>Differences in access to and preferences for using patient portals and other eHealth technologies based on race, ethnicity, and age: a database and survey study of seniors in a large health plan</td>
<td>2016</td>
<td>Qualitative</td>
<td>Administrative database n=231082; survey respondents n=2602</td>
</tr>
<tr>
<td>Haggstrom et al.</td>
<td>Lessons learned from usability testing of the VA’s personal health record</td>
<td>2011</td>
<td>Mixed Methods</td>
<td>Patients n=24</td>
</tr>
<tr>
<td>Haun et al.</td>
<td>Evaluating user experiences of the secure messaging tool on the Veterans Affairs’ patient portal system</td>
<td>2014</td>
<td>Qualitative</td>
<td>Veterans n=33</td>
</tr>
<tr>
<td>Hilton et al.</td>
<td>A cross-sectional study of barriers to personal health record use among patients attending a safety-net clinic</td>
<td>2012</td>
<td>Quantitative</td>
<td>Patients n=338</td>
</tr>
<tr>
<td>Lober et al.</td>
<td>Barriers to the use of a personal health record by an elderly population</td>
<td>2006</td>
<td>Qualitative</td>
<td>Publicly subsidized housing project residents n=38</td>
</tr>
<tr>
<td>Luque et al.</td>
<td>Barriers and facilitators of online patient portals to personal health records among persons living with HIV: formative research</td>
<td>2013</td>
<td>Qualitative</td>
<td>Persons living with Human Immunodeficiency Virus n=90</td>
</tr>
<tr>
<td>Lyles et al.</td>
<td>“I want to keep the personal relationship with my doctor”: understanding barriers to portal use among African Americans and Latinos</td>
<td>2016</td>
<td>Qualitative</td>
<td>Focus groups n=10; total participants n=87</td>
</tr>
<tr>
<td>McInnes et al.</td>
<td>Development and evaluation of an internet and personal health record training program for low-income patients with HIV or hepatitis C</td>
<td>2013</td>
<td>Mixed Methods</td>
<td>Low-income veterans with HIV or hepatitis C n=14</td>
</tr>
<tr>
<td>Mishuris et al.</td>
<td>Barriers to patient portal access among veterans receiving home-based primary care: a qualitative study</td>
<td>2014</td>
<td>Qualitative</td>
<td>Veterans n=14; surrogates n=2; staff n=3</td>
</tr>
</tbody>
</table>

Table 2. List of eligible studies included in systematic review.
Table 2. Continued.

<table>
<thead>
<tr>
<th>Reference</th>
<th>Title</th>
<th>Year</th>
<th>Study Design</th>
<th>Sample Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nijland et al.²⁵</td>
<td>Factors influencing the use of a Web-based application for supporting the self-care of patients with type 2 diabetes: a longitudinal study</td>
<td>2011</td>
<td>Mixed Methods</td>
<td>Enrolled patients n=50; nonenrolled patients n=300</td>
</tr>
<tr>
<td>Roblin et al.²⁶</td>
<td>Disparities in use of a personal health record in a managed care organization</td>
<td>2009</td>
<td>Mixed Methods</td>
<td>Patients n=1777</td>
</tr>
<tr>
<td>Ronda et al.²⁷</td>
<td>Reasons and barriers for using a patient portal: survey among patients with diabetes mellitus</td>
<td>2014</td>
<td>Qualitative</td>
<td>Patients n=1390</td>
</tr>
<tr>
<td>Sarkar et al.²⁸</td>
<td>The literacy divide: health literacy and the use of an internet-based patient portal in an integrated health system—results from the diabetes study of northern California (DISTANCE)</td>
<td>2010</td>
<td>Qualitative</td>
<td>Patients n=14102</td>
</tr>
<tr>
<td>Sarkar et al.²⁹</td>
<td>Social disparities in internet patient portal use in diabetes: evidence that the digital divide extends beyond access</td>
<td>2011</td>
<td>Quantitative</td>
<td>Patients n=14102</td>
</tr>
<tr>
<td>Schickedanz et al.³⁰</td>
<td>Access, interest, and attitudes toward electronic communication for health care among patients in the medical safety net</td>
<td>2013</td>
<td>Qualitative</td>
<td>Patients n=416</td>
</tr>
<tr>
<td>Sox et al.³¹</td>
<td>Patient-centered design of an information management module for a personally controlled health record</td>
<td>2010</td>
<td>Quantitative</td>
<td>Needs assessment n=15; initial usability testing n=10; performance-testing n=7</td>
</tr>
<tr>
<td>Tieu et al.³²</td>
<td>Barriers and facilitators to online portal use among patients and caregivers in a safety net health care system: a qualitative study</td>
<td>2015</td>
<td>Qualitative</td>
<td>Patients n=11; caregivers n=5</td>
</tr>
<tr>
<td>Tieu et al.³³</td>
<td>Online patient websites for electronic health record access among vulnerable populations: portals to nowhere?</td>
<td>2016</td>
<td>Qualitative</td>
<td>Patients n=23; caregivers n=2</td>
</tr>
<tr>
<td>van der Vaart et al.³⁴</td>
<td>Measuring actual eHealth literacy among patients with rheumatic diseases: a qualitative analysis of problems encountered using Health 1.0 and Health 2.0 applications</td>
<td>2013</td>
<td>Qualitative</td>
<td>Health 1.0 test patients n=15; Health 2.0 test patients n=6</td>
</tr>
<tr>
<td>Ward³⁵</td>
<td>A cautionary tale of technology: not a substitute for careful collaboration and effective communication</td>
<td>2012</td>
<td>Qualitative</td>
<td>Patient n=1</td>
</tr>
<tr>
<td>Wen et al.³⁶</td>
<td>Consumers' perceptions about and use of the internet for personal health records and health information exchange: analysis of the 2007 Health Information National Trends Survey</td>
<td>2010</td>
<td>Quantitative</td>
<td>US civilian noninstitutionalized adults n=7674</td>
</tr>
<tr>
<td>Yamin et al.³⁷</td>
<td>The digital divide in adoption and use of a personal health record</td>
<td>2011</td>
<td>Quantitative</td>
<td>Patients n=75056</td>
</tr>
<tr>
<td>Zarcadoolas et al.³⁸</td>
<td>Consumers' perceptions of patient-accessible electronic medical records</td>
<td>2013</td>
<td>Qualitative</td>
<td>New York City residents n=28</td>
</tr>
</tbody>
</table>

Patients with chronic illnesses often relied on the patient portal to request prescription refills or to prepare for their next clinic visit.⁴,¹⁵,²⁵-²⁶,³⁷

Six studies identified iterative feedback loops as critical facilitators. Patient portals that went through multiple rounds of usability testing and relied on user feedback prior to modifications often found greater success in patient adoption and use.¹¹,¹³,¹⁶,¹⁸,³¹,³⁸ Being able to make timely evidence-based improvements was also key to steadily increasing user enrollment and maintaining continued use by patients.

**Solutions**
Eight studies promoted policy changes at the public and institutional level to stimulate increased computer and internet access, and to emphasize the impact of the digital divide as a public health concern. 4,13,24,28,30,34-35,37 Eleven studies proposed training and support for patients on digital basics and on the patient portal. In two studies, authors advocated for the continuation of traditional health information resources in print form.14,24 Authors from thirteen studies advocated for more aggressive tailored marketing strategies toward disadvantaged populations. The key was to indiscriminately highlight all advantages of the patient portal, and not solely emphasize its potential as a self-management tool. Self-management was often viewed more as burden than benefit among disadvantaged, chronically ill patients.4,13-16,18-19,24,26-29,32 Transparency about patient portal capabilities, security features, and costs of services was also encouraged.

Redesigns in patient portal interfaces were mentioned in seventeen studies.4,8-10,12,15-18,24-25,28,30,32,34-35,38 Authors advocated for standardization in the creation of future patient portals. Easy-to-use, easy-to-navigate interfaces and simpler language were all encouraged to attract patients across socioeconomic groups. Six studies indicated the need for more consumer health informatics research.10,12,26,29,33-34

Discussion

Patient portals and PHRs were conceived as revolutionary and transformative patient engagement tools, but the reality has been disappointing given low enrollment and usage equitably across populations. Efforts have been made in recent years by researchers to understand the factors that contribute to the success or failure of consumer health technology adoption. Our systematic review evaluated all publications since 2000 that specifically addressed barriers, facilitators, and solutions to optimal patient portal and PHR enrollment and use. Popular themes were identified.

Negative attitudes by patients posed one of the most common barriers. Negative attitudes ranged from outright refusal to engage with the technology, to apathy or ambivalence. Without the provision of a clinically relevant context to the data displayed, patient-accessible laboratory work or imaging results came across to patients as meaningless, confusing, and in the worst-case scenario, even detrimental. On the other hand, patients’ perception of the patient portals’ value increased after the patient portals’ capabilities were showcased and demonstrated, particularly when done so by their providers. For instance, patients who were older or had multiple comorbid conditions represented users who saw great value in having an online resource to keep track of their many medications and provider recommendations. Moving forward, healthcare organizations will need to make greater efforts to recruit providers in the promotion of consumer health technologies without these efforts being seen as yet another clinical documentation burden.

Toward that end, marketing and outreach need to occur indiscriminately, while being sensitive to the likelihood that users from different socioeconomic statuses and cultures may have different expectations of the patient portal. These solutions reflect the bare minimum requests by researchers to prevent the digital divide from widening. The onus falls on informatics societies, such as the American Medical Informatics Association, to not only raise the bar for consumer health technologies by establishing and promoting gold standard guidelines for the design and implementation of patient portals and PHRs, but to also hold vendors and healthcare organizations accountable.

Having a governing body oversee standardization of patient portals and PHRs is more important than ever, especially since our systematic review has revealed that the other most common barrier to patient portal and PHR registration and use is the lack of a user-friendly interface. While providing staffing and resources to help patients interact with the patient portal interface has proven beneficial, particularly in alleviating stress or anxiety associated with patient portal enrollment and navigation for “digital immigrants (individuals who do not or cannot access computers and the internet on a regular basis, e.g. the poor and the elderly),” the reality remains unchanged that many interfaces are long overdue for a facelift.

Future Directions

Consumer health informatics is a growing discipline. Our review of the literature has provided patient portal and PHR vendors and the institutions that purchase them a succinct summary of all available evidence-based strategies to recruit and retain patient portal and PHR users. Ongoing research, particularly in the domain of human factors and usability testing, would be beneficial for the continued optimization of consumer health technology adoption. 39-40
Figure 2. Themes captured from reviewed articles.
Light grey columns represent barriers. Light green columns represent facilitators. Light blue columns represent solutions. A shaded cell indicates that the barrier/facilitator/solution was present in a study. Theme counts are totaled at the bottom of each column.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Negative attitudes by patient</td>
<td>13</td>
<td>13</td>
<td>11</td>
<td>1903</td>
<td>1903</td>
<td>1903</td>
<td>1903</td>
<td>1903</td>
<td>1903</td>
<td>1903</td>
<td>1903</td>
<td>1903</td>
<td>1903</td>
<td>1903</td>
<td>1903</td>
<td>1903</td>
<td>1903</td>
<td>1903</td>
<td>1903</td>
<td>1903</td>
<td>1903</td>
<td>1903</td>
<td>1903</td>
<td>1903</td>
<td>1903</td>
<td>1903</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Conclusion

We conducted a systematic review and identified several themes in barriers and facilitators of patient portal and PHR adoption and use. Common barriers were negative attitudes by patients and interface problems. Frequently reported facilitators were perceptions of benefit and the provision of patient portal and PHR training. Interface redesign and improved marketing and outreach were popular solutions. More research is needed in human factors and usability testing to understand how to engage and motivate patients toward improved health outcomes. Sufficient penetration of patient portals and PHRs into patients’ routine healthcare practices will provide researchers and providers the opportunity to harness the intellect, energies, and care of patients to achieve mutual healthcare goals.

References


Identifying Falls Risk Screenings Not Documented with Administrative Codes Using Natural Language Processing

Vivienne J Zhu, MD, MS,1 Tina D Walker, BSN,RN,2 Robert W Warren, MD, MS, MPH2 Peggy B Jenny, BSN, RN,2 Stephane Meystre, MD, PhD,1 Leslie A Lenert, MD, MS1

1Biomedical Informatics Center and 2 Information Solutions at Medical University of South Carolina, Charleston, South Carolina

Abstract
Quality reporting that relies on coded administrative data alone may not completely and accurately depict providers' performance. To assess this concern with a test case, we developed and evaluated a natural language processing (NLP) approach to identify falls risk screenings documented in clinical notes of patients without coded falls risk screening data. Extracting information from 1,558 clinical notes (mainly progress notes) from 144 eligible patients, we generated a lexicon of 38 keywords relevant to falls risk screening. 26 terms for pre-negation, and 35 terms for post-negation. The NLP algorithm identified 62 (out of the 144) patients who falls risk screening documented only in clinical notes and not coded. Manual review confirmed 59 patients as true positives and 77 patients as true negatives. Our NLP approach scored 0.92 for precision, 0.95 for recall, and 0.93 for F-measure. These results support the concept of utilizing NLP to enhance healthcare quality reporting.

Background
Falls are a leading cause of injury in the elderly population and a significant morbidity and mortality risk factor. Approximately 35-40% of generally healthy people over 65 fall at least once each year.1 The risk of falling increases proportionately with age. At 80 years, over 50% of seniors fall annually.2 Over 20% of falls cause a serious injury such as hip fracture or brain trauma.3,4 In 2012-2013, 55% of all deaths due to unintentional injury among the elderly came from falling.5 The estimated direct medical costs for fall injuries in the U.S. alone are $31 billion annually.6 Clinical screening of fall risk can significantly prevent falls.7,8 However, less than half of the elderly patients who fall discuss the fall with their health care providers.9 In 2001, The American Geriatrics Society and British Geriatrics Society (AGS/BGS) published a clinical practice guideline on falls prevention and management in older persons. The AGS/BGS guideline specifies that, at least annually, providers should ask all elderly patients if they had a fall or had no fall during the prior year.10 Moreover, falls risk screening is a core quality measurement required by the Centers for the Medicare and Medicaid Services (CMS) payment incentive program under the Affordable Care Act (ACA).11 Primary Care Providers (PCP) can receive reimbursement for falls risk screening through voluntary participation in the Physician Quality Reporting System (PQRS).12 In 2015, the Medicare Access and CHIP Reauthorization ACT (MACRA) became law, offering value-based alternative payments for reimbursing physician services through a Merit based Incentive Payment System (MIPS).13 In performance year 2017, falls risk screening remains a high priority measure in MIPS.14 Both PQRS and MIPS require reporting the percentage of patients aged 65 years and older who had a risk screening for falls within the prior 12 months. They both use a set of CPT or HCPCS codes to define the denominator and numerator for this quality measurement. However, coded data alone may not completely or accurately reflect clinical activities that providers performed during their practice. Clinical narratives, such as visit notes, admission notes, progress notes, consultation notes, and nursing notes contain important information and are increasingly available in electronic form. However, obtaining structured information from free-text clinical narratives is a major challenge. Natural language processing (NLP) is a technology that uses computer-based linguistics and machine learning approaches to extract and classify information automatically from free-text data.15 NLP offers an opportunity to identify clinical information from narrative documents and holds the promise of enhancing data availability and quality for diagnostic, patient safety, clinical decision support, and quality performance reporting.16-19 In this study, we developed an NLP approach to identify documented falls risk screening in clinical notes of patients lacking discretely coded falls risk screening. We also formally evaluated the NLP algorithm performance against a gold standard (i.e. domain expert manual review). We hypothesized that this NLP approach could correctly identify more falls risk screening in electronic health records (EHR) than the quality metrics based on by administrative codes. Confirmation of this hypothesis would support using NLP approaches to enhance current quality reporting systems.
**Methods**

**Data resource and NLP software**

The primary study setting was the Medical University of South Carolina (MUSC). MUSC is an academic medical science center with inpatient, outpatient, and emergency facilities serving Charleston, South Carolina, and surrounding areas. MUSC has had the EpicCare EHR system (Epic Systems Corp., Verona, WI) in place for outpatient care since 2012 and for inpatient care since 2014. A Research Data Warehouse (RDW) copies the Epic data warehouse and serves as the data repository for clinical research. This study tests the feasibility of identifying mentions of falls risk screenings in clinical notes using NLP for the defined patient population. Clinical narratives collected during 2015 for MUSC Medicare outpatients are available for analyses, including progress notes and consult notes. This study was approved by the MUSC data access committee. We used commercial NLP software (Linguamatics I2E version 4.4, Cambridge, United Kingdom) licensed by MUSC to index, parse, and query each clinical note for this project. Linguamatics I2E (I2E) applies concept-based indexing techniques to electronically identify key words/phrases from text documents and map them to concepts in the UMLS Metathesaurus. Then, I2E queries retrieve information for reports meeting a user-defined set of criteria through a user-friendly interface to define syntactic and semantic representations.

**Measurement of falls risk screening**

For performance year 2015, MUSC used screening for future falls risk (CMS193v3) to report the percentage of Medicare patients 65 years of age and older who were screened for falls risk during the measurement period. The numerator is the number of patients who were screened for future falls risk at least once during the measurement period. Two CPT codes, 3288F and 1100F, are required on the claim form to submit this numerator option. The denominator is the number of patients from the initial Medicare outpatient population. Exceptions for the denominator are documented medical reasons not to screen for falls risk, including the patient being non-ambulatory or wheelchair bound. We developed an NLP algorithm and conducted a performance evaluation based on CMS193v3.

**Development of the lexicon for falls risk screening**

In practice, providers record clinical findings and activities as narratives for each patient encounter, including symptoms, health histories, physical exam results, consultation services, and care plans. In these clinical narratives, NLP can identify mentions of falls risk screening that providers performed during the encounter. A domain expert, a nurse analyst with extensive quality measure experience, manually reviewed 144 patient charts and generated an initial list of terms that were commonly used to represent falls risk screening. These initial terms included “fall,” AMPAC (Activity Measure for Post Acute Care), and ADLs (Activities of Daily Living measures) plus mention of fall. Without domain expert engagement, the NLP informatics team then developed a draft full terminology set based on this initial list. We also generated modifiers (e.g. “evaluation,” “associated symptoms,” “risk screen,” “history of,” etc.). I2E provides default detection of pre- and post-negation, a collection of regular terms that are negative mentions (e.g. “no,” deny,” “negative”). However, the I2E default clinical negation detection is not sufficient to exclude a “false” falls risk screening in clinical notes. In fact, when a patient is identified as “no fall” or “denied falls” in clinical notes, a falls risk screening had likely been conducted during the encounter. Therefore, we considered such negative mentions of falling as true cases of falls risk screening. Another area of ambiguity included words associated with “falling” that might represent lab value decreases (e.g., “hgb fall,” “bone density fall,” “ejection fraction fall”), vital sign reductions (e.g., “weight fall,” “BP fall”), seasons (e.g., “last fall,” “in the fall”), problems (e.g., “fell on ice,” “fall with fracture”), or others (e.g., “fall asleep”, “drain fall”, “fell out”). In context these terms are not relevant to a fall screening and were reconsidered as valid negations for this study. We also included in our negation criteria terms that reliably indicate a non-ambulatory patient (e.g., “wheelchair bound”, “bed bound”) because a non-ambulatory patient is excluded from the denominator for CMS falls risk screening measures. For each term, we utilized the I2E morphologic variants functions and I2E built-in ontology to generate a set of spelling variants, acronyms, and abbreviations; we then queried these terms against clinical notes to extract any relevant lexical representations to form an enhanced and refined list. The domain expert and the NLP informatics team came to consensus agreement, creating the final lexicon. A final lexicon list was imported to the I2E customized macros, which allows for reuse and refining.

**Development of NLP algorithms to identify falls risk screenings**

We developed a set of NLP queries to identify falls risk screening mentions in clinical notes (progress notes and consults notes), using the following criteria: A) mentions of falls by the lexicon; and B) excluding fall mentions with negations. These NLP queries were designed to capture semantic information, syntactic patterns, and clinical negations in order to translate a documented falls risk screening to structured data elements including: 1) patient MRN;
2) falls risk screening; 3) note create date; 4) note updated date; 5) author type; 6) note ID; and 7) type of clinical note. We used I2E 4.4 to index, query, flag, and count the number of query hits within each type of clinical note. We used all clinical notes for 144 patients to develop an NLP algorithm for each variable, evaluated the results of the NLP queries independently and in combination against the gold standard of expert chart review. These NLP and chart review evaluations were done independently by a domain expert, who was blinded to the patient source. Discrepancies between query results and the manual expert review led the informatics team to iteratively refine and develop the I2E query algorithms until sensitivity and specificity could not be further improved or could not reach predefined thresholds. A multiple I2E query combining these 7 queries was established to produce a structured output table to store information extracted by the 7 queries.

**NLP algorithm performance evaluation**

For clinical notes from all 144 patients, we compared the results generated from the I2E multiple query to the results from the gold standard. A domain expert further validated the results for both positive and negative cases of NLP-detected falls risk screening by manual chart review. The average times for falls risk screening case identification by both the I2E multiple query and manual review of clinical notes were reported and compared. Furthermore, we calculated three standard performance measures for the NLP algorithm: the precision, recall, and $F$-measure. Precision (exactness) is the proportion of true positives to the total number of algorithm-identified cases; in contrast, recall (completeness) is the proportion of true positives that are retrieved by algorithms. $F$-measure is a weighted harmonic mean of precision and recall ($F$-measure $=2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$); an $F$-measure reaches its best value at 1 and worst at 0. Finally, for all false positives and negatives generated by the NLP algorithms, the reasons for false classification were manually determined and summarized to improve the algorithm.

**Results**

**Falls risk screening lexicon**

Using the Epic Medical Shared Savings Program (MSSP) abstraction assistant tool, we identified 144 patients who were Medicare outpatient beneficiaries and had no discrete fall screening results recorded during 2015. The I2E multiple query processed 1,588 documents (1,575 progress notes and 13 consult notes) from these 144 unique patients within 0.5 seconds comparing with 10 hours by manual review. The average number of document per patient was 11 (max: 106; minimum: 2). We developed a set of I2E queries to search the initial key words ("fall," "ADLs," "AMPC," etc.) against these 1,588 documents. After iterative evaluations between the keywords hit and the original documents, we developed a lexicon of falls risk screening and negations. The final lexicon presented wide variations (Table 1). A total of 38 terms associated with falls risk screening resulted in 397 hits in 223 documents from 93 unique patients. The leading keywords were "fall" and its morphologic variants (236 hits, 67.4%), "fall risk" (42 hits, 11.7%), and "denies any falls/no falls/ Denies falling/no recent falls" (22 hits, 6.2%). Among 26 pre-negations, the leading terms were "fracture fall," "this fall," and "in the fall." Among 35 post-negations, the leading terms were "fall at," "fall in," "fall on," and "fall of [a year]". The most common note in which these instances occurred was a progress note, and the most common author types were physician, physical therapist, and physician assistant (Figure 1).

![Figure 1: Note type and author type.](image-url)
Table 1. Lexicon of falls risk screening and frequency (example)

<table>
<thead>
<tr>
<th>Terms relevant to falls risk screening</th>
<th>Pre-negation</th>
<th>Post-negation</th>
</tr>
</thead>
<tbody>
<tr>
<td>fall</td>
<td>113</td>
<td>13</td>
</tr>
<tr>
<td>falls</td>
<td>47</td>
<td>6</td>
</tr>
<tr>
<td>falling</td>
<td>29</td>
<td>6</td>
</tr>
<tr>
<td>Fall Risk</td>
<td>24</td>
<td>5</td>
</tr>
<tr>
<td>fell</td>
<td>18</td>
<td>4</td>
</tr>
<tr>
<td>fall risk</td>
<td>15</td>
<td>2</td>
</tr>
<tr>
<td>Fallen</td>
<td>13</td>
<td>2</td>
</tr>
<tr>
<td>Fall</td>
<td>13</td>
<td>2</td>
</tr>
<tr>
<td>LE AMPAC</td>
<td>12</td>
<td>2</td>
</tr>
<tr>
<td>denies any falls</td>
<td>11</td>
<td>2</td>
</tr>
<tr>
<td>No falls</td>
<td>8</td>
<td>1</td>
</tr>
<tr>
<td>AM-PAC Raw Score</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>fall</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Denies falling</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>No recent falls</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Denies any fractures or falls</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>History of falls</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Falls risk scale score</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Risks for falls</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>risk of fall</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>No risk of fall</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

NLP algorithm performance

The multiple I2E query combining seven I2E queries produced a structured output table which extracted a patient MRN, note ID, keywords for falls risk screen, the sentences where keywords hit, note type, author type, and note creation date. It also provided a link to the original document, which the NLP developer and domain expert could review during the development and evaluation phases (Figure 2). Among 144 patients, I2E query identified 93 patients with a likely mention of falls risk screening from 223 documents. Applying both pre- and post-negations, NLP identified 62 patients (40.2%) with falls risk screening from 118 documents. The domain expert’s manual review identified 64 patients who had falls risk screening and 80 patients who did not have a falls risk screening. The I2E algorithm and manual review identified, in common, 59 patients who had falls risk screenings and 77 patients who had no falls risk screening. Three non-ambulatory patients were identified by both NLP and manual review. The I2E query for falls risk screening had a precision of 0.92, recall of 0.95, and an F-measure 0.93 (Figure 3). A major reason for false negatives and false positives was that our NLP approach could not clearly distinguish a fall event as the reason of a clinical visit from a falls risk screening in some clinical notes (Table 2).

Figure 2: Example of I2E query output
**Figure 3. Results of manual review and I2E algorithm**

| Table 2. Detailed information of false positives and false negatives of falls risk screening |
|---------------------------------|---------------------------------|
| False Negatives                 | False Positives                 |
| No falls but reports some near falls at home | chest pain related to the fall and movement; but nothing ... |
| She denies any trauma or falls in the last 6 months, | the fall - she has no recollection of events |
| One week ago, she tripped and fell on her knees her | |
| Denies falls or fractures, urinary stones, heat... | |

**Discussion**

Using Lingumatics I2E, we developed an NLP lexicon and algorithms to detect mentions of falls risk screening in clinical notes (mostly progress notes) automatically for patients who had no coded falls risk screening reported to PQRS through Epic MSSP. Our NLP algorithms sufficiently extracted and aggregated necessary quality outcomes for falls risk screening and achieved high performance as evaluated with our gold standard. Our study results confirmed that NLP can enhance current quality reporting approaches for falls risk screening. The potential value of NLP enhancements for quality regulatory reporting is extremely high for measures such as annual falls risk screening. Nevertheless, we strongly caution against the use of current NLP technology and algorithms as we have utilized here for official measure reporting. However, automated NLP assessments could be used periodically throughout a quality measure reporting period to help gauge individual and group provider performance and encourage needed discrete documentation. Moreover, some regulatory quality metrics do allow manual abstraction; for example, limited and official chart review could confirm a positive NLP result for falls risk screening within clinical notes that had not been coded. That might well result in significantly better and more accurate provider group performance on the regulatory measure.

Information about falls risk screening is commonly documented in clinical notes, and completely and accurately extracting mentions of falls risk screening from clinical narratives using NLP is an achievable although challenging goal. To the best of our knowledge, this is the first report of NLP–based extraction for quality reporting on falls risk screening. Other studies have focused on falls detection, with variable accuracy. For example, Toyabe reported a high accuracy (F-measure: 0.91) NLP approach to identify inpatient injuries from image order entries; he also noted that his NLP algorithm for identifying inpatient falls in progress notes only achieved an F-measure of 0.12 because of the high number of false-positive cases.\(^{23}\) Elsewhere, Shiner et al. reported about an NLP algorithm to detect falls in progress notes with modest accuracy (specificity: 0.80; sensitivity: 0.44), as evaluated by manual review.\(^{24}\) While these studies were designed to detect a “fall incident,” our NLP algorithms were designed to detect a “falls risk screening” event from clinical notes, an NLP task posing more challenges. Initially, we used terms related to “fall risk” and “fall screening” to probe a falls risk screening; however, in clinical practice, providers may not use these terms to document falls risk screening even when they have actually performed it. For example, providers may document screening with the following statements: “Patient has no fall in the past year” or “no fall.” Providers may also document clinical information relevant to falls as “there have been some falls, memory loss,” or “fall or dizzy when standing.” Prior types of NLP algorithms would not likely detect these cases nor do typical coding procedures. Therefore, we used the term “fall” as the major search term in order to mimic a common way that providers document a conversation about falls with patients. Before applying negations, a hit of “fall” is much more frequent than hit of “fall risk” (236 vs. 42 mentions in our corpus).
We applied negation detections to classify a mention of “fall” into either a positive case or a negative case of falls risk screening. At the same time, it was important to keep in mind that the term “fall” may have other linguistic meanings that would cause false positives for “falls risk screening” identification. For example, a description of a “fall event” (“fall on wall,” “fall with broken arm”); a season of the year (“last fall,” “fall 2014”); or “fell” as a typo for “feel.” We established a set of pre-negation and post-negation detection rules in order to eliminate false positives; however, these negations also introduced false negatives. For quality reporting, we expect a highly specific (avoiding false positive) NLP algorithm by sacrificing sensitivity. After iterative evaluation and refinement, our final I2E algorithm achieved 0.92 precision and 0.95 recall, which indicated that our NLP approach could effectively identify falls risk screening in clinical notes of patients who had no coded falls risk screening recorded in their EHR.

Our NLP approach accurately identified 59 patients without coded falls risk screening (out of 144) who had a falls risk screening documented in their clinical notes. These patients were previously misclassified as having no falls risk screening performed by their PCPs. This result indicates that quality reporting based on administrative coded data alone may not accurately measure the quality of care providers performed. After years of PQRS implementation, CMS noted that potential data quality issues exist among reporting systems, such as missing data elements and inaccurate or incomplete quality measure calculations. CMS also highlighted that completeness and accuracy of data are critical to the accurate calculation of a quality performance score.23 Researchers have suggested multimodal approaches rather than relying on a single data resource to capture clinical events.26, 27 Recently, quality reporting studies, including colonoscopy quality metrics, postoperative complications identification, and advanced adenoma detection, demonstrated that the NLP-based approach offers a powerful alternative to either unreliable administrative data or labor-intensive manual chart reviews.28-30 In order to reduce non-reporting issues for measuring falls risk screening, one potential strategy is to incorporate NLP-extracted information on quality indicators that exist in the clinical notes as the supplemental data source for current quality reporting systems.

Limitations
Four major limitations are identified for this study. First, the study population (negatives by coded data) only included 144 patients, and we estimate that relatively small number of patients could be identified as cases of falls risk screening; therefore, we did not split the clinical notes into a training and testing set. In order to obtain the NLP performance for the testing set, we plan to include 2014-2016 data and randomly sample a training and testing set to reevaluate our NLP algorithm. Second, we only used clinical notes from patients who had no administrative coded data recorded for falls risk screening to develop our lexicon and NLP algorithms. Indeed, providers may document a falls risk screening differently in clinical notes for patients who had coded falls risk screening. On the other hand, providers may possibly not document a falls risk screening in clinical notes at all because these screening activities are assumed to be reported through a coded form. In fact, MUSC also uses the Morse Fall Scale to assess falls risk for hospitalized patients; and this information is usually not documented in clinical notes. Therefore, our NLP approach would not reliably identify patients who had been assessed with the Morse Fall Scale. A further study focusing on NLP strategies identifying falls risk screening for cases confirmed by coded data may help refine the lexicon and NLP algorithms. Finally, this study was designed for the context of MUSC quality reporting purposes. Our lexicon and NLP algorithms may not be generalizeable to other institutions without customization and evaluation. However, the concept of NLP-based approaches to enhance PRQS reporting may well be applicable to other healthcare settings.

Conclusions
Information about falls risk screening can commonly be found in clinical notes of patients lacking such screening recorded by coding systems for quality reporting purposes. For PQRS quality reporting, we developed an NLP approach that automatically and effectively identified and extracted information about falls risk screening found in clinical notes. The results support the concept that using both structured coded data and clinical narratives for quality reporting is superior to the current reporting approach based on administrative coded data alone. Future work will focus on replicating the study findings using a larger sample and developing NLP-based identification strategies for other quality matrix. NLP augmentation of current regulatory quality measure assessment methodologies may result in improved care and more accurate quality reporting.

Acknowledgement
This research was supported in part by an NIH/NCATS grant to the South Carolina Clinical and Translational Research Institute (SCTR) (5UL1TR00062-05). We also would like to thank Paul Sobanski and Ryan Siegle at Epic Systems Corporation.
References
27. Olsen S, Neale G, Schwab K, Psaila B, Patel T, Chapman EJ, Vincent C. Hospital staff should use more than one method to detect adverse events and potential adverse events: incident reporting, pharmacist surveillance and local real-time record review may all have a place. Qual Saf Health Care. 2007 Feb;16(1):40-4.
Spark-MCA: Large-scale, Exhaustive Formal Concept Analysis for Evaluating the Semantic Completeness of SNOMED CT

Wei Zhu, Licong Cui, PhD, Guo-Qiang Zhang, PhD
1Institute for Biomedical Informatics, University of Kentucky, Lexington, KY
2Department of Computer Science, University of Kentucky, Lexington, KY
3Department of EECS, Case Western Reserve University, Cleveland, OH

Abstract

The completeness of a medical terminology system consists of two parts: complete content coverage and complete semantics. In this paper, we focus on semantic completeness and present a scalable approach, called Spark-MCA, for evaluating the semantic completeness of SNOMED CT. We formulate the SNOMED CT contents into an FCA-based formal context, in which SNOMED CT concepts are used for extents, while their attributes are used as intents. We applied Spark-MCA to the 201403 US edition of SNOMED CT to exhaustively compute all the formal concepts and sub concept relationships in about 2 hours with 96 processors using an Amazon Web Service cluster. We found a total of 799,868 formal concepts, within which 500,583 are not contained in the 201403 release. We compared these concepts with the cumulative addition of 22,687 concepts from the 5 “delta” files from the 201403 release to the 201609 release. 3,231 matches were found between those suggested by FCA and those from cumulative concept addition by the SNOMED CT Editorial Panel. This result provides encouraging evidence that our approach could be useful for enhancing the semantic completeness of SNOMED CT.

Introduction

Ontologies (a.k.a. terminology systems) serve as a knowledge source in many biomedical applications including information extraction, information retrieval, data integration, data management, and clinical decision support. The quality of an ontology is a key property that determines its usability. As evidenced by their fast paced evolution (e.g., SNOMED International releases a new version almost every 6 months), ontological systems are often incomplete and under-specified. Therefore, one quality issue is semantic completeness. For example, in SNOMED CT, the concept Structure of muscle acting on metatarsophalangeal joint (with identifier 707861009) was not present in the 201403 release but was added in the 201603 release. To study semantic completeness, Jiang and Chute used Formal Concept Analysis (FCA) as a tool to construct contexts from normal form presentations in SNOMED CT and analyzed the resulting lattice hierarchy for unlabeled nodes. These unlabeled nodes signify semantically incomplete areas in SNOMED CT which can serve as candidate pool of new concepts for inclusion in SNOMED CT.

Given their size and complexity, performing such an FCA-based semantic completeness analysis for large ontologies is computationally as well as methodologically challenging. In fact, constructing concept lattices from formal contexts is believed to be a PSPACE-complete problem. Hence, performing formal concept analysis on large biomedical ontologies in their entirety, such as on SNOMED CT with over 300,000 concepts and 1,360,000 relations, was considered impractical. Because of the computational challenge, Jiang and Chute had to randomly select only 10% of the contexts from the subbranches of two largest domains of SNOMED CT to perform an FCA-based analysis. Although such a limited sample size may serve to demonstrate feasibility, an exhaustive analysis would be needed to achieve the full potential of the approach.

Most existing algorithms for computing formal concepts are not scalable to very large contexts. The process of generating formal concept hierarchies and constructing concept lattices from large contexts, after formal concepts are identified, is an additional computationally costly step. Due to such challenges, no scalable approaches have been proposed to exhaustively audit the semantic completeness of large biomedical ontologies using FCA, even with the aid of cloud computing technology. To address the computational challenge involved in constructing concept lattices from large contexts, we propose Spark-MCA, a Spark-based Multistage algorithm for Concept Analysis. Spark-MCA implements our proposed new FCA-based algorithms within the Apache Spark distributed cloud computing framework to provide a scalable approach for exhaustively analyzing the semantic completeness of large biomedical ontologies.

Another challenge that impedes the quality assurance of semantic completeness is the lack of reference standards.
Because of the inherent discovery-oriented nature and associated challenges involved in quality assurance work for ontological systems, there is an inevitable lack of “reference standards.” In semantic completeness auditing, case studies are commonly used for validating proof-of-concept. However, for large scale ontological systems, case studies may not be sufficient. We adopt RGT, Retrospective Ground-Truthing, as a surrogate reference standard for evaluating the performance of auditing methods for semantic completeness. The key idea of RGT is to extract the cumulative concept changes during the evolution of a developing ontology. In this study, for the purpose of evaluating semantic completeness, we focused on the addition and deletion of concepts. We used the cumulative inserted concepts in “delta” files of SNOMED CT between two releases, 201403 and 201609, as the RGT to evaluate the performance of Spark-MCA.

The contributions of this paper are: (1) the introduction of Spark-MCA, scalable distributed algorithms for exhaustively evaluating the semantic completeness within feasible computational times; (2) the application of RGT for validating the performance of FCA-based semantic completeness auditing method. Our results show that Spark-MCA provides a cloud-computing feasible approach for evaluating the semantic completeness of SNOMED CT using formal concept analysis.

1 Background

SNOMED CT Release Cycle. Developed and maintained by SNOMED International, SNOMED CT is the premier comprehensive clinical terminology for healthcare-related industries worldwide. As required by the meaningful use provisions of the Health Information Technology for Economic and Clinical Health (HITECH) Act, SNOMED CT is suited for use in clinical documentation in electronic health records (EHR). Clinical data coded in SNOMED CT can be used for multiple purposes. For example, providers can send SNOMED CT encoded data securely during transitions of care to other providers or upon discharge, or share data with patients themselves. This decreases barriers to the electronic exchange of critical patient information and positively impact patient safety.

SNOMED CT adopted description logic (DL) to formally represent concept meanings and relationships. A normal form of concept expression is a structured combination of subtype relationships (i.e. is-a) and attribute relationships (e.g., associated morphology, causative agent, finding site, part of). For example, the definition of Shunt of cerebral ventricle to extracranial site is shown in Figure 1 using such a normal form.

![Figure 1: SNOMED CT concept expression for Shunt of cerebral ventricle to extracranial site.](image)

Each SNOMED CT release comes with a “delta” file, which identifies the individual changes that occurred between releases. Table 1 is a fragment of the “delta” file in the SNOMED CT 201609 release. In the “active” column, an entry 1 denotes that a concept (e.g., 128282005) becomes active from this release (i.e., an addition), while 0 denotes that a concept (e.g., 128570000) becomes inactive from this release (i.e., a deletion). We use this file to retrieve cumulative concept changes between releases as RGT to evaluate our Spark-MCA.

<table>
<thead>
<tr>
<th>id</th>
<th>effectiveTime</th>
<th>active</th>
<th>moduleId</th>
<th>definitionStatusId</th>
</tr>
</thead>
<tbody>
<tr>
<td>128282005</td>
<td>20160731</td>
<td>1</td>
<td>900000000000207008 900000000000073002</td>
<td></td>
</tr>
<tr>
<td>128570000</td>
<td>20160731</td>
<td>0</td>
<td>900000000000207008 900000000000074008</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: A fragment of the delta file in the SNOMED CT 201609 release.
Formal Concept Analysis. Formal concept analysis (FCA) has been advocated as a mechanism to audit the quality of SNOMED CT.\textsuperscript{4,11–13} FCA is a mathematical method for knowledge engineering. It can be used to derive a concept hierarchy or formal ontology from a collection of objects and their attributes.\textsuperscript{14,15} The input data for FCA is a binary relation between a set of objects and a set of attributes. This relation is represented in the form of a formal context, \( I = (X, Y, R) \), where \( X \) is a set of objects, \( Y \) is a set of attributes, and \( R \) is a relation \( R \subseteq X \times Y \). Each formal context \( I = (X, Y, R) \) induces two operators: \( \uparrow : 2^X \to 2^Y \) and \( \downarrow : 2^Y \to 2^X \). The operators are defined, for each \( A \subseteq X \) and \( B \subseteq Y \):

\[
A^\uparrow = \{ y \in Y | \forall x \in A: (x, y) \in R \}, \quad B^\downarrow = \{ x \in X | \forall y \in B: (x, y) \in R \},
\]

where \( A^\uparrow \) is the set of all attributes shared by all objects in \( A \), and \( B^\downarrow \) is the set of all objects sharing all attributes in \( B \). A formal concept is a pair \( (A, B) \) with \( A \subseteq X \) and \( B \subseteq Y \) such that \( A^\uparrow = B \) and \( B^\downarrow = A \). The collection of all such pairs, together with the natural partial order \( (A_1, B_1) \leq (A_2, B_2) \) iff \( A_1 \subseteq A_2 \) \( (B_2 \subseteq B_1) \), form a complete lattice.\textsuperscript{14} The structure of the concept lattice corresponds to the hierarchy of formal concepts and provides generalization and specialization between the concepts, i.e., \((A_1, B_1)\) is more specific than \((A_2, B_2)\) (or \((A_2, B_2)\) is more general).

There are two basic algorithmic strategies for computing formal concepts: (a) forming intersections by joining one attribute at a time, and (b) repeatedly forming pairwise intersections starting from the attribute concepts. Kuznetsov and Obiedkov provided a survey and comparative evaluation of algorithms for FCA.\textsuperscript{16} Bordat\textsuperscript{17} used strategy (a), while Chein\textsuperscript{18} used strategy (b). In this study, we use the idea of Troy, Zhang, Tian’s\textsuperscript{19} multistage algorithm for constructing concept lattices (MCA) and adapt and extend it in order to take advantage of the scalable distributed Spark framework.

We use ontological definitions for constructing a formal context. One can think of ontological concepts as objects, and ontological relationships as attributes in constructing a formal context. For example, with respect to Figure 1 we can use [69483009,Shunt of cerebral ventricle to extracranial site (procedure)] as an object and the other lines as its attribute entries.

Cloud Computing. Apache Spark is an advanced and high-end upgrade to Hadoop, aimed at enhancing scalability and performance using Resilient Distributed Datasets (RDDs) which facilitates in-memory computations on large clusters in a fault-tolerant manner.\textsuperscript{20} While both share the key-value conceptual framework, Spark is generally faster than Hadoop for MapReduce because of the way it processes data. It completes the full data analytic operations in-memory and in near real-time: reading data from the cluster, performing the requisite analytic operations, and writing results to the cluster. The in-memory processing strategy makes Spark more suitable for iterative algorithms. However, designing efficient algorithms in a distributed way requires a different strategy than that for the traditional approach, with more attention paid to data locality, job break-down, and trade-offs between parallelism and communication latency.

2 Method

2.1 Overview

In this paper, we introduce Spark-MCA, a distributed algorithm based on MCA, using the theoretical basis of FCA and adapting it to the Spark framework. Spark-MCA contains two parts: distributed MCA for obtaining all concepts, and distributed big set operation for lattice diagram construction. Spark-MCA takes in the original concepts and relations from SNOMED CT in the format described in the previous section as source formal context and generates the entire concept lattice from this context. We illustrate the main steps of our algorithm in the following subsections with the help of an example dataset extracted from SNOMED CT (Table 2).

We performed two types of evaluation: (1) partial validity, and (2) computational performance. To demonstrate that our Spark-MCA can be used in auditing semantic completeness and the results of our algorithm can provide a viable candidate pool of new concepts in a developing ontological system, we apply RGT by comparing the results of Spark-MCA with the concepts added to SNOMED CT versions. For computational performance, we tested our algorithm for its scalability by comparing the running times for datasets of different sizes. We also evaluated our algorithm on its performance on the same dataset but using different numbers of working computer processors.
<table>
<thead>
<tr>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
<th>e</th>
</tr>
</thead>
<tbody>
<tr>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
</tbody>
</table>

Table 2: a: Shunt of cerebral ventricle to extracranial site, b: Neck repair, c: Ventricular shunt to cervical subarachnoid space, d: Creation of ventriculo-jugular shunt, e: Creation of subarachnoid/subdural-jugular shunt. 1: Creation of subarachnoid/subdural-jugular shunt (isa), 2: Brain ventricle structure (Procedure site - Direct), 3: Shunt action (Method), 4: Operation on neck (isa), 5: Surgical repair of head and neck structure (isa), 6: Repair - action (Method), 7: Neck structure (Procedure site - Direct), 8: Creation of connection from ventricle of brain (isa), 9: Repair of spinal meninges (isa), 10: Cervical subarachnoid space (Procedure site - Direct), 11: Spinal cord cerebrospinal fluid pathway operation (isa), 12: Cerebral ventricular shunt to venous system (isa), 13: Structure of jugular vein (Procedure site - Direct), 14: Structure of meningeal space (Procedure site - Direct), 15: Creation of vascular bypass (isa), 16: Anastomosis of veins (isa).

2.2 Model SNOMED CT Using the FCA

SNOMED CT uses name-value to describe composite expressions. Take [405813007 | Procedure site - Direct (attribute)] → 35764002 | Brain ventricle structure (body structure)] in Figure 1 as an example. Here, 405813007 | Procedure site - Direct (attribute) is the name part, while 35764002 | Brain ventricle structure (body structure) is the value part. We use two steps to construct a formal concept from such SNOMED CT expressions. First, we include SNOMED CT concepts objects and each of their name-value pair as one attribute. For example, Figure 1 is encoded as 69483009: 116680003 → 88834003; 405813007 → 35764002; 260686004 → 424208002, where we used identifiers for simplicity. Second, to enrich the attributes, we include its ancestors’ attributes by leverage transitive closure of the isa relation (identifier 116680003). With the second step, Figure 1 is encoded as 69483009: 116680003 → 138875005; 405813007 → 35764002; 260686004 → 424208002; 260686004 → 129264002, where 138875005 is the root of the sub-hierarchy and dose not have any ancestors.

2.3 Spark-MCA

Algorithm for Distributed MCA

The distributed part of Spark-MCA is built on top of MCA. The idea behind MCA is to perform multistage intersection operations on each pair of concepts from the initial concept set consisting of all objects, until no more new concept is generated. Each of the stages is independent of the previous ones and does not involve any concepts in the previous stages. For notational preparation, define $S \cap T := \{s \cap t | s \in S, t \in T\}$, where $S$ and $T$ are collections of subsets. With respect to a given formal context $(X, Y, R)$, the MCA algorithm involves the following iterative steps:

$$S_0 := \{X\},$$
$$S_1 := \{\{x\} | x \in X\},$$
$$S_{i+1} := (S_i \cap S_i) - \bigcup_{1 \leq k \leq i} S_k$$

for $i \geq 1$. One important intermediate step for each stage is to remove all existing concepts $\bigcup_{1 \leq k \leq i} S_k$ when forming $S_{i+1}$. This way, only newly generated (and necessary) concepts are kept for subsequent stages, resulting in potentially large savings in computational cost.

The strategy involved in designing a distributed algorithm is to decompose a complex job into a sequence of jobs and to distribute particular parts of the data to be processed on different processing nodes. Our idea is to leverage the independence of stages involved in MCA and design a distributed MCA in the following way: Instead of iteratively performing intersections of object concepts pairs, $(S_i \cap S_i) - \bigcup_{1 \leq k \leq i} S_k$ on a single computer in each stage, we employ the MapReduce paradigm to split the iteration of pairwise intersections into a number of partitions. Each partition is then processed on an independent compute node by a Map function. The outputs from the Map functions are then collected, shuffled, and sent to compute nodes to perform Reduce functions. The output of Reduce functions are split
again to repeat the above steps. While iterations involved in MapReduce become very expensive in a standard Hadoop MapReduce implementation due to the I/O latency between the MapReduce steps, the distributed MCA implemented using the Spark framework can take advantage of in-memory processing to reduce the I/O latency.

Data in Spark are generally represented in the form of MapReduce key-value pairs \(<k, v>\) and stored as resilient distributed datasets (RDDs). Algorithm 1 describes the concept formation phase in Spark-MCA to generate all concepts from a given context. First we initialize the primitive concept set using \(S = \{x | x \in X\}\). Then, each loop in the \(while\) block (lines 3 - 11) performs one pairwise intersection. In each iterative step, we first obtain all pairs of current concepts as the \(keys\) for the input to Map (line 5). Then we split the pairs into a number of partitions and send them to compute node. Every node performs Map tasks of intersection on \(x^{\uparrow}\) and \(x\) (lines 6 and 7). The Reduce tasks collect all \(keys\) of Map’s results as candidate concepts to perform next iteration, and add the outcome to the final result (lines 8 - 10). When no new concepts generated, the algorithm stops and returns the set of cumulated formal concepts \(C\).

<table>
<thead>
<tr>
<th>Stage</th>
<th>C1</th>
<th>C2</th>
<th>C3</th>
<th>C4</th>
<th>C5</th>
<th>C6</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>{1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16}</td>
<td>{1,2,3}</td>
<td>{4,5,6,7}</td>
<td>{1,2,3,4,5,6,7,8,9,10,11}</td>
<td>{1,2,3,4,5,6,7,12,13,14}</td>
<td>{1,2,3,4,5,6,7,13,15,16}</td>
</tr>
</tbody>
</table>

Table 3: The newly generated formal concepts in each stage for the formal context in Table 2 using Algorithm 1.

Table 3 shows the result of new formal concepts generated in each stage. Stage 0 generates all primitive concepts represented by intents. Stage 1 outputs two new concepts: C7 and C8. Stage 2 does not output any new concept, and thus the concepts generation phase stops.

Algorithm for Performing Big Set Operations
One criterion for the well-formedness of an ontology is that its hierarchical structure forms a lattice.\(^{11}\) By constructing lattice graphs, we can locate those newly generated concepts in the big picture of the entire ontology and help further inspection of the concept neighbors. With lattice graphs, our result can be a used for other structure-based auditing methods, as well as for visualization.

After all formal concepts are obtained, we have a nature transitively closed relationship with respect to set inclusion. For graph rendering, we need to construct the minimal, irreducible set of subset relations as an irredundant representation of set inclusion (modular transitivity). To do so, we use distributed set operations to construct lattice graphs from the output of the concept concept construction part of MCA (Algorithm 1). We design our lattice construction algorithm by leveraging distributed big set operations using the following strategy adapted from Zhang et al.\(^{13,21}\):

1. get all subset relations of one concept, 
2. remove other relations except for the minimal supersets.

More formally, given a set \(X\) in a collection of subsets \(C\), we use \(\uparrow X\) to denote the set of all common ancestors of \(X\), i.e.,

\[
\uparrow X := \{a \in C | \forall x \in X, x \subseteq a\}.
\]

Thus, \(\uparrow X\) represents the strict upper closure of \(X\). When \(X\) is a singleton, i.e., \(X = \{x\}\), we write \(\uparrow x\) for \(\uparrow \{x\}\). Similarly, we define \(\downarrow X := \{a \in C | \exists x \in X, x \subseteq a\}\). We have,\(^{13,21}\) for any set \(X, Y \subseteq C\),

\[
\downarrow X = \bigcap_{a \in X} \uparrow a, \quad \downarrow Y = \bigcup_{b \in Y} \uparrow b.
\]

Hence, the set of minimal upper bounds of \(X\) can be obtained using the formula \(\uparrow X - \uparrow(\uparrow X)\).
The distributed big set operations involve two MapReduce jobs: Marking Subset Relations (Algorithm 2) and Transitive Reduction (Algorithm 3). The input of Algorithm 2 is the collection of all formal concepts (intents, say) obtained from Algorithm 1. First we initialize a HashSet \( C' \) to store all concepts and distributed to all computational computers (line 2). Then all concepts are split into partitions to perform MapReduce tasks. In each task, we collect the subsets of concept \( c_i \) by iteratively checking each element in \( C' \) to see if it is a subset of \( c_i \) (lines 3-12). The output of Marking Subset Relations is a collection of concept-subsets pairs \((c_i, a_c)\), which will serve as the input of Transitivity Reduction. In Algorithm 3, concept and subsets pairs are first cached in \( CS' \) and sent to all nodes (line 2). Then each MapReduce task removes those subsets that are the subsets of the current concept on each pair. This is achieved by removing the union of the subset list \( S_i \) for each subset \( i \) with concept \( c \) in \( S_c \) (lines 3-10).

<table>
<thead>
<tr>
<th>Algorithm 2: Marking Subset Relations</th>
<th>Algorithm 3: Transitivity Reduction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Function Marking Subset Relations(C)</td>
<td>Function Transitive Reduction(CS)</td>
</tr>
<tr>
<td><strong>Data:</strong> Formal concept set ( C )</td>
<td><strong>Data:</strong> Concept-subsets pairs ( CS )</td>
</tr>
<tr>
<td><strong>Result:</strong> Concept-subsets pairs ( S )</td>
<td><strong>Result:</strong> Edge Pairs ( E )</td>
</tr>
<tr>
<td>2 initialize HashSet ( C' = Cache(C) )</td>
<td>initialize HashMap ( CS' = Cache(CS) )</td>
</tr>
<tr>
<td>3 Function Map(( c \in C )) ( = ) ( a = \emptyset )</td>
<td>4 ( s' = \emptyset )</td>
</tr>
<tr>
<td>5 foreach el ( \in C' ) do</td>
<td>5 foreach el ( \in s ) do</td>
</tr>
<tr>
<td>6 if el ( \subset c ) then</td>
<td>7 ( s' \leftarrow CS'[el] )</td>
</tr>
<tr>
<td>7 ( a \leftarrow el )</td>
<td>end</td>
</tr>
<tr>
<td>8 end</td>
<td>9 RDD&lt; ( Int, Set[Int] ) &gt; pair</td>
</tr>
<tr>
<td>10 end</td>
<td>11 Function Reduce(pair)</td>
</tr>
<tr>
<td>11 Function Reduce(c(a,c)) ( = ) ( CS \leftarrow (c,a) )</td>
<td>12 return ( E )</td>
</tr>
<tr>
<td>12 return ( CS )</td>
<td>13</td>
</tr>
</tbody>
</table>

Table 4 lists all the formal concepts of the formal context given in Table 2. If we equip the associated order on the concepts collection, we obtain the concept lattice shown in Figure 2. In the lattice graph, \( C_1 \) contains all the intents and \( C_0 \) contains all the extents. Concepts \( C_1, C_2, C_3, C_4, C_5 \), in blue, are primitive concepts. Concepts \( C_7, C_8 \), in red, are suggested new concepts.

<table>
<thead>
<tr>
<th>Concept</th>
<th>Extent</th>
<th>Intent</th>
</tr>
</thead>
<tbody>
<tr>
<td>( C_1 )</td>
<td>{}</td>
<td>{1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16}</td>
</tr>
<tr>
<td>( C_2 )</td>
<td>{a}</td>
<td>{1,2,3}</td>
</tr>
<tr>
<td>( C_3 )</td>
<td>{b}</td>
<td>{4,5,6,7}</td>
</tr>
<tr>
<td>( C_4 )</td>
<td>{c}</td>
<td>{1,2,3,4,5,6,7,8,9,10,11}</td>
</tr>
<tr>
<td>( C_5 )</td>
<td>{d}</td>
<td>{1,2,3,4,5,6,7,12,13,14}</td>
</tr>
<tr>
<td>( C_6 )</td>
<td>{e}</td>
<td>{1,2,3,4,5,6,7,13,15,16}</td>
</tr>
<tr>
<td>( C_7 )</td>
<td>{d,e}</td>
<td>{1,2,3,4,5,6,7,13}</td>
</tr>
<tr>
<td>( C_8 )</td>
<td>{c,d,e}</td>
<td>{1,2,3,4,5,6,7}</td>
</tr>
<tr>
<td>( C_9 )</td>
<td>{a,b,c,d,e\}</td>
<td>{}</td>
</tr>
</tbody>
</table>

Table 4: The list of formal concepts generated from the formal context given in Table 2.

In Spark-MCA, each concept is represented by its intent (and refers neither to the original context, nor to any extent). One can easily incorporate a data structure for looking up the extent of a concept, after all concepts are determined. In this study, we maintain two hash maps: one is \( C \), which, for each concept \( x \in X \), collects all its attributes; similarly, the other is \( A \), which, for each attribute \( y \in Y \), collects all concepts that contain a particular attribute. Hence, the intersection \( \bigcap a_i^+ \), where \( a_i \) is the attribute of concept \( c_i \), contains all extents of a concept. To identify the labels of formal concepts, we compare their attributes with the primitive concepts list. Concepts corresponding to the primitive concepts are naturally labeled, while the others concepts can be treated as a candidate pool of new concepts for enhancing semantic completeness.
2.4 Evaluation Method

SNOMED CT provides a “delta” file, which contains changes of concepts and relations associated with each release. To evaluate the performance of Spark-MCA, we constructed RGT by obtaining concepts that were added in the releases. Obtaining an appropriate set of changes across multiple SNOMED CT releases is not a mere job of performing unions. For the purpose of evaluation, we constructed a reference set involving two steps: (1) extract added concepts (active value: 1) from all 5 “delta” files from 201403 release to 201609 release; (2) remove deleted or revised concepts (active value: 0) from the added concept sets obtained in the first step. These remaining concepts are selected as a RGT for evaluating the result of Spark-MCA. Since the suggested concepts by Spark-MCA have no labels, we matched the corresponding concepts using their attribute sets (intents).

For scalability analysis, we executed Spark-MCA on Amazon Web Services (AWS) with different hardware configurations. For this study, we selected M4, the latest generation of general purpose instances as the main working configuration. The hardware configuration for M4 is as follows: (1) 2.3 GHz Intel Xeon® E5-2686 v4 (Broadwell) processors or 2.4 GHz Intel Xeon® E5-2676 v3 (Haswell) processors, (2) EBS-optimized by default.

3 Results

3.1 Semantic Completeness Analysis

We applied Spark-MCA to the SNOMED CT 201403 release and found a total of 500,583 formal concepts suggested by Spark-MCA that were not included in this release. For evaluation, RGT identified 22,687 concepts as additions based on the 5 “delta” files from 201403 to 201609. A total of 3,231 concepts were found in the intersection of the two groups of concepts. This means that Spark-MCA correctly identified 3,231 concepts as missing concepts in the SNOMED CT 201403 release. Table 5 shows the numbers of identified concepts in representative subhierarchies. In this table, “FCA-New” denotes the number of suggested new concepts by Spark-MCA; “Delta” denotes the cumulative number of concept additions based on the “delta” files; “IS-A Only” denotes the number of concepts in “delta” files with no attributes; “New attributes” denotes the number of concepts in the “delta” files whose definition involved attributes that are not in primitive concept list; and “Matched” denotes the number of concepts in the “delta” files that were also found using Spark-MCA.

<table>
<thead>
<tr>
<th>Subhierarchy (# of concepts)</th>
<th>FCA-New</th>
<th>Delta</th>
<th>IS-A Only</th>
<th>New attributes</th>
<th>Matched</th>
</tr>
</thead>
<tbody>
<tr>
<td>Body structure (30,623)</td>
<td>11,311</td>
<td>743</td>
<td>225</td>
<td>0</td>
<td>353</td>
</tr>
<tr>
<td>Clinical finding (100,652)</td>
<td>153,401</td>
<td>10,121</td>
<td>309</td>
<td>5,563</td>
<td>1,825</td>
</tr>
<tr>
<td>Pharmaceutical biologic product (16,797)</td>
<td>7,696</td>
<td>2,099</td>
<td>106</td>
<td>1,634</td>
<td>109</td>
</tr>
<tr>
<td>Procedure (54,091)</td>
<td>325,114</td>
<td>2,893</td>
<td>178</td>
<td>777</td>
<td>875</td>
</tr>
<tr>
<td>Situation with explicit context (3,723)</td>
<td>1,718</td>
<td>991</td>
<td>12</td>
<td>751</td>
<td>45</td>
</tr>
<tr>
<td>Specimen (1,475)</td>
<td>1,283</td>
<td>200</td>
<td>10</td>
<td>94</td>
<td>24</td>
</tr>
<tr>
<td>Event (3,683)</td>
<td>14</td>
<td>23</td>
<td>23</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Staging and scales (1,308)</td>
<td>6</td>
<td>123</td>
<td>123</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 5: Comparison of Spark-MCA results and retrospective ground truthing with respect to main subhierarchies.

3.2 Scalability Analysis

Scalability is one of the hallmarks of cloud computing. We performed two scalability tests for our Spark-MCA by analyzing computational time for different subhierarchies of SNOMED CT and for different AWS configurations, respectively. For the first test, we selected 4 largest subhierarchies in SNOMED CT 201403 release: Body structure, Clinical finding, Procedure, and Pharmaceutical biologic product. We performed the test on an AWS cluster with 96 processors. The computational time for each subhierarchy is shown in Table 6. For the second test, we repeatedly ran Spark-MCA on the Body structure subhierarchy with different configurations by increasing the number of processors in the AWS cluster. Figure 3 shows the time taken for computing concepts, and Figure 4 shows the time taken for concept lattice graphs. We found a linear decrease of the computational time when the number of computation processors ranged from 1 to 30. When the number of processors approached 100, the computational time tends to cease to decrease, indicating a threshold at which additional processors would not be helpful.
3.3 Evaluation

We applied a Precision-Recall metric to the comparison of our results and RGT. Table 7 shows the $P_1$, $R_1$, and $R_2$ scores in terms of subhierarchies (see the table caption for the definitions of the scores). The $P_1$ values are not so good and we did not calculate the precision by removing concepts in “delta” with no attributes or new attributes. In any case, this would not cause a big difference because of the large number of results from Spark-MCA. For smaller “delta” files, precision may not be so significant. From Table 7 we can see that in the 3 largest hierarchies, Body structure, Clinical finding, and Procedure, our Spark-MCA can identify 47.5%, 18.0%, and 30.2% of all newly added concepts, respectively. If we removed those concepts without attributes or with new attributes, the proportions become even larger. This indicates that Spark-MCA could be a reliable approach for addressing semantic completeness.

<table>
<thead>
<tr>
<th>Subhierarchy</th>
<th>$P_1$ (%)</th>
<th>$R_1$ (%)</th>
<th>$R_2$ (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Body structure</td>
<td>3.12</td>
<td>47.5</td>
<td>68.1</td>
</tr>
<tr>
<td>Clinical finding</td>
<td>1.19</td>
<td>18.0</td>
<td>43.0</td>
</tr>
<tr>
<td>Pharmaceutical biologic product</td>
<td>1.41</td>
<td>5.2</td>
<td>30.4</td>
</tr>
<tr>
<td>Procedure</td>
<td>0.27</td>
<td>30.2</td>
<td>45.1</td>
</tr>
<tr>
<td>Situation with explicit context</td>
<td>2.62</td>
<td>4.5</td>
<td>19.7</td>
</tr>
<tr>
<td>Specimen</td>
<td>1.87</td>
<td>12</td>
<td>25</td>
</tr>
<tr>
<td>Event</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Staging and scales</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 7: $P_1$: ratio of Spark-MCA resulting concepts that can be found in the “delta” files. $R_1$: ratio of delta concepts that can be identified by Spark-MCA, including all concepts from “delta.” $R_2$: ratio of delta concepts that can be identified by Spark-MCA, removing concepts in “delta” with no attributes or new attributes.

From Table 6, we can see that larger formal contexts required more computational time using our algorithm. However, we point out that because of the density and complexity, we cannot assume that a context with more objects and attributes would necessarily take more computational time. For example, the Clinical finding subhierarchy had more objects (concepts) and attributes, but it needed less computational time than did the Procedure subhierarchy. For Procedure, the process was completed in about 1.5 hours on 96 processors, which might also be completed in days using a sequential approach. From Tables 3 and 4, we can see that as the number of processors increase, the time taken for processing the same dataset declined significantly in the range of 1 to 30 computer processors. Due to the shuffle stage in Spark, the total computational time became stable as the number of processors increase. In conclusion, our computational experiments did demonstrate the scalability of our Spark-MCA algorithm.
4 Discussion

Limitation of Spark-MCA. Although our results showed Spark-MCA to be a practical method in evaluating semantic completeness, there remain a number of limitations. First, Spark-MCA cannot find those added concepts in “delta” involving newly added attributes. For example, Difficulty swimming (714997002) is added with a new attribute 363714003—Interprets (attribute) = 714992008—Ability to swim (observable entity) in the 201603 release, but Spark-MCA did not include it. Other types of lexical or structure auditing methods can be useful supplements to Spark-MCA in auditing semantic completeness. Second, it is less useful to apply the FCA approach to terminological systems with very minimal semantic definitions (e.g., limited to one type of relationship). Subhierarchies in SNOMED CT involving only the isa relation where Spark-MCA suggested no new concepts, even though these were present in the delta files (numbers given in parenthesis indicate additions in the delta files) are: Environment or geographical location (125), Observable entity (418), Organism (1,118), Physical force (0), Physical object (631), Qualifier value (649), Record artifact (32), Model Component (69), Social context (33), Special Concept (0), Substance (2,420).

SNOMED CT versions used for RGT. We used the latest five versions of SNOMED CT (the March 2017 version was released too late to be included in this study) for the feasibility demonstration of our method. This choice was purely a matter of convenience and did not result from methodological or computational limitations of our approach. In fact, it would be desirable to use all the available SNOMED CT versions (in Release Format 1 as well as Release Format 2), with an appropriate starting release as the time-point for computing the formal concepts. All the performance measures would only improve, and the RGT reference set would become larger.

Limitation of RGT. To understand why some concepts suggested by Spark-MCA were not included in SNOMED CT, and some were, we need domain experts assessment and case study to improve gain insights. Such insights may help enhance the method further.

Comparison with related results. We examined the example subhierarchy (hypophysectomy) provided in Jiang and Chute's work. We found our results to be in agreement. However, because Jiang and Chute did not provide details about all their findings, we could not perform a more thorough comparison.

Sensitivity of the FCA-based model. FCA-based approach is in general known to be sensitive to the density and complexity of the input formal context. Even though Spark-MCA performed reasonably well for SNOMED CT, we did not perform similar computational experiments on other terminology systems.

5 Conclusion

In this study, we introduced Spark-MCA, a scalable approach for evaluating the semantic completeness of SNOMED CT using an FCA-based method on top of the Spark cloud-computing framework. We formulated SNOMED CT into a formal context in FCA and then used Spark-MCA to exhaustively compute the formal concepts of the context as well as the associated subset relations. We the applied Retrospective Ground-Truthing to assess the performance of Spark-MCA. Our results show that Spark-MCA provides a cloud-computing feasible approach for evaluating the semantic completeness of SNOMED CT using formal concept analysis.

6 Acknowledgment

This work was made possible by the University of Kentucky Center for Clinical and Translational Science (Clinical and Translational Science Award UL1TR001998). This work was also supported by the National Science Foundation under MRI award No. 1626364. We thank the Amazon Web Services (AWS) for its Cloud Credits for Research program that enabled the implementation of our distributed algorithms in Spark.

References


User-Centered Design and Usability Assessment of an EHR Integrated Automated SOFA Calculator App

Christopher A. Aakre, MD, Jaben E. Kitson, Man Li, Vitaly Herasevich, MD, PhD
Mayo Clinic, Rochester, Minnesota

Abstract
Sepsis is a frequently encountered, life-threatening condition with significant in-hospital mortality. A recent change to the clinical definition of sepsis introduced the need for frequent recalculation and comparison of the Sequential Organ Failure Assessment (SOFA) score. Automation can overcome this clerical burden. We describe the user-centered user-interface design and usability assessment of an EHR-integrated app for a previously validated near real-time automated SOFA scoring algorithm that meets clinicians’ needs.

Introduction
In March 2016, the Third International Consensus Definitions for Sepsis and Septic Shock (SEPSIS-3) redefined sepsis to include a change in SOFA score ≥ 2 points over baseline. Subsequently, the SOFA score must be frequently recalculated to identify sepsis—increasing clerical burden. In response to the increased task-load, Mayo Clinic has developed and validated an automated SOFA score calculator algorithm. In this work, we describe a user-centered design process for the creation of an “app” that integrates this automated SOFA score calculator algorithm into our local critical care EHR.

Methods
We utilized an agile development process to create a user interface (UX) for an automated SOFA score calculator integrated into our local critical care EHR. Clinician stakeholders were interviewed to determine (1) essential and (2) nice to have features and how to display each SOFA component. UX prototyping was performed using Pencil, an open source graphical user interface prototyping tool, and modified iteratively with feedback from prospective end-users until consensus was attained. The final UX’s usability was assessed with the Computer Systems Usability Questionnaire by clinician end-users with real world usage of the app over a two-month period in the medical intensive care unit.

Results
Among the 12/50 (24%) usability survey respondents, our user-centered UX design process resulted in > 75% favorability of survey items in the domains of system usability, information quality, and interface quality.

Conclusion
By engaging end-user stakeholders throughout the UX design process, we successfully created a user-centered UX for an automated SOFA score calculator app integrated into our local critical care EHR. Emerging interoperable platforms may allow for rapid creation and distribution of clinical decision support algorithms and calculators as apps. App developers will need to consider UX usability and user-centered design during creation of these tools.

Figure. Percent of responses categorized as favorable, unfavorable, or neutral within each domain from the post-implementation Computer Usability Scale Questionnaire (Respondents = 12).
Cranky Comments: Detecting Clinical Decision Support Malfunctions Through Free-Text Override Reasons

Skye Aaron1, Adam Wright, PhD1,2,3

1Brigham & Women’s Hospital, Boston, MA; 2Harvard Medical School, Boston, MA; 3Partners HealthCare, Boston, MA

Introduction: Electronic health record (EHR) systems often provide a free-text area for clinicians to enter an override reason when overriding a clinical decision support (CDS) alert. Override reasons may be used for auditing purposes but typically are not systematically shown to other providers or to EHR technical support. Nonetheless, it has been found that a significant fraction of override reasons are addressed to a recipient who will never be shown the comment.1 Free-text override reasons have previously been studied to categorize types of reasons provided, to determine the clinical appropriateness of alert overrides, and to investigate whether override reasons contain clinically useful information.1 However, no study has used override reasons as a method of detecting alerts that do not function as intended. There is growing evidence that most institutions with CDS will, at some point, have alerts implemented in their EHR that are operating incorrectly, and these malfunctions may go undetected for long periods of time.2 This project investigates the usefulness of free-text override reasons to detect alert malfunctions and prototypes techniques to automatically detect malfunctions by analyzing free-text override reasons.

Methods: Free-text override reasons were collected for all best practice alerts (BPAs), excluding medication alerts, that fired from 5/2015-3/2017 at Partners HealthCare System. BPAs were flagged for review using four detection methods: (1) total number of alert overrides with free-text override reasons; (2) frequency of words that express frustration (as selected by a human reviewer); (3) a high ratio of negative to positive comments as categorized by a Naïve Bayes classifier trained on pre-labeled movie reviews; and (4) similarity of comments within a BPA based on the cosine similarity measure. For each of the detection methods, a human evaluator then reviewed the top 10 flagged BPAs and categorized them as “likely broken,” “likely not broken,” and “unclear.” Flagged BPAs categorized by the human evaluator as “likely broken” were further investigated using the rule logic and data from patient charts to determine whether the alert was malfunctioning and could be fixed.

Results: There were 36,210 free-text override reasons submitted for 151 unique BPAs during the time period. Each of the detection methods yielded BPAs the reviewer classified as “likely broken”: 3 of the top 10 for method 1, 8 of the top 10 for method 2, 6 of the top 10 for method 3, and 6 of the top 10 for method 4. Additionally, all four methods uncovered alerts that were working as designed but could be improved using advice in the override reasons. Table 1 shows three rule malfunctions discovered by the detection methods, along with sample override reasons.

Table 1. Rules that are likely broken based on user-provided override reasons.

<table>
<thead>
<tr>
<th>Rule description</th>
<th>Sample override reasons suggesting a rule is broken</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient is due for a Cyclosporine Level</td>
<td>“stupid reminder-N/A for ophthalmic CyA” “why?” “inaccurate warning patient on topical restasis does not need level”</td>
</tr>
<tr>
<td>Patient is at risk for developing DKA. Please order long-acting insulin.</td>
<td>“I ALREADY GAVE INSULIN!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!” “he hed [sic] it already - so why do you bother me with this advisory you idiots” “wrong!!! She gets them per GI!! Stop these stupid warnings that are inaccurate”</td>
</tr>
<tr>
<td>Patient is taking digoxin and potassium level is less than 3.0</td>
<td>“please fix this advisory! K+ is the same as potassium” “serum K+ is potassium level- pease [sic] FIX THIS!!!” “Inaccurate warning. K+ is 4.4 today.”</td>
</tr>
</tbody>
</table>

Discussion: Free-text override reasons were useful for discovering BPA malfunctions. Several BPA malfunctions that were discovered have been fixed, and more will be investigated. Although the sentiment analysis method using positive and negative movie reviews is preliminary, it nonetheless uncovered six BPAs that were likely to be broken. Next steps for this project will refine the human evaluation of flagged alerts and improve the application of sentiment analysis by using context-specific training data as well as other types of classifiers.

Conclusion: Free-text override reasons can be used to detect CDS alert malfunctions and to identify alerts that are functioning as designed but could be tailored or improved.

References:
Characterizing a Learning Curve for Annotating Data for Training and Validation of Natural Language Processing Systems

Glenn L. Abastillas, MA RN¹, Spencer D. Morris, MS², Jessica A. Boten, MPH², Tumenbayar Tumurchudur, MA¹, Kinjal Vora, MD-MPA³, Paul A. Fearn, PhD²

¹Information Management Systems, Incorporated, Rockville, MD; ²National Cancer Institute, Rockville, MD; ³Vasta Global, New York City, NY

Background
The National Cancer Institute’s (NCI) Surveillance, Epidemiology, and End Results (SEER) Program provides cancer statistics representing 28% of the US population. To accelerate population-based research to improve patient outcomes in cancer, the NCI is developing a natural language processing (NLP) pipeline to extract structured information from unstructured pathology reports in SEER data. NLP systems often need a large set of human-annotated data for training and testing of algorithms and systems, which is resource-intensive and time-consuming to produce¹,². There is no standardized method used to forecast effort needed for annotation, which would be useful for efficient allocation of resources to NLP projects. We aim to characterize a learning curve (LC) on our data for developing such a method informed by learning curve research from both manufacturing and informatics industries.

Methods
We analyzed 100 de-identified pathology report documents annotated by a pair of two-person annotator-reviewer teams for elements relevant to biomarkers of interest. Annotators extracted information in the form of highlighted text and codified data (annotations) from each document. Reviewers performed quality checks on their annotations per iteration. We analyzed the empirical cumulative distribution (ECD) of documents completed against time required for completion and compared the ECD difference subset by all annotator and reviewer pairs (n = 4).

Results & Discussion
The ECD in Figure 1 highlights a difference in the LC depending on annotator-reviewer team configuration. There is a significant difference in LCs depending on the reviewer who checked annotations (p < 0.05; p = 0.001) rather than the annotator who produced them (p > 0.05; p ≈ 0.92). This may indicate that the review process matters greatly in completing quality annotations. We intend to analyze more incoming data to further develop our analysis and LC.

![Figure 1](image)

Figure 1 Each panel represents our current LC model according to each pair of reviewers (R1 and R2) and annotators (A1 and A2). The color lines represent documents with reviewer-identified revisions (teal) and without (pink). Easier learned tasks tend to have a curve approaching the top-left of the chart. Harder learned tasks tend to have a curve approaching the bottom-right.

References
2. Lughofer E. Hybrid active learning for reducing the annotation effort of operators in classification systems. Pattern Recognition 2012;45:884-896.
Extraction of Patient Temporal Patterns and Clusters from Clinical Data

Samir E. Abdelrahman1, Julio C. Facelli1, Bruce E. Bray1,2, Rashmee U. Shah2, Guilherme Del Fiol1
Department of Biomedical Informatics1 and Division of Cardiovascular Medicine2, School of Medicine, The University of Utah

Introduction: Descriptive data mining techniques, such as temporal pattern mining and clustering analysis, may provide clinicians with deeper understanding of the associations and/or correlations among clinical variables that may predict patient outcomes. Temporal patterns and clustering of patient data evolve due to changes in the patient’s clinical status and proper analytical techniques for their analysis should be informative and give stable results over time. Here we report the application of our Frequency pattern mining, Clustering, and Graph database (FCG) method1 to extract temporally stable and accurate patterns and clusters of patients, as described by clinical and demographic variables measured in an Intensive Care Unit (ICU) setting. The accuracy of patterns and clusters is measured by their ability to correctly discriminate between positive (not-survival) and negative (survival) cases at any given time.

Methods: We developed a temporal descriptive framework with two steps: data preprocessing and tree pruning and clustering using FCG. In the data-preprocessing step, all numeric values were categorized into hourly “low”, “normal”, or “high” levels, using standard reference values. If a patient has multiple readings for the same variable within the hour, we considered the last reading as representative of the patient status in that hour. In the FCG step, the data is represented as a regulated hierarchical tree of frequencies. The tree is cut at a certain level node to remove the patterns of all the preceding nodes, which are deemed as non-informative. Then the clustering analysis is conducted for the pattern’s successive nodes. A graph database was used, as it can provide a permanent storage to support retrieving data for visualizing multiple trees/clusters in a scalable manner.

We analyzed data from each hour of the first 48 hours of ICU patients admitted to a cardiac service, using their mortality outcome as the dependent variable. We extracted demographic information and 100 clinical indicators1 such as vital signs and laboratory tests for 1,000 patients that were randomly selected from the Medical Information Mart for Intensive Care (MIMIC III) database with mortality rate = 30%. The data-preprocessing step produced 480,000 values for each patient. The FCG default values were utilized such that the minimum numbers of readings found in the patient records data was used for defining the cut tree levels and one iteration was performed for merging the clusters of each pattern in that hour. We tested four different tree cut levels (i.e. pattern lengths) of 1, 2, 3, and 4 patterns for which we tested the accuracy of the clusters for each hour. For each cluster, we used the F-measure of the negative cases and the cluster purity (i.e., its ability of cluster to present the most positive or negative cases) (p < 0.05)1 as measures of the performance of the method.

Results: Fig.1 shows the mean accuracies and their standard deviations for every cut-off level during the 48-hour period. The standard deviations represent the stability over the whole timeframe of that cut-off level.

Discussion: Although, we used the framework default values, the analysis produced stable results of the average 48-timeframe accuracies for all cut-off levels considered here. However, the accuracy of the results increase significantly with increasing cut-off levels. However, the best cut-off level depends on the user input threshold.

Conclusion: The proposed framework provides stable and more accurate results with increasing retrieval features as the cut-off level increases. Since the patterns and clusters are accurate overtime with respect to patient outcome, this approach may be useful for clinical decision support in the ICU. We are currently using different strategies to present time variabilities of patient readings. Further algorithmic enhancements will be developed.

References:
1. Samir Abdelrahman, and Bruce Bray, Frequency tree clustering for ICU mortality analytics using graph database: circulatory system diseases case study, IEEE Conference on Bioinformatics and Biomedicine (BIBM), 2016.
Dissemination of HIV, HCV and STD Clinical Evidence to Primary Care Providers – An Analysis of User Profiles and Course Evaluations of a Statewide Online Clinical Education Program

Meredith Abrams, MPH; Dongwen Wang, PhD
Arizona State University, Scottsdale, AZ

Introduction

Rapid advances in HIV, HCV, and STD research have generated important clinical evidence. Timely and effective dissemination of such evidence to primary care clinicians is essential for translating the latest research findings into the daily care of patients. With the wide use of information technology, continuing professional development courses are increasingly offered online. The New York State (NYS) HIV-HCV-STD Clinical Education Initiative (CEI) online training program has developed hundreds of multimedia learning modules, online CME/CNE courses, interactive case simulation tools, and other internet-based resources since its launch in 2009. These resources have been disseminated to tens of thousands healthcare providers from 170+ countries. This study analyzes the profiles of clinicians participating in the CEI online CME/CNE course during the study period. For each parameter of the user profile and seven categories of training topics, we performed univariate analyses with chi-square test to identify the potentially significant factors for positive course evaluations.

Methods

We compiled CEI online training user profiles and course evaluation data from November 2014 through October 2015. We collected clinicians’ personal and professional background such as demographics, education level, primary professional discipline, practice years, employment setting, employment location, caseload, and clinical services provided. We measured course evaluation through clinicians’ self-reported data on usefulness/relevance of information, ease of comprehension, knowledge of the trainer, appropriateness of format, knowledge level before and after training, intention to utilize the learned knowledge, and plan to change clinical practice. We included in the analyses only those clinicians with complete user profiles and passed at least one online CME/CNE course during the study period. We recorded significant differences in knowledge increase by: (1) training topic (p<0.01); (2) race (p<0.01); (3) education level (p<0.01); (4) discipline (p<0.01); (5) employment setting and location (p<0.01); (6) HIV, STD, and HCV caseloads (p<0.01); and (7) most HIV, HCV, and STD services (p<0.01). We recorded significant differences in knowledge increase by: (1) training topic (p<0.01); (2) race (p<0.01); (3) discipline (p<0.01); (4) practice years (p<0.01); (5) employment setting and location (p<0.01); (6) HIV, STD, and HCV caseloads (p<0.01); and (7) most HIV, HCV, and STD services (p<0.01). We recorded significant differences in intention to change practice by: (1) course topic (p<0.01); (2) race (p<0.01); (3) ethnicity (p<0.01); (4) education level (p<0.01); (5) discipline (p<0.01); (6) practice years (p<0.01); (7) employment setting (p<0.01); (8) HIV, HCV, and STD caseload (p<0.01); and (9) most HIV, HCV, and STD services (p<0.01).

Conclusion

Clinicians had very positive evaluations of the CEI online courses. We have identified potential factors for positive impact to knowledge increase, intention to use knowledge, and plan to change clinical practice. Future development of online CME/CNE courses may need to incorporate these factors for more effective knowledge dissemination.
A Resource for Clinical Semantic Textual Similarity

Naveed Afzal, PhD, Yanshan Wang, PhD, Feichen Shen, PhD, Liwei Wang, MD, PhD, Hongfang Liu, PhD
Department of Health Sciences Research, Mayo Clinic, Rochester, MN

Introduction
There is a growing need for automated methods to better synthesize patient data from electronic health records (EHRs) and reduce the cognitive burden in clinical decision-making process for providers. Patient data can be scattered in several heterogeneous sources. Tools that can aggregate data from diverse sources and minimize data redundancy, organize and present the data in a user-friendly way to reduce the cognitive burden are desired. One necessary task for extracting and consolidating information is to recognize semantically similar text snippets, a natural language processing (NLP) task known as semantic textual similarity (STS), which computes the semantic similarity between text snippets. In the general English domain, the SemEval STS share tasks have been organized since 2012 to develop automated methods for the task. Clinical text contains highly domain-specific terminologies and thus specific NLP tools and resources tailored to the clinical domain are needed for analysis, interpretation and management of clinical text. In this study, we describe our effort to create an STS resource consisting of sentence pairs from clinical notes with semantic similarity assigned by clinical experts.

Methods
The construction of a dataset by gathering naturally occurring pairs of sentences with different degree of semantic equivalence itself is a very challenging task. We selected unique sentences from 3 million de-identified clinical notes of patients receiving their primary care at Mayo Clinic. For sentence pair’s selection, we used various surface lexical similarity measures that enable us to find sentence pairs sharing some level of surface similarity. We used three surface level similarity measures: sequence-matching algorithm, cosine similarity and Levenshtein distance. All methods assign a score between 0 to 1 where 0 indicates no similarity and 1 otherwise. We average these three scores to get the final surface similarity score for a sentence pair. We experimented with different score ranges and empirically selected all sentence pairs where the average score was greater or equal to 0.45. In order to build sentence pair’s dataset that would reflect a uniform distribution of similarity ranges, we randomly selected equal number of sentence pairs from five scales of surface similarity range [0.45 – 0.95] from the dataset. We sampled 1,250 pairs and filtered sentence pairs that contain personal health information and that left us with 1068 sentence pairs. After the sentence pair selection phase, two clinical experts, each with many years of experience, were asked to annotate each sentence pair on the basis of their semantic equivalence using a scale of 0-5 (low to high similarity) using the provided guidelines and examples. A similarity score of 0 denotes complete semantic dissimilarity between two sentences. A similarity score of 1 shows that two sentences are not equivalent but are topically related to each other, while similarity score of 2 indicates that two sentences agree on some details mentioned in them. The similarity score of 3 implies that there are some differences in important details described in two sentences while a score of 4 represents that the differing details are not important. The score of 5 represents that two sentences are completely similar. The final similarity score for each sentence pair was obtained by averaging both annotators’ similarity score.

Results
Given a sentence pair, an STS system would need to return a similarity score. For a baseline system, we produced final score by averaging aforementioned surface similarity measures for each sentence pair. The system performance was evaluated using the Pearson correlation coefficient between the system output scores and the average human scores; we obtained a Pearson correlation coefficient score of 0.68 for baseline system.

Conclusion
The ability to organize concepts on the basis of their similarity or relatedness to each other is an essential step in the human mind and in many NLP applications. In this study, we have illustrated our effort to address a gap in existing resources for the study of STS in clinical domain. We selected sentence pairs from clinical domain using various surface-based similarity measures and each pair was annotated on the basis of their semantic similarity by human annotators. The use of STS can significantly reduce excessive redundant information that results in information overload, cognitive burden and difficulties in effective decision-making process at point-of-care.
Introduction
Lack of timely follow up on abnormal test results is common and linked to missed or delayed diagnosis, resulting in negative patient outcomes and increased malpractice\(^1\). To examine the safety culture surrounding test result follow up, we used qualitative and quantitative patient safety data in the ambulatory setting. We inspected how patients and providers estimated the frequency for providing timely test results; how patient satisfaction with their provider correlated with their provider’s response time to test result messages; and qualitative themes in patient complaints and safety reports filed by clinic staff. While this data has been used individually for quality improvement, we believe that examining these metrics together offer valuable insight into test result management culture.

Methods
Data was requested for outpatient clinics affiliated with Brigham Health, a tertiary academic medical center. To examine how providers estimated their frequency for providing timely test results, we analyzed data aggregated on the clinic level from Agency for Healthcare Research and Quality Surveys on Patient Safety Culture (AHRQ SOPS) from 2012 (1,482 surveys) and 2014 (1,739 surveys). To examine how patients perceive the timeliness of returned results and their satisfaction with their provider, we utilized data from the Consumer Assessment of Healthcare Providers and Systems (CAHPS) Clinician and Group Survey for patients who had an office visit between 12/1/15 – 5/31/16 (8,707 surveys returned), aggregated on the clinician level. Additionally, provider response times to test message results was calculated for messages received between 6/1/15 – 8/26/16. Providers were matched to CAHPS survey responses via provider ID. Patient complaint data pertaining to test results categories from the Brigham Health Patient Family Relations department from 1/1/11 to 4/30/16 (2,083 cases), and safety reports filed by Brigham employees between 1/1/11 to 9/18/16 (6,381 reports) were analyzed for narrative trends.

Results
Preliminary analysis of SOPS data showed a weighted average score of 0.88+/-.14 in 2012 and 0.83+/-.12 in 2014 for Brigham clinics for a question asking about critical abnormal result follow up\(^2\). The SOPS score indicates the proportion of returned surveys with a favorable response. Analysis of CAHPS data on how frequently a patient perceives that the provider’s office follows up on test results showed that 73.8% were returned with the patient marking “Always”\(^3\). We also correlated the percentile rank of individual providers with the median amount of time the provider took to respond to high priority messages. With a cut-off of 20 messages and 20 returned surveys, we found that as response time decreased, satisfaction increased (p=0.0073). We isolated 144 complaint cases and 737 safety reports that were caused by delayed test results. The most common narrative trend for patient complaints was breakdown in communication between provider or clinic and patient. The most common issue for safety reports was delivery or transportation of collected samples.

Discussion/Conclusion
We investigated the culture of test result follow-up using tools typically used to examine patient satisfaction and experience and staff culture. Pairing together CAHPS, a survey tool that measures patients’ perception of perceived care, and SOPS, a survey tool that examines employees’ perception of given care, allowed us to gather quantitative data on test result follow-up culture. Patients and staff perceived test result follow-up similarly, with the average practice scoring around 0.8 and 74% of patients saying that they always receive prompt follow-up. Similarly, we found a slight correlation between how long providers take to respond to message and how satisfied the patient is with their care. Additionally, while examining patient complaints and staff safety reports, we saw similar narrative trends. Using these four sources of data allowed us to examine multiple perspectives in follow-up culture and to identify possible explanations for inappropriate follow-up. We hope to further explore these data sources to identify solutions to inappropriate follow-up.

References
Releasing De-Identified Clinical, Imaging, and Genomic Data from the VA to External Repositories for the APOLLO Network

Samuel Ajjarapu MD1, Frank Meng PhD1,2, Danne C. Elbers MS1, Nhan V. Do MD1, Robert Hall MS1, Karen Pierce-Murray RN1, Luis E. Selva PhD1, Beth Katcher MS1, Brett C. Johnson PhD1,4, Andrew Zimolzak MD1,3,4, Corri DeDomenico BS1, Mary Brophy MD1,3, Louis Fiore MD1,3

1MAVERIC, VA Boston Healthcare System, Boston, MA
2Medical Imaging Informatics, Dept. of Radiological Sciences, UCLA, Los Angeles, CA
3Department of Medicine, Boston University School of Medicine, Boston, MA
4BD-STEP Program, VA and NCI

Introduction

Much of the recent advancements in medicine have been driven by data-centric approaches that utilize big data sets to derive knowledge for informing treatment options. As the direction of this progress continues, one of the limiting factors will be the availability of large-scale repositories that contain heterogeneous patient data. Because of the sensitivity of releasing PII, many data sets continue to be housed behind firewalls and are not easily accessible by most researchers. This greatly limits their utility because potential users typically lack the time and resources needed to successfully overcome the various hurdles for gaining access. The work described in this poster is being done under the auspices of the APOLLO (Applied Proteogenomics Organizational Learning and Outcomes) Network, a collaboration between the NCI, the DoD, and the VA for studying the proteogenomic pathways that impact disease. In support of this effort, the VA has been actively engaged in establishing regulatory and technical workflows that facilitate the release of de-identified patient clinical, imaging, and genomic data from various sources to partner institutions, specifically the Genomic Data Commons (GDC) and The Cancer Imaging Archives (TCIA) of the NCI. We believe that the lessons learned and the various artifacts produced from our efforts to release this data will prove useful to other researchers and decision-makers who desire to make their own data available to the broader community.

Methodology

Transferring VA data to external repositories consisted of several major tasks: obtaining patient consent to release their data; curating the data from multiple sources both within and external to the VA; de-identifying the data to remove all PII; and obtaining all regulatory approvals for the established processes and workflows. The following sections will describe these steps in more detail.

Patient Consent. Formal consent was obtained from patients through an established protocol for gathering a research cohort based on the VA’s Precision Oncology Program. Patients were given a choice to either have their data released internally within the VA, externally to other repositories, or not released at all.

Data Curation. Clinical data elements were selected from existing protocols and data sets specifically for lung cancer patients. Sources of data elements included ORIEN, MED-C, DoD, GDC, and APOLLO. The VA’s Corporate Data Warehouse was used as the primary source for clinical data. Imaging data consisted mainly of chest CT studies and were obtained from individual PACS systems throughout the VA network. Genomic FASTQ files that resulted from tumor sequencing were obtained directly from the vendors. All three data streams were linked using the patient’s ID.

De-Identification. Clinical data values contained no PII and dates were de-identified using the TCGA’s date obfuscation method. Software from TCIA was used to scrub all PII from the DICOM headers of images. The genomic FASTQ files contained no PII within its content and thus did not require additional de-identification.

Regulatory. Several regulatory requirements were satisfied to ensure proper protection of veterans’ privacy, including signing data use and transfer agreements with the NCI, subject matter expert certification and information security approvals for the de-identification processes, and risk analyses for using electronic submission portals.

Conclusion

Our experience in releasing VA patient data has shown that challenges are more often regulatory than technical. Our hope is that the increasing reliance on rich data sets for researchers to advance clinical care and the desire of patients themselves to share their own data will drive positive change in healthcare privacy policy and help establish protocols and workflows that facilitate the sharing of medical data.
Bladder cancer is among the most common forms of cancer, with nearly 17,000 deaths projected this year [1], and is the fourth leading cancer in men [2]. We developed a natural language processing (NLP) application that focused on (a) the severity of pathology findings and (b) the existence of finding mentions in pathology reports.

Methods

Using Department of Veterans Affairs (VA) electronic medical records, we identified 12,874 patients with bladder cancer diagnosed between 01/01/2005 – 12/31/2012 and created an initial set of 1,476,063 documents. Based on frequency of documents, we identified and filtered on surgical pathology reports created within 30 days before and up to 120 days following the cancer diagnosis that contained at least one of the following keywords: “ureteral”, “urethral”, or “bladder”. This resulted in 11,030 documents for 7,490 patients. 600 documents were then randomly selected and annotated by two urologists for 7 variables: histology, grade, carcinoma in situ (CIS), invasion type, invasion depth, muscle in specimen, and stage. The differences were adjudicated through a series of discussions and the assistance of a third urologist. We then split the annotated set of 600 documents into three parts: 300 (50%) of documents as a training set, 150 (25%) of documents as a development set, and 150 (25%) of documents were set aside for final validation.

Using the training and development sets of annotations, we built an NLP pipeline for the Apache Unstructured Information Management Architecture Asynchronous Scaleout (UIMA AS) [3], utilizing the libraries and tools contained in the Leo framework [4]. We implemented three modules in the pipeline: a regular expression module to detect mentions of each of the 7 variables, negation pattern matching, and term filtering to exclude negated text. Instances of the variables found in the text were aggregated to the most severe finding. For example, if a single document was found to have two specimens; one with no cancer and one with a histological urothelial classification, the output would have a single record for urothelial.

Results

After processing the test set of 150 documents, the hierarchical application performance was validated (Table 1). An error analysis of the final testing set showed a small number of random errors consisting of words and pattern combinations that were not previously seen in the training set and consequently, not captured in the testing set. There were no systematic errors that affected multiple separate documents.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Precision</th>
<th>Recall</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grade</td>
<td>0.957</td>
<td>0.918</td>
<td>0.933</td>
</tr>
<tr>
<td>Invasion</td>
<td>0.855</td>
<td>0.877</td>
<td>0.867</td>
</tr>
<tr>
<td>Depth</td>
<td>0.788</td>
<td>0.631</td>
<td>0.820</td>
</tr>
<tr>
<td>Histology</td>
<td>0.972</td>
<td>0.958</td>
<td>0.960</td>
</tr>
<tr>
<td>Stage</td>
<td>0.708</td>
<td>0.708</td>
<td>0.887</td>
</tr>
<tr>
<td>Muscle in Specimen</td>
<td>0.887</td>
<td>0.790</td>
<td>0.827</td>
</tr>
<tr>
<td>CIS</td>
<td>0.950</td>
<td>0.905</td>
<td>0.980</td>
</tr>
</tbody>
</table>

Discussion

NLP applications such as this, which focus on severity of cancer data in patient reports can collectively produce significant results. Categorical output may vary, according to precision, recall, and accuracy.

Table 1. Hierarchical values validation summary

References

Tochi C. Amadi, MPH1, Craig W. Johnson, PhD1

1The University of Texas School of Biomedical Informatics, Houston, Texas

Introduction
The increase in resistance of pathogens to antibiotic medication has become a major public health concern. Surveillance of changes in antibiotics resistance is imperative to inform clinicians on the appropriate use of existing medication for treatment, as well as guide the development of new antimicrobial agents1. Pseudomonas aeruginosa is an increasingly multi-drug resistant pathogen with a high rate of mortality in critically ill patients. Surveillance must be carried out to monitor any changes in susceptibility to common antibiotics2. Common medications for treating P. aeruginosa infections include gentamicin from the aminoglycoside antibiotics class, Imipenem, a carbapen antibiotic, and Ciprofloxacin from the quinolone class. The research question of the present study addressed whether changes in resistance of P. aeruginosa to these three antibiotic medications occurred during the period from 2006 to 2012 using laboratory data obtained from teaching hospital facilities.

Materials and Methods
We obtained laboratory test results from the Cerner Health Facts database collected from inpatient teaching hospitals. Sample sizes for the three antibiotics were: gentamicin = 512, Imipenem = 514, ciprofloxacin = 506. The study researched differences in resistance of Pseudomonas aeruginosa between 2006 and 2012. We used three hierarchical binomial logistic regression analyses to isolate effects uniquely associated with time on resistance, while controlling for gender. Time, coded as years, was treated as a categorical variable using 2006 as the reference category. We dichotomized resistance (susceptible vs. resistant). Resistance statuses originally coded as intermediate were classified as resistant. Gender used males as the reference category. To isolate and control for effects associated with gender, it was added in a first block. Years, the variable under investigation, was added in a second block. The three statistical analyses used SPSS version 24 (experimentwise $\alpha = .05$) and excluded year 2009 due to insufficient data.

Results and Discussion
The logistic regression analyses did not indicate any significant changes in resistance in the years following 2006 at the $\alpha = .05$ experimentwise error rate. The overall model was not significant for any of the three medications (Gentamicin: $\chi^2 = 6.97, p = .32$; Ciprofloxacin: $\chi^2 = 3.50 p = .74$; Imipenem: $\chi^2 = 4.04, p = .67$). Years did not significantly predict resistance for any of the individual medications as shown in table 1.

Table 1. Test Statistics for Antibiotic Medications: Gentamicin, Ciprofloxacin, Imipenem (2006 – 2012)

<table>
<thead>
<tr>
<th>Antimicrobial</th>
<th>Statistic</th>
<th>Gender</th>
<th>Year</th>
<th>2007</th>
<th>2008</th>
<th>2010</th>
<th>2011</th>
<th>2012</th>
<th>Constant</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gentamicin</td>
<td>Wald</td>
<td>0.04</td>
<td>6.969</td>
<td>0.093</td>
<td>1.82</td>
<td>1.793</td>
<td>4.722</td>
<td>1.433</td>
<td>7.372</td>
</tr>
<tr>
<td></td>
<td>p-value</td>
<td>0.841</td>
<td>0.223</td>
<td>0.76</td>
<td>0.177</td>
<td>0.181</td>
<td>0.03</td>
<td>0.231</td>
<td>0.007</td>
</tr>
<tr>
<td></td>
<td>Odds Ratio</td>
<td>0.953</td>
<td>0.872</td>
<td>0.506</td>
<td>0.573</td>
<td>0.379</td>
<td>0.561</td>
<td>0.364</td>
<td></td>
</tr>
<tr>
<td>Ciprofloxacin</td>
<td>Wald</td>
<td>0.412</td>
<td>3.109</td>
<td>0</td>
<td>0.003</td>
<td>0.071</td>
<td>1.275</td>
<td>0.119</td>
<td>3.152</td>
</tr>
<tr>
<td></td>
<td>p-value</td>
<td>0.521</td>
<td>0.683</td>
<td>0.994</td>
<td>0.959</td>
<td>0.79</td>
<td>0.259</td>
<td>0.73</td>
<td>0.076</td>
</tr>
<tr>
<td></td>
<td>Odds Ratio</td>
<td>0.881</td>
<td>1.003</td>
<td>1.023</td>
<td>0.903</td>
<td>0.638</td>
<td>0.861</td>
<td>0.533</td>
<td></td>
</tr>
<tr>
<td>Imipenem</td>
<td>Wald</td>
<td>0.236</td>
<td>3.73</td>
<td>0</td>
<td>0.841</td>
<td>0.272</td>
<td>0.176</td>
<td>0.188</td>
<td>11.108</td>
</tr>
<tr>
<td></td>
<td>p-value</td>
<td>0.627</td>
<td>0.589</td>
<td>0.993</td>
<td>0.359</td>
<td>0.602</td>
<td>0.675</td>
<td>0.664</td>
<td>0.001</td>
</tr>
<tr>
<td></td>
<td>Odds Ratio</td>
<td>0.899</td>
<td>0.996</td>
<td>1.57</td>
<td>1.258</td>
<td>0.825</td>
<td>1.235</td>
<td>0.256</td>
<td></td>
</tr>
</tbody>
</table>

*2006 is the reference category for year; Male is the reference category for gender

Conclusion
The study revealed no evidence that P. aeruginosa resistance to Gentamicin, Ciproflaxin, and Imipenem changed significantly between 2006 and 2012. Thus, it does not indicate an urgent need to replace or modify these medications. Further studies with other antibiotics and larger sample sizes can use logistic regression to analyze trends in resistance.

References
Harmonizing User-defined Phenotypic Variables using Latent Semantic Analysis (LSA) to Improve Data Discoverability in dbGaP

Liz Amos, MLIS, Satyajeet Raje, PhD, Masato Kimura, PhD, Preeti Gokal Kochar, PhD, Victoria Wilder, MLS, Ben Busby, PhD

U.S. National Library of Medicine, National Institutes of Health, Bethesda, MD, USA

Background
The NLM Database of Genotypes and Phenotypes (dbGaP) provides a rich source of studies with genotypes, phenotypes and associations between them1. As part of the submission process, dbGaP requires researchers to submit a data dictionary of study variables including their short descriptions. A recent study2 highlights the need to harmonize these user submitted variables to allow researchers to effectively find studies that share the same or similar variables in dbGaP. The study documents a manual effort to map variables in dbGaP to a standardized set of measures, viz. the PhenX toolkit. However, the rapidly expanding size of dbGaP variable dictionary (200,000+ variables by early 2017) renders sustaining such manual post-coordinated efforts infeasible. In this project, we attempt to (semi-)automate the harmonization of variables in dbGaP using the user-submitted variables.

Methods
A. Measuring conformity in original study. The original study provides PhenX IDs (by RTI International) and LOINC codes (by NCBI staff) for 20635 variables in dbGaP. To assess conformance within the original mapping, we perform a pairwise comparative analysis of the PhenX and LOINC mappings for all variables.

B. Clustering variables using LSA. We use Latent Semantic Analysis (LSA) to compute a pairwise similarity score between the 20635 variables in the original study (using the gensim python library3). We leverage the user submitted variable descriptions as textual input. We cluster together variables based on the calculated similarity score (threshold of 0.8 was empirically chosen). Thus, each variable has a list of “duplicates” identified by LSA. To validate the output of the LSA method, three evaluators manually reviewed 10% (~2000) randomly sampled duplicate variables independently. We also compare the results of our LSA based method with the existing PhenX and LOINC mappings.

Results / Discussion
A. Measuring conformity in original study. The results of the pairwise comparison indicates that the LOINC mapping has a more generalized grouping of variables compared to PhenX (610 unique LOINC codes vs. 1414 unique PhenX IDs). Moreover, LOINC codes conform completely to the PhenX mapping (i.e. all pairs with same PhenX ID have the same LOINC code). It should be noted that LOINC codes were missing for 3556 variables.

B. Clustering variables using LSA. Manual evaluation showed that the LSA method can identify “duplicate” variables with very high accuracy (96% precision when all 3 evaluators agree that a result is valid). The inter-rater agreement among evaluators was also high (kappa = 0.07). However, we observed poor conformance between the LSA derived clusters with existing PhenX ids and LOINC codes for some concepts. This means that the LSA is able to cluster variables that were not originally clustered by the LOINC or PhenX IDs. Further analysis revealed this is mainly due to discrepancies between semantic and syntactic characteristics. For instance, implicit synonymy among terms in a terminology (e.g. “sex” and “gender” have same LOINC code) cannot be clustered by LSA. Further refinement would be needed to account for such conditions.

Conclusion
Initial results using LSA for harmonization of variables in dbGaP shows promising results. We are currently exploring ways to extend our method to map to other standard terminologies, such as SNOMED CT. We are also studying the impact of using harmonized variables to the number and quality of cohorts retrieved from dbGaP.

References


Decoding the Behavior Change Techniques in Mobile Health Apps

Marcy G. Antonio, PhD (c), MPH, BSc1, Karen L. Courtney, PhD, MSN, BSN, BA1, and Judith T. Matthews, PhD, MPH, RN2

1 University of Victoria, Victoria, British Columbia, Canada; 2 University of Pittsburgh, Pittsburgh, Pennsylvania, United States

Background

It was estimated in 2016 that 165,000 health-related mobile apps would be downloaded 1.7 billion times [9]. Most of these apps are rarely used and have limited evidence of supporting and improving health outcomes [1-3]. Further, the majority of health apps have been developed with little attention to successful techniques for behavior change [2,3]. This study used the Behavior Change Technique Taxonomy (BCTT) developed by Michie et al. [4] to classify behavior change techniques employed in a sample set of mobile health apps, as a first step toward designing evidence-based health apps that are likely to influence users’ behaviors and generate positive health outcomes.

Aim

To develop a codebook based on the BCTT, and to apply this codebook in the identification of behavior change techniques in commercially available health apps.

Methods

The codebook is being developed in two phases. The completed first phase involved incorporating the 93 behavior change techniques (BCTs) defined by the BCTT into a coding framework specific to mobile health apps. The framework includes individual examples of health app features that match with each BCT code, as well as examples of mobile health app interventions that insufficiently align with specific codes. The ongoing second phase entails psychometric assessment of the inter-rater reliability of the codebook. Trained coders are independently identifying the BCTs in 10 randomly selected health apps that focus on chronic disease management. The identified BCTs are being examined for level of agreement to determine areas of the classification scheme that require clarification or revision. Through this iterative process coders will evaluate a series of apps until inter-rater agreement is substantial, as signified by a Kappa coefficient of at least .75.

Results

This poster will present the process for developing the coding framework, training coders, and evaluating the inter-rater reliability of the resulting codebook. It will also provide classification examples from the various mobile health apps, as well as identify their current strengths and limitations in incorporating BCTs.

Conclusion

The BCTT has been useful for identifying the critical techniques present within behavioral change interventions [9]. To date, the majority of the studies that have applied this taxonomy have focused on in-person interventions. This work applies the taxonomy to mobile health applications which offer unique opportunities for delivery of BCTs that may not be possible with in-person interventions. Classification of the behavior change techniques used in mobile health apps will aid in future research on effective mobile health app design aimed at changing behavior and improving health outcomes.

References


This project was supported by grant number R01HS022889 from the Agency for Healthcare Research and Quality. The content is solely the responsibility of the authors and does not necessarily represent the official views of the Agency for Healthcare Research and Quality.
Reference Range Number Lines Preferred by Latino Adults for Display of Clinical and Patient-Reported Outcome Data

Adriana Arcia, PhD, RN¹
¹Columbia University, New York, NY

Introduction
Tailored infographics can be used in patient- or consumer-facing informatics resources to support communication of clinical and patient-reported outcome data but additional research has been needed to identify optimal graphical formats for displaying various types of data. Reference range number lines are a graphical format in which reference ranges (e.g., blood pressure; scoring categories from a depression screening instrument) are marked off in labeled segments along a number line (e.g., Figure 1B). We worked with over 100 Latino adults with varying levels of health literacy. The first study was with residents of northern Manhattan using variables covering a variety of biometric and self-reported outcomes; the second was with family caregivers of dementia patients and focused on variables specific to dementia and caregiving. In each study, the aim was to develop tailorable infographic designs that would be preferred and perceived as easily comprehensible by the target audience. After the design phase, we programmed the designs into a novel system, Electronic Tailored Infographics for Community Engagement, Education, and Empowerment (EnTICE²), to automatically generate tailored infographics based on data provided.

Methods
We used similar methods across two studies. First, we identified multiple graphical formats (e.g., reference range number line, illustrated table) appropriate for the display of each variable. Next, we created one or more infographic designs for each format such that we prepared multiple designs for every variable. Then, we conducted participatory design sessions (N = 27) in both English and Spanish with small groups from the intended audiences to solicit preferences between designs and suggestions for their improvement. Infographic designs preferred by participants were iteratively refined between sessions until participants expressed satisfaction and had no further substantive changes.

Results
Participants across languages, target group, and literacy level consistently preferred designs that included reference range number lines over other graphical formats. In the first study, participants suggested combining the number lines for body mass index (BMI) and waist circumference with illustrations of body silhouettes. In the second study, refinements to designs that incorporated reference range number lines were focused on a) the imagery used to anchor the reference ranges, b) word choice for the text labels, and c) improvements to numerical scaling.

Conclusion
Reference range number lines are acceptable to and perceived as easily comprehensible by Latino adults with varying levels of health literacy. Although intended to meet the needs of a particular population, the strong preference for this graphical format across in-group differences suggests the format may have broad appeal. Therefore, it should be tested in additional groups to determine if it can be appropriately adopted as a standard patient-facing format for the display of values for which reference ranges apply.

Figures 1A & 1B. A score for the Kessler Psychological Distress instrument is displayed at left (1A) in an illustrated tabular format and at right (1B) in a reference range number line format (final version). Participants preferred the number line format and sun image to the tabular format and blue clouds /rainbow image. Score ranges were shifted from 6-30 to 1-25 because participants found the original scaling confusing.
Introduction
The Health Information Technology for Economic and Clinical Health (HITECH) Act earmarked $30 billion to incentivize the meaningful use of EHR and technology. One of the core pillars of the HITECH Act is engagement of patients and families in their health. The Centers for Medicare and Medicaid (CMS) meaningful-use requirement recommends that hospitals use health information technology (HIT) with features that will engage patients in their care. As hospitals work towards meeting these requirements, it has become clear that different hospitals are at different stages in their adoption. While some are hesitant and slow to adoption, others have adopted technologies with a wide range of patient engagement features. Therefore, it is important to investigate if a hospital’s financial standing determines the number of features or add-ons they include in their adopted information technologies. This study explores the relationship between hospital financial strength and breadth of patient engagement features offered through HIT.

Method
Data on hospital adoption of patient engagement HIT features were obtained from the AHA IT supplement, and financial data were obtained from the Centers for Medicare and Medicaid (CMS) cost report. Total margin and operating margin were used as measures of financial strength, and a summated scale of the identified patient engagement features (from 0 – 11) was created. Multinomial regression models with a longitudinal panel sample of 1,612 hospitals (2012 – 2013) were used. Only non-governmental acute care hospitals that provided information on the patient engagement features adopted for both years were included. Four categories of hospitals based on the breadth of patient engagement features adopted were identified: high adoption (9 – 11), medium adoption (5 – 8), low adoption (1 – 4), and no adoption. It was hypothesized that hospital financial strength will be positively and significantly associated with breadth of patient engagement features offered to patients (e.g., provide patients with an electronic copy of their discharge instructions and record, provide patient-specific education resources, and view and download their health/medical record). Variables identified from previous studies to be associated with innovation adoption were included in the model as control variables. They are: total number of beds, teaching status, system membership, ownership status, geographic location, competition, per capita income, Medicare discharges per inpatient days, average length of stay, occupancy rate, and FTE per 1000 inpatient day.

Results
From 2012 to 2013, the hospitals included in this study had adopted on average 5.08 out of a total of 11 patient engagement features. This increased from 4.62 features in 2012 to 5.54 features in 2013. Low adoption, medium adoption, and high adoption hospitals offered on average 3.11, 6.38, and 9.60 patient engagement HIT features respectively. This study further revealed that financial strength is important in determining the number of HIT features for patient engagement a hospital will adopt. Hospitals with high adoption had better total margins and operating margins compared to hospitals with no adoption (β= –0.088, p < 0.01) and (β= –0.066, p < 0.05), but only better total margins compared to hospitals with low adoption (β= –0.054, p < 0.01). There was however no significant difference in financial strength between high adoption and medium adoption hospitals.

Conclusion
The HITECH Act encourages the adoption of HIT in anticipation that it will help improve health care quality, reduce health disparities, and promote patient-centered delivery of medical care. This study presents a categorization of hospitals based on the number of patient engagement features offered through their adopted HIT. It also shows an overall positive trend in the number of features hospitals offer to engage patients. Notwithstanding, organizations will often weigh political pressures in light of their economic realities. This study suggests that even with the incentives provided to hospitals, administrators are obligated to consider their financial position in their strategic adoption of technology for patient engagement. While a hospital’s financial standing may not be the only consideration for administrators it is a top priority, and must be fully explored to reduce hindrances to achieving nationwide adoption.

References
Integrating a patient display into the primary care exam room to improve patient engagement

Onur Asan, PhD1, Paul Iglar, BS1, Jeanne Tyszka, MS,1 and, Rabia Copuroglu, BS1
1Center for Patient Care and Outcomes Research, Division of General Internal Medicine, Medical College of Wisconsin, Milwaukee, WI, USA

Abstract

This study was a mixed method interventional study to improve patient-centered communication and patient engagement with using electronic health records in primary care settings. We set up a double screen monitor in the exam rooms, with one being a patient specific display. We videotaped 24 visits and interviewed all patients as well as 8 physicians after the visits. The findings showed the potential of this approach to improve patient engagement, patient understanding of the data as well as transparency in the visit.

Introduction

The use of electronic health records (EHRs) in primary care has increased dramatically in the last decade. Increasingly, research in primary care focuses on how to use EHRs in a patient-centered way and design more patient centric EHRs in order to enhance patient-clinician communication and patient engagement during the visit. Several studies suggest that the potential pitfalls of EHRs can be avoided if physicians use EHRs as a communication and patient education tool during the visit, including sharing the screen with patients. Screen sharing might help to improve real-time patient-clinician communication; facilitate more accurate documentation, shared decision making, shared understanding, and patient involvement; it could also reduce the patient alienation that results when physicians focus on the computer screen. In this study, we had an intervention of a patient specific screen in the exam room (double screen). With the additional screen, patients were able to see all activities ongoing in the EHRs, and providers were able to share any information with the patient in much easier way. The goal of this study was to understand perception (including perceived usefulness and ease of use) and use of this patient-centered display, from both patients and providers in primary care environment.

Methods

We conducted a mixed method study to understand perceptions of providers and patients as well as their interaction with the double screen monitor in the visit. We videotaped 24 visits and using two videos from different angle. We also interviewed all 24 patients as well as 8 physicians who participated to the study. Purposive convenience sampling method was followed to recruit providers. We used a special tool called Noldus-Observer to code all videos to quantify interaction in the visit. We also used content analysis to identify themes in the interview.

Results

The video analysis showed duration of patients engaged with the second screen as well as total interaction time of doctors with EHR and patient. It was also interesting to observe that patients mostly looked at their screen while provider is typing to see what is typed in. The qualitative piece of the data showed that there is potential of having second screen to improve patient engagement, medication reconciliation, patient understanding of the data and patient trust in providers. Most of the patients liked the 2nd screen and felt more empowered since they could see every EHR related action of providers. Some providers felt that it improved number of questions prompted by patients. It was also reported that providers used patient screen to educate some of the patients with visuals and it was well- received. Most patients felt the second screen was useful, however almost half of the providers did not feel in the same way. We also identified some improvement points. Since the primary EHR screen is clutter, patient screen was also clutter. The size of the monitors are large, so it created some barriers given the desk and layout of the room was small. The future designs should make it more simplified and less clutter and more patient centric. Overall, patients acknowledged their pleasure to have access to their own data and transparency in the visit compared to visits before.

Discussion/Conclusion

More collaborative technologies in health care settings have great potential to support patient-centered care and communication by providing accessible, and actionable health information. The findings of this exploratory study can guide future exam room patient centred EHR redesign and intervention studies.
Healthcare Claims Data - Challenges in Acquisition and Analysis

Bhanu Bahl, Ph.D¹, Daniel Runt, MSMI¹, Nealia Khan, MPH¹, Douglas Macfadden, MS¹

¹The Harvard Clinical and Translational Science Center, Harvard Medical School, Boston, Massachusetts

Introduction

The Prevention and Wellness Trust Fund (PWTF) evaluation is a project established by the Massachusetts Department of Public Health (MDPH) with Harvard Catalyst team to estimate the impact of a variety of evidence-based interventions on both clinical and cost outcomes for the four principal conditions: pediatric asthma, hypertension, tobacco cessation, and falls in older adults. In this regard, the HC informatics team built a HIPAA compliant technical infrastructure for collecting anonymized patient data from a variety of sources, including the Massachusetts All-Payer Claims Database (APCD), Case Mix Index, death registry, and electronic health records through MDPHNet. In this poster we will illustrate the process and challenges encountered in the healthcare claims data capture, data cleaning, data validation, and overall data processing. Several policies and procedures regarding the sharing, transmission, storage, and distribution of data files were put in place including HIPAA complaint data security measures.

Methods

Implemented data governance procedures for secure data transfer, processing, and use as documented in the Data Use Agreement. The research data was treated as Level 4, the University’s second highest security classification, sufficient for identifiable human research. Rules for data validation and cleaning were created. Several SAS procedures were developed for converting raw data files into analysis data sets which included, but were not limited to, derived data fields, handling missing data, data linking, and data aggregation. All the coding scripts were kept within a Git repository. Rules and processes were further optimized after feedback from the statisticians.

Challenges

The delivery of the data was a raw text data dump, with processing to be done using the SAS¹ software, version 9.4. The objective was to locate individual datum in a large dataset, verify their quality, and decide to keep, derive, or discard each element. The size of the datasets, which was several terabytes, as well as the remote configuration – and multiple users accessing the same dataset – each presented performance issues. The rules for grouping ICD 9 codes to define a condition of interest, e.g. elderly falls, posed a challenge in the variant data.

Results and Conclusion

Solutions that lead to efficient processes for data extraction and handling were: well-defined data requirements, creation of small analysis-driven datasets, multiple iterations of data pulls during the streamlining process, and the creation of subsets of data. This streamlined the creation of analysis-specific files which has multiple benefits:

1. Removing end-users from the process of data cleaning and processing, resulted in high confidence analytical datasets, that were easy to use by statisticians
2. Employing this iterative process helped quantify the ratio of data used vs. data received, thereby reducing time spent on the challenge of finding multiple needles in a large haystack
3. Reducing the file size by removing redundant data, reduced computational burden and wait time for users.
4. Implementing the 'small data' approach, resulted in improvement in the quality of data, efficiency in process, and better reproducibility.
5. Optimized Algorithm for data sub-setting, cleaning, merging, and validation
Local Causal Networks Discover Predictive Cytokine Biomarkers of Scleroderma

Ali S. Bakhtiari¹ PhD, Galina S. Bogatkevich² M.D., Ph.D., Alexander V. Alekseyenko¹,³ PhD

¹ Biomedical Informatics Center, Department for Public Health Sciences, ² Division of Rheumatology and Immunology, Department of Medicine, ³ Department of Oral Health Sciences, Medical University of South Carolina, Charleston, South Carolina.

Introduction

Scleroderma is an autoimmune disease with established relationship between immune cytokines and its prognosis (1). The task of linking the cytokines to scleroderma diagnostically is important in simplifying detection of the disease. The purpose of this work is to discover cytokine biomarkers of scleroderma using local causal neighborhood (LCN) learning techniques. LCN methods identify biomarkers in the causal vicinity of the target variable (e.g. scleroderma status), which under broad assumptions results in most parsimonious and highly predictive models.

Materials and Methods

Our study cohort consists of 40 cases of positive scleroderma patients and 24 healthy controls. The study has been approved by Medical University of South Carolina (MUSC) IRB. We measured 29 bronchoalveolar lavage fluid (BAL) cytokines in each of the study subjects. We have log-transformed the raw cytokine measurements and imputed the missing values from their 10 nearest neighbors identified by K-nearest neighbors (KNN) algorithm. Adjustments for sex and race have been performed by fitting a linear regression model to individual cytokine models and extracting the residuals. We have performed univariate Welch t-tests to find the cytokines significantly associated with scleroderma status. Next, we have used the HITON-PC algorithm (2) from causal explorer toolbox (3) to identify the local causal network (LCN) of the scleroderma status. In order to assess the stability of the LCN estimate, we performed 1,000 iterations of cross validation. In each iteration, we have randomly selected 70% of the dataset and obtained the LCN biomarkers identified by HITON-PC. The remaining 30% of the data has been withheld for testing. The frequency of the appearance of the inferred cytokines from the whole dataset in the individual subsampled runs, acts as a metric for the stability of the model. The same 1,000 iterations have been used for predictive analysis. We used the LCN biomarkers as predictors for the logistic regression model of scleroderma status. The performance of each model has been assessed in the testing set using the area under the curve (AUC) of the receiver operating characteristic (ROC) curve.

Results

From the 29 studied cytokines, 8 are significantly associated with scleroderma status (P<0.05). After adjusting for multiple comparison using false discovery rate (FDR) method (4), 3 cytokines remained significant (FDR=0.05). HITON-PC algorithm has identified 2 cytokines to be in the LCN of the scleroderma status. Cross validation analysis of the HITON-PC algorithm have shown that the 2 inferred cytokines are frequently present in the LCN of the subsampled dataset as well. The predictive performance of the LCN biomarker under logistic regression model is estimated to achieve an average classification success rate of 68.46%, mean sensitivity of 65.49% and mean specificity of 73.44%. AUC of ROC curve is 73%.

Conclusion

In this work, we have demonstrated the use of LCN for discovering cytokines biomarkers of scleroderma status. We have shown the merit of using the LCN, by rigorously testing the model’s predictive power.

References

Patient Generated Health Data Governance at a Cancer Center

Erdeta Bani, MPA\textsuperscript{1}, Rimma Belenkaya, MS, MA\textsuperscript{1}, Olga Strachna, MS\textsuperscript{1},
Daniel Stein MD, PhD\textsuperscript{1}, Peter Stetson, MD, MA\textsuperscript{1}

\textsuperscript{1}Division of Health Informatics, Memorial Sloan Kettering Cancer Center, New York, NY

Introduction
As the number of patient data collection instruments grows, and the sets of questions asked to patients proliferate\textsuperscript{1}, the need for governance becomes critical. ONC has only just begun developing a policy framework to identify best practices, gaps and opportunities for implementation and integration of Patient Generated Health Data (PGHD) into clinical workflows. Internally, health care organizations do not normally have governance around the process for PGHD implementation and integration\textsuperscript{2}. Practical guidelines on how to optimally standardize and systemically govern integration of PGHD into clinical workflow/clinical systems remain a knowledge gap.

Methods
We have implemented a PGHD Governance Program at Memorial Sloan Kettering (MSK) to support the systematic collection and analysis of patient generated data via electronic questionnaires. Governance of the PGHD integration requires collaboration of various stakeholders\textsuperscript{2}. At MSK, PGHD Governance is represented by two interconnected entities: PRO (patient reported outcomes) Working Group and eForms Committee. The PRO Working Group is represented by clinicians, researchers, epidemiologists, biostatisticians and informaticians. The eForms Committee is a larger multidisciplinary taskforce comprised of SMEs in Informatics, Patient & Caregiver Engagement, Nursing Operations, Health Information Management, Biostatistics, Mobile Application Development, and representatives from the PRO Working Group. The role of PRO Working Group is to create a conceptual framework for PRO integration. The role of eForms Committee is to oversee all aspects of implementation of the conceptual framework and apply these principles beyond PRO questionnaires to all PGHD. Thus the eForms Committee ensures standardization of questions/question sets across instruments; optimal patient experience with questionnaires that do not introduce survey fatigue; appropriate secondary uses of PGHD; and monitoring usage and value of electronic instruments over time. We include representatives of the patient's care team in all aspects of PGHD integration.

Results
Our PGHD Governance was launched in January 2016. The eForms Committee developed a workflow for receiving and triaging all PGHD requests. It oversaw enforcement of this workflow, meeting monthly to address new requests, and to evaluate the effectiveness and impact of the PGHD on clinical practice while ensuring data integrity in the legal record. In its first year, the PRO Working Group identified goals for collecting PROs; chose appropriate instruments and adjusted them to serve specific goals and populations; established patient cohort parameters; determined setting and timing for administering assessments; chose scoring methodology; designed reports and analytical methods to interpret results; determined thresholds for care team notifications and escalation rules; and evaluated impact of the PRO intervention on the practice. To date, the collaboration between these two entities has standardized PRO-CTCAE, exercise, smoking and pain question sets in a number of instruments used with new, post-operative, palliative, and survivorship patients. It used analytic dashboards to evaluate use and value of questionnaires such as EORTC-QLQ30. It has converted more than 80 paper questionnaires to electronic ones applying these implementation principles.

Conclusion
Although we are at a very early stage of the PGHD Governance implementation, our approach has already improved our process of electronic PGHD instrument adoption. Including representatives of care teams in all aspects of the governance helped our clinicians understand information collection bottlenecks, improve their workflows and communication with patients and evaluate the impact of the PGHD intervention on the clinical practice. As our PGHD governance program matures, we expect our approach to PGHD governance can be applied more broadly to healthcare institutions, beyond cancer centers, that are looking to tackle this challenge.

References:
1. Vande Loo, SJ. and North, F. BMC Medical Informatics and Decision Making (2016) 16:41
The Time is Right for Clinical Data Platforms that Reduce Clinician Time Reviewing Patient Care Data Elements: Emerge as a proof of concept

Denise Barchas RN MSN1, Angela Lipshutz MD MPH1, Kathleen Liu MD PhD1, Kathleen Turner RN1, Charlotte Garwood RN MS2, Kelly Timothy RN MSN1, Amie Lencioni RN1, Michael Gropper MD PhD1, and Hildy Schell-Chaple RN PhD1

1University of California San Francisco, SF, CA; 2Vanderbilt University, Nashville, TN

Introduction

The electronic health record (EHR) is intended to provide clinicians with secure, instant, and current patient-specific clinical data. Despite this intent, EHR usability and data display is suboptimal and may lead to time wasted for clinicians using the EHR, ultimately affecting patient care delivery1. The purpose of this study was to evaluate the time spent by intensive care unit (ICU) clinicians to locate specific clinical harm prevention intervention elements for venous thromboembolism events (VTE), delirium, ventilator associated events (VAE), central line-associated blood stream infections (CLABSI), ICU-acquired weakness (IAW), and care discordant with goals (GOC). Clinicians were asked to locate and review data in the EHR as if they were performing routine chart review and compare it to time spent reviewing the innovative Emerge platform. The Emerge platform transforms EHR data elements to a separate harms monitor that summarizes essential patient data in a graphical fashion (Figure 1).

Methods

Clinicians were recruited from the adult medical-surgical ICU at a large academic medical center. Participants were asked to locate specific harm prevention care elements in the EHR (EPIC Verona, WI) and also the Emerge platform for a single patient. The primary measure was time spent to review care interventions by harm type and total time for review of all 23 data elements. A Wilcoxon signed-rank test was used to analyze the difference between time spent in the EHR and Emerge platform. An alpha level of less than .05 was considered statistically significant. Analyses were performed using SPSS statistical software, version 23 (SPSS, Inc).

Results

Based on power calculations to detect clinically relevant differences, 23 ICU nurses, nurse practitioners, and fellow physicians participated in this study. There was a significant reduction in total time spent reviewing care elements using the Emerge platform compared to the EHR (Median (IQR) in minutes, 5:00 (4:17, 6:51) vs. 8:41 (6:30, 9:37), p < .001. Less time was spent reviewing all harm data using Emerge compared to the EHR, and significant reductions were found for review of VTE, VAE, and GOC element (Table 1).

Figure 1. Emerge platform Table 1. Time difference between EHR and Emerge (minutes:seconds)

<table>
<thead>
<tr>
<th>Data Elements by Harm</th>
<th>EHR Median (IQR)</th>
<th>Emerge Median (IQR)</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>VTE</td>
<td>1:13 (0:57, 1:38)</td>
<td>0:56 (0:35, 1:19)</td>
<td>.02</td>
</tr>
<tr>
<td>Delirium</td>
<td>1:09 (0:57, 1:37)</td>
<td>0:47 (0:32, 1:04)</td>
<td>.07</td>
</tr>
<tr>
<td>VAE</td>
<td>2:46 (2:17, 3:31)</td>
<td>1:38 (1:22, 2:03)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>CLABSI</td>
<td>0:16 (0:15, 0:23)</td>
<td>0:15 (0:11, 0:19)</td>
<td>.11</td>
</tr>
<tr>
<td>GOC</td>
<td>1:15 (0:56, 1:43)</td>
<td>0:24 (0:17, 0:34)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>IAW</td>
<td>1:00 (0:50, 1:29)</td>
<td>0:57 (0:36, 1:06)</td>
<td>.12</td>
</tr>
<tr>
<td>Total Time</td>
<td>8:41 (6:30, 9:37)</td>
<td>5:00 (4:17, 6:51)</td>
<td>&lt;.001</td>
</tr>
</tbody>
</table>

Conclusion

Well-designed innovations that transform data from the EHR and provide an intuitive visual display can reduce clinician time spent reviewing patient data in the EHR to ensure the plan of care is implemented. The time saved for a single patient multiplied by the typical patient load of 16 ICU patients at our hospital equals 56 minutes. The Emerge platform is a model for innovative platforms that can improve efficiency, usability, reduce time facing the computer, and potentially impact safety and quality of care.

Reference


1942
Distributed Health Data Networks: Extending the PopMedNet\textsuperscript{TM} Informatics Infrastructure to Extend Existing Query Functionality to New Data Sources

Kimberly Barrett, MPH\textsuperscript{1}, Jessica M. Malenfant, MPH\textsuperscript{1}, Kyle Erickson\textsuperscript{1} Adam Paczuski\textsuperscript{1}, Zachary Wyner\textsuperscript{1}, MPH, Chayim Herzig-Marx, PhD\textsuperscript{1}, Jeffrey S. Brown, PhD\textsuperscript{1}
\textsuperscript{1}Harvard Pilgrim Health Care Institute, Boston, MA

Introduction: Distributed health data networks are increasingly used to conduct clinical and observational research, medical product safety surveillance and to supplement existing public health surveillance activities. Several of these distributed networks utilize the open source platform, PopMedNet\textsuperscript{TM} (PMN), to facilitate research of electronic health data. PopMedNet (PMN) enables distributed analyses through a flexible infrastructure that allows for various methods of querying data such as, modular programs, file distributions, and menu-driven queries (MDQs). Not only does PopMedNet streamline querying data, but it is compatible with multiple distributed networks- The FDA’s Sentinel system, the NIH Health Care Systems Research Collaboratory, PCORI’s PCORnet, National Cancer Institute’s CRNnet, and the Massachusetts Department of Public Health’s MDPHnet are active PMN-based networks. They have different common data models (CDMs), goals, and priorities, but all benefit from the functionality provided by PMN and related software applications.

Methods: This presentation describes the integration of multiple tools to validate that the PMN distributed querying infrastructure can be used for non-PMN data sources.

Menu-Driven Query: The Menu-driven Query (MDQ) query composer interface was developed to support “query terms” that can be shared across networks. The system supports querying filters (referred to as terms) and result grouping options (referred to as stratifications). The MDQ interface is Common Data Model aware, allowing a user to see only the terms and associated value sets applicable to the Common Data Model they are querying. While this is an assumed ability when an investigator is querying within a single network, this change helps extend the collaboration opportunities across networks with different data models.

After distribution, queries are transmitted to the PopMedNet (PMN) DataMart Client in a JSON string, that is then parsed by the PMN model adaptor for that user’s Common Data Model and compiled into a LINQ expression tree that Microsoft Entity Framework 6 translates into SQL queries appropriate for the target database. This process enables data holders to keep data in the relational database management system (RDBMS) of their choice, reducing barriers to participation. With this improvement, it is no longer necessary to develop custom SQL code for each RDBMS. As previously described, the Rapid Data Characterization Tool (RDAT) initially developed for the Sentinel network in PMN was enhanced to allow querying of new terms for age, sex, height and weight distributions which will benefit PCORnet and others. This tool continues to be used by multiple networks. It uses a data characterization data model that is agnostic to the Common Data Models of the individual networks.

Test Case Inserter: As the capabilities of the MDQs expand, a method of verifying the function and analysis is necessary. The Test Case Inserter tool, a purpose-built command line program, enables users to easily insert test data into a relational database management system (RDBMS), such as Oracle or SQL Server without requiring knowledge of SQL programming syntax for each RDBMS. The test data is formatted according to a Common Data Model, and the Test Case Inserter tool then “inserts” all of the test information into a RDBMS. The tool can be configured to read any data model so long as the .csv adheres to a standard format, originally based on the PCORnet data model. Once the data is inserted into the RDBMS, the researcher can run a MDQ against the data. Users are able to the run a PMN query and validate that the results match what is expected according to what they added via the TCI tool.

Results: We can demonstrate that the TCI can be used to accurately update databases that adhere to existing common data models, such as PCORnet as well as ad hoc databases. This process allows us to validate that the new PopMedNet MDQ’s are accurate and consistent across different database flavors and the functionality can be repurposed for multiple data models. One challenge discovered that should be addressed is the requirement to modify the table and column format from all caps to lowercase depending on the workflow step. For TCI, the target database needs to be all caps, but for the PMN queries, the target database needs to be lowercase.

Conclusion: Ongoing and future work will involve using additional data models that have fields that already exist as queryable terms in the menu-driven query interface and explore how we can continue to extend this functionality. Distributed data networks have proven valuable in conducting public health surveillance, research, and evaluation activities. This work has shown how PMN and allied tools can be extended and scaled for use in various work streams and across networks, regardless of data model or technical environments.
Comparative Analysis of Intravenous to Oral Transitions at Two Tertiary Care Hospitals in the U.S. and Switzerland

Patrick E. Beeler MD\textsuperscript{1,2,3}, David W. Bates MD MSc\textsuperscript{1,2}, Jürg Blaser PhD\textsuperscript{3}

\textsuperscript{1}General Internal Medicine, Brigham and Women’s Hospital, Boston, MA, USA; \textsuperscript{2}Harvard Medical School, Boston, MA, USA; \textsuperscript{3}Research Center for Medical Informatics, University Hospital Zurich, Switzerland

Introduction: Early intravenous (IV) to oral (PO) transitions increase the quality of care and decrease costs. Automated reminders successfully promoted earlier IV-to-PO transitions of antibiotics.\textsuperscript{1} Yet, no comprehensive data on IV-to-PO transitions across all drug classes have been published, although such data could inform on opportunities of reasonable transitions. Our objective was to generate knowledge on IV-to-PO transitions performed at two large, comparable hospitals, one in the U.S. and one in Europe, thereby increasing the generalizability and meaning of the results.

Methods: Data were extracted from the clinical data warehouses at the Brigham and Women’s Hospital (BWH; Boston, MA, USA; preliminary data 1/2014-4/2014) and at the University Hospital Zurich (UHZ; Switzerland; 8/2009-4/2014). These leading tertiary care academic medical centers are particularly suitable for comparative studies due to their similar size (859 and 850 beds) and Case Mix Index of 1.47 and 1.54, respectively. Data were processed using SQL. Drugs and free text orders were mapped to the Anatomical Therapeutic Chemical (ATC) Classification System (WHO, Geneva, Switzerland) ensuring international comparability. At the end of an IV therapy the time frame ±36 hours was beginning PO administrations of the same drug agent. This approach allowed for calculating the interquartile range (IQR): half of PO drugs started within -0.5 h to +3.0 h (UHZ) and 0 h to +7.5 h (BWH) after IV discontinuation. By using these IQRs we could identify IV-to-PO transitions among identical and different drug agents. Aggregated lists of IV-to-PO transitions were reviewed to sort out unreasonable drug agent sequences.

Results: Analgesics accounted for the vast majority of IV-to-PO transitions (Table 1). The most frequent transition of an antimicrobial at the University Hospital Zurich was IV amoxicillin/clavulanic acid → PO amoxicillin/clavulanic acid (n=3,844), at the Brigham and Women’s Hospital IV metronidazole → PO metronidazole (n=315).

Table 1. Most frequent IV-to-PO transitions at the two tertiary care hospitals in the U.S. and Switzerland

<table>
<thead>
<tr>
<th>IV drug</th>
<th>PO drug</th>
<th>Count</th>
<th>IV drug</th>
<th>PO drug</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hydromorphone</td>
<td>Oxycodeone</td>
<td>2,000</td>
<td>Acetaminophen</td>
<td>Acetaminophen</td>
<td>16,416</td>
</tr>
<tr>
<td>Hydromorphone</td>
<td>Acetaminophen</td>
<td>1,498</td>
<td>Dipyrone</td>
<td>Dipyrone</td>
<td>12,807</td>
</tr>
<tr>
<td>Hydromorphone</td>
<td>Hydromorphone</td>
<td>1,036</td>
<td>Amoxicillin/clavulanic acid</td>
<td>Amoxicillin/clav. acid</td>
<td>3,844</td>
</tr>
<tr>
<td>Metoprolol</td>
<td>Metoprolol</td>
<td>646</td>
<td>Esomeprazole</td>
<td>Esomeprazole</td>
<td>3,837</td>
</tr>
<tr>
<td>Pantoprazole</td>
<td>Omeprazole</td>
<td>645</td>
<td>Dipyrone</td>
<td>Acetaminophen</td>
<td>3,746</td>
</tr>
<tr>
<td>Famotidine</td>
<td>Famotidine</td>
<td>557</td>
<td>Potassium</td>
<td>Potassium</td>
<td>2,623</td>
</tr>
<tr>
<td>Morphine</td>
<td>Oxycodeone</td>
<td>455</td>
<td>Acetaminophen</td>
<td>Dipyrone</td>
<td>2,536</td>
</tr>
<tr>
<td>Lorazepam</td>
<td>Lorazepam</td>
<td>435</td>
<td>Heparin</td>
<td>Phenprocoumon</td>
<td>2,093</td>
</tr>
<tr>
<td>Hydromorphone</td>
<td>Acetylsalicylic acid</td>
<td>406</td>
<td>Furosemide</td>
<td>Torasemide</td>
<td>1,891</td>
</tr>
<tr>
<td>Morphine</td>
<td>Acetaminophen</td>
<td>403</td>
<td>Diclofenac</td>
<td>Diclofenac</td>
<td>628</td>
</tr>
</tbody>
</table>

Discussion: We found major distribution differences of IV-to-PO transitions. No opioid analgesics were among the top transitions at the UHZ, contrasting with the BWH where 6 of the transitions involved opioids. Regarding the opioid epidemic, our study shows ways in IV-to-PO transitions of analgesics towards non-opioids, and early transitions may be supported by automated reminders. Dipyrone’s approval in the U.S. was withdrawn in 1977 due to its association with agranulocytosis, however, this potent analgesic is frequently used at the UHZ. Considering the mortality related to the opioid epidemic, one might argue for a reevaluation of dipyrones applicability in the U.S. In conclusion, comparative analyses of medication orders revealed surprising cultural and pharmacoepidemiological differences.

A Simple and Efficient Method for the Management of Multiple Electronic Health Record-Driven Phenotype Projects

Cosmin A. Bejan, PhD¹, Joshua C. Denny, MD, MS¹²

¹Department of Biomedical Informatics, ²Department of Medicine, Vanderbilt University Medical Center, Nashville, TN

Abstract
Based on our experience of working with the Electronic Health Record (EHR) at Vanderbilt University Medical Center (VUMC), we developed a simple method of good scientific practice for the management of multiple phenotype projects. Our method relies on the Google Sheets application (https://docs.google.com/spreadsheets/) that allows for cooperative editing of documents by multiple investigators at the same time. Our framework proved to be efficient when working with multiple teams of investigators from both inside and outside VUMC.

Background
Researchers in biomedical informatics are often involved in multiple EHR-based phenotype projects that may happen simultaneously or may be in different phases of their development life cycle. Each of such phenotype projects involves one or multiple investigators, requires the design and implementation of one or multiple phenotype algorithms, and usually results in executing dozens of experiments with various configuration parameters.

Method
The office suite made freely available by Google as both web and mobile applications has quickly become popular for its simplicity. These applications, including Google Docs, Sheets, and Slides, allow for an efficient collaboration where users can simultaneously edit documents, can see character-by-character document changes by other users in real time, can track the past edits, and are able to revert the documents through a revision history mechanism. Out of these applications, we adapted Google Sheets to create a practical framework for the management of our phenotype projects. In general, for each project, we created a spreadsheet that we shared with all the investigators participating in the project. A standard document contains one main sheet with all the experiments performed for the corresponding project and multiple auxiliary sheets with information used in these experiments. Figure 1 depicts one of our projects that involves sensitive time searches for the discovery of rare adverse events associated with various medications.[1] For instance, to identify patients who experienced severe allergic reactions such as Stevens-Johnson Syndrome and Toxic Epidermal Necrolysis (SJS/TEN) caused by phenytoin, an investigator added the medication in ‘MED lists’ and SJS/TEN billing codes in ‘ICD lists’. As shown in Figure 1, time constraints for filtering patients with adverse drug reactions occurring after a medication are specified in the main sheet. The configuration of each experiment is automatically processed, the corresponding phenotype algorithm is executed, and the extracted patient cohort is returned to the investigator for further analysis. Additional information sheets integrated in other projects include: 1) annotation guidelines and relevant examples for textual information extraction projects; 2) the evaluation of phenotype algorithms; and 3) SQL queries for cohort size estimations.

Conclusion
We presented a simple and easy to implement method that proved to be efficient for the management of intensive, collaborative phenotype projects. The type of phenotype projects tested with this method ranges from applications that combine structural clinical information to projects mainly based on natural language processing technologies.

References
Implementation of a Knowledge Base Driven Perioperative Metrics Framework and Governance Program

Rimma Belenkaya, MS, MA1, Olga Strachna, MS1, Christopher Stromblad, PhD1, Kinjal Shah, MPH1, Hyok-Hee Yoo1, Natalia Shapiro1, Alexander Tanchoco1, Marina Kerpelev1, Ilya Sher1, Jonathan Wills1, Shai Ahmed1, Peter Stetson, MD1, Daniel Stein, MD, PhD1.

1Division of Health Informatics, Memorial Sloan Kettering, New York, NY.

Introduction

We have implemented a knowledge base driven enterprise framework for perioperative metrics, developed at Memorial Sloan Kettering to support operational analytics and research in the perioperative domain.

Consistent generation of key perioperative metrics is challenging, due to the lack of consistent definitions, redundant reports produced by multiple teams, multiple underlying data sources and lack of transparency in metric derivations. The framework we developed covers the entire metric life cycle and ensures consistency of metrics generation and reporting across the enterprise, and increases efficiency in providing stakeholders the data and insights they need.

Methods

We created a framework that includes governance, knowledge base, an integrated harmonized data repository, and data visualization components. The governance body oversees all aspects of the metric development from definition and certification, to datamart design and visualization creation, and it addresses any changes in the underlying data or business processes that may affect the metric.

The knowledge base is made up of two components: A Metric Ontology and a Report & Dashboard Catalog. The Metric Ontology captures detailed metric definitions, classification, relationships between the metrics, and links the underlying data sources and metadata. The Report & Dashboard Catalog contains all the perioperative reports, annotated with rich set of attributes. The knowledge base associates metrics with reports. It is based on a semantic extension of MediaWiki utilizing a set of nested templates and forms to collect and display metric attributes. It features ontology-driven search interface and tools for evaluation of report usability.

The integrated data repository built as a star schema, is a single source of data for derivation of the certified metrics. Its dimensions contain metadata defined in the knowledge base. ETL (extract transform and load) processes that populate the datamart are vetted to accurately harmonize data from multiple sources.

Visualization of the data is the ultimate purpose of the framework, helping the stakeholders to answer key questions about perioperative outcomes, quality of care, operations, and efficiency. Our dashboard design process follows human-centered design principles. The dashboard elements are driven by the metadata defined in the knowledge base. These visual elements also link back to the metric definitions completing the framework loop.

Results

Our framework was launched in July 2016. Since then, the knowledge base was populated with 90 perioperative metrics and over 400 perioperative reports. The top priority productivity and efficiency metrics have been revised, defined, and certified. During this process, new metric variants were introduced that enabled different perspectives on perioperative productivity and efficiency. The first version of the data repository was built and populated. Executive and Ambulatory Case Volume dashboards were implemented, and Scheduling Accuracy dashboard is being designed. We are integrating these instruments into the workflows of the key stakeholders. The next critical metrics, utilization and quality and outcome measures, are underway.

Discussion

Our framework has enabled our Department of Surgery to capitalize on existing informatics and analytic resources to support clinical operations and research. The process of metric development is complex and requires appropriate planning, given the need to coordinate a diverse group of stakeholders and to achieve consensus. Also, while the framework reorganizes existing reports, it does not address problems associated with new data and report requests. We plan to address these limitations next by adopting a Semantic Web platform that supports formalized ontology and collaborative process of metric development and implementing process of triaging new data and report requests.
Honoring Patient’s Data Sharing Preferences: Implementation Challenges

Elizabeth Bell, MPH; Diana Guijarro, BA; Imho Jang; Wenhong Zhu, PhD; Tyler Bath, BS; Gwang-No Yun; Masud Rahman; Chao Jiang, MS; Xiaqian Jiang, PhD; Lucila Ohno-Machado, MD, MBA, PhD; Hyoneui Kim, RN, MPH, PhD

Department of Biomedical Informatics, University of California, San Diego, La Jolla, CA

Introduction and Background

iCONCUR (informed CONsent for clinical data and biosample Use for Research) is a web-based consent management system that allows patients to indicate their data sharing preferences in a tiered manner (https://ctriconcur.ucsd.edu). We are currently revising iCONCUR based on the lessons learned from a pilot study and will test it with a larger sample of patients recruited from the two academic health systems. During this study period, the sharing preferences of participating patients will be honored – i.e., data that they do not wish to share will be excluded from the datasets prepared for researchers. The first step to implementing the process of honoring patient’s data sharing preference is to establish mapping between the data items in the data sharing preference taxonomy and the location of the Clinical Data Warehouse (CDW) where the data items are stored. Based on this mapping, the data that patients do not wish to share is automatically filtered out when a clinical dataset is prepared for a researcher (Fig 1). UCSD Medical Center uses Epic EHR system. Clarity is the database of Epic EHR. Research data has been prepared from the CDW that has the same structure of Clarity (i.e., its copy).

Methods

The data sharing preference taxonomy covers with 10 content areas such as socio-demographic, living environment and life style, medical history, family history, medications, test results, procedures, clinical encounters, vital signs and assessment, and sexual and pregnancy history. We first identified where the information related to the data items listed in the data sharing preference taxonomy is documented in Epic EHR. Then we tracked how those data fields of Epic are ETL-ed (Extract-Transfer-Load) into the CDW to pinpoint the data tables and columns needed to link back to the data sharing preference taxonomy.

Results

Nine content areas were mapped to the fields of the EHR and subsequently to those in the CDW. The content area not fully mapped is living environment and lifestyle, which includes drinking, smoking, and recreational drug use histories, as well as diet habit, exercise and activity, stress level, and social engagement. The former 3 items were fully mapped as they were documented in a structured format. However, the latter 4 were left unmapped, as they do not have dedicated structured fields and are mostly documented in free-text notes.

Discussion

We noted a few complex cases where a single data item from the sharing preference taxonomy is mapped to multiple tables and fields in the CDW. For example, medical history item is linked to three tables (problem list, visit diagnosis, medical history) in the CDW, and pregnancy history is to two tables. Currently, the CDW at UCSD is transitioning into a standardized structure following the Observational Medical Outcome Partnership Common Data (OMOP-CDM) Model. This new implementation will streamline the process of honoring patient’s data sharing preferences by simplifying the mapping between the sharing preference taxonomy and the CDW. For example, in the new OMOP-CDM based CDW, medical history is stored in the single table “Condition Occurrence.” An additional challenge that we identified is handling inter-relationships between medical history and the other types of related information (e.g., test results, medication, and therapeutic procedures) that could be used to infer the medical history even if medical history is not shared. We plan to build a smart honoring system that recognizes such relationships and filters out this related information, even if a patient forgets to indicate the removal of such information along with removing medical history.

Acknowledgment: This work was supported in part by grant R01HG00802 (NIH/NHGRI)
Self-Service Cohort Discovery across Five Academic Health Centers: Usage and User Evaluations of the University of California Research eXchange

Douglas S. Bell, MD, PhD1, Lisa Dahm, PhD2, Nicholas Anderson, PhD3, Ida Sim, MD, PhD4, Pralav Dessai1, Marianne Zachariah1, Lucila Ohno-Machado, MD, PhD5

1David Geffen School of Medicine at UCLA, Los Angeles, CA; 2University of California, Irvine, Orange, CA; 3University of California, Davis, Davis, CA; 4University of California, San Francisco, San Francisco, CA; 5University of California, San Diego, San Diego, CA

Introduction

The University of California Research eXchange (UC ReX) is a partnership among the five University of California medical campuses, aimed at enabling multi-campus research projects. In late 2013, we launched the UC ReX Data Explorer, an implementation of the SHRINE1 system, to support cohort discovery queries across the federated clinical data warehouses of the five campuses, which contain records from nearly 15 million patients in total.

Methods

We analyzed usage logs after excluding data from project personnel who performed system testing or user training. We programmed the SHRINE web client to randomly pop-up a survey invitation 30 seconds after query results appear, for 5% of queries. The invitation contained a link to a REDCap survey with questions about the specific query results and the user experience. The survey was active from 2/2015 to 5/2016.

Results

The system went into production 10/2/2013. In the 41 months through 2/28/2017, 426 distinct users ran a total of 8317 cohort queries (19.5 queries/user). Patterns of growth in users and usage differed among the five UC sites (Figure 1).

Figure 1. Growth in new users (left panel), and query volumes (right panel) at each of the UC ReX sites.

Users completed the pop-up survey for 60 queries, representing 26% of an estimated 231.5 opportunities (5% of 4630 queries during the time the survey was active). Users rated 75% of query results as clinically “expected” or “believable,” as opposed to “not likely” (17%), or “not possible” (7%). They rated the results as meeting their goals “completely” or “mostly” for 58% of queries, “somewhat” for 40%, and “not at all” for only 1 result (2%). The Data Explorer was rated as “extremely” (33%) or “somewhat” (42%) easy to use in most responses, whereas 10% rated it “somewhat” or “extremely” difficult to use. Ninety percent of users agreed or strongly agreed with the statement that they would recommend the Data Explorer to their follow researchers. User comments were mixed.

Conclusion

Most users of the UC ReX Data Explorer found the system to be useful and usable for self-service cohort discovery across the UC health systems. Investigators are making growing use of the system but the majority of clinical researchers are not using it. Programs to accelerate the uptake of cohort discovery are warranted.

References

Barriers to Prompting Beta Blocker Titration in Heart Failure at the Point of Care
Paige Benally, BS 1, Michael Smith, PhD 2, Charnetta Brown, MS 3, Natalie Kelly, MBA 4, Jose Nativi-Nicolau, MD, MS 5, Charlene Weir, PhD, RN 4, 6, Salim Virani. MD, PhD 3, Megha Kalsy, PhD 6, Jennifer H. Garvin, PhD, MBA 1, 4, 6

1 Department of Pediatrics, University of Utah, Salt Lake City, UT; 2 Engineering, Universidad de las Americas Puebla, MX; 3 Center of Innovation, Houston Veteran Affairs HSR&D, Houston, TX; 4 IDEAS Center, Salt Lake City, UT; 5 Division of Cardiology, University of Utah, Salt Lake City, UT; 6 Department of Biomedical Informatics, University of Utah, Salt Lake City, UT.

Introduction: Chronic heart failure (CHF) is a prevalent condition among Veterans within the Veteran Affairs (VA) health care system. A high percentage of Veterans are not receiving guideline-directed beta blocker medical therapy. We examined the barriers and facilitators related to evidence based care among primary care providers to inform the design of a clinical reminder in the VA electronic health record that would prompt appropriate ordering and titration of beta blockers in primary care Patient Aligned Care Teams (PACT).

Methods: Semi-structured usability interviews adapted from a cognitive task analysis perspective were used to elicit provider’s mental models of the care process, and the utility and usability of the clinical reminder prototype. Clinicians from PACTs from 2 VA medical centers and related community clinics were recruited for the interviews (4 physicians, 2 nurses, and 3 pharmacists). Providers were asked to participate from both rural and urban settings in two states. Each interview focused on workflow, clinical decision-making, institutional and contextual barriers and clinician’s attitudes toward using guidelines. Our research team reviewed the transcripts to answer the research questions, “Why are CHF patients in the VA being undertreated with respect to guideline recommended beta blocker use?”, and “What are the barriers to beta blocker titration?”, using applied thematic analysis. 3 Analysis was undertaken to identify themes in the narrative using an iterative approach, beginning with identification of basic PARIHS constructs and then development of emergent themes.

Results: Four themes were identified during preliminary analysis. There is: 1) Confusion about who is responsible for completing the reminders; 2) Concern about patient factors in tolerating titration; 3) Concern about patient comorbidities; 4) Concern about incomplete communication between providers and patients.

Conclusions: Our results provide insight to barriers for beta blocker titration within the PACT in VA. Potential changes to the next version of the reminder may include a link to the patient’s problem list and modifications in the titration story related to titrating to maximal tolerated doses given specific patient factors.

Acknowledgements: Supported by VA HSR&D project #CRE 12-037 from the United States (U.S) Department of Veterans Affairs VA IDEAS 2.0 HSR&D Research Center. The views expressed are those of the authors and not those of the Department of Veterans Affairs or affiliated institutions. We thank the VA stakeholders for their valuable contributions and the Departments of Pediatrics and Biomedical Informatics, as well as the Native American Research Internship at the University of Utah.

References:
Implementation and Adherence Evaluation of Intrusive Alerts for Adjusting Dose to Renal Function

Francisco M. Bernabeu MD1,2,3, María S. Hernández Caffot MD, PhD1,2, Nadime N. Sufan, Pharm.1,2, Fabiana Ellena MD2,3 Bayan V. Khorsandnia MD1,2
1) Departamento de Informática en Salud, 2) Instituto Universitario para las Ciencias Biomédicas de Córdoba (IUCBC), 3) Servicio de Clínica Médica, Hospital Privado Universitario de Córdoba (HPUC); Córdoba, Argentina

Introduction
Medication errors can trigger adverse events associated with increased mortality, excess length of hospitalization, and extra costs1. Among those are the ones related to the dosage, due to lack of adjustments required for different pathological conditions. Within these, renal impairment plays an important role since about 60% of the drugs are excreted in this way2. In the present study we analyze the implementation of a Clinical Decision Support System (CDSS) with an intrusive drug adjustment alert according to the renal function, and evaluate their adherence and the ability to avoid potential adverse drug events.

Material and Methods
A retrospective study was performed at the HPUC, during 6 months (July - December, 2016). A CDSS of maximum allowable dose was implemented for the 248 most frequently prescribed drugs, 94 of which required dose adjustment for renal impairment. The maximum adjusted doses of each drug were obtained according to creatinine clearance. The alert forced the user to make the decision to modify, cancel or, if justified because the patient was on dialysis, critical or terminal state, continue with the dose. All the hospitalized patients with a creatinine clearance less than 60 ml/min were included. Prescriptions carried out in the oncology sector and the pediatric population were excluded. Medication errors due to inadequate adjustment to renal function and the decisions made by the doctor were analyzed. Adherence was considered adequate (AA) when, in response to the alert, the user chose to cancel, modify or continue, as long as the justification was correct, corroborating the information through the review of medical records. If the justification was incorrect, the adhesion was considered inadequate (AI).

Results
We registered 629 renal dose adjustment alerts, 74 (11.7%) modified the dose of the drug (AA) and 155 (24.7%) canceled the prescription (AA). In 400 cases (63.6%) the physician decided to continue, justifying 136 (34%) cases as being on dialysis; 250 (62.5%) as critical patients and 14 (3.5%) as terminal patients. A sub-analysis by group of patients on dialysis showed that 33 justifications were incorrect as they corresponded to non-dialyzable drugs and 13 to patients who were not on dialysis (AI). From critically ill patients, 51 prescriptions were incorrect because they were non-critical drugs and 118 alerts were non-critical patients (AI), only 81 alerts were adequate in this subgroup (AA). 6 out of 14 terminal justifications were correct (AA). A total of 406 out of 629 (64.5%) alerts were considered as AA and 223/629 (35.5%) as AI.

Conclusion
Our CDSS allowed us to avoid 229 drug adverse events (corresponding to modified and canceled alerts) out of 629 maximum dose alerts according to renal function within 6 months. Despite the evidence in favor of the use of advanced CDSS, its adequate adherence continues to be suboptimal, with 64.5% of them in our study showing a high attempt rate of overriding (35.5%).

References
Development of a large-scale, whole body CT image database

Shamsi Daneshvari Berry, PhD, MS1,2, Heather J.H. Edgar, PhD2

1University of Mississippi Medical Center, Jackson, MS; 2University of New Mexico, Albuquerque, NM

Abstract

Some research questions cannot be answered because the data is simply not available. Such is the case for many projects that require a large, diverse, high quality dataset. A sample of over 11,000 whole body CT images has been created through a National Institute of Justice grant. After determining the metadata to associate with the images, a large scale database of CT scans and associated metadata is currently under development.

Introduction

Often researchers cannot answer the question of interest because the necessary dataset does not exist. In addition, the technological requirements for datasets grows as the sets themselves become larger. This is the case for questions requiring a representative samples of high quality Computed Tomography (CT) images and associated health and lifestyle data. Currently no such dataset exists.

Decedent Database

The Office of the Medical Investigator (OMI) is a state-wide, centralized medical examiner’s office for the entire state of New Mexico. Any individual who dies in the state in a sudden, violent, untimely, or unexpected manner, and any person who is found dead and the cause of death is unknown, is routed to the OMI for a possible autopsy. In 2010, 5,249 deaths were processed by the OMI, 51% of those underwent autopsy, accounting for 35% of the deaths within the state.1,2

The Center for Forensic Imaging at the OMI was awarded a grant from the National Institute of Justice in 2010 to evaluate whether CT scans can supplant or supplement the traditional autopsy. As a result, every decedent that underwent an autopsy received a high resolution, head-to-toe CT scan. To date this work has produced over 11,000 whole-body, 3-D CT images. However, there were no plans for the images after the 2010 National Institute of Justice grant was completed. As there was no appropriate minimum dataset (MDS), a modified Delphi method was implemented to determine the MDS to associate with the images to make them useful to the widest variety of researchers. A list of 59 metadata variables was developed with a high level of consensus by querying experts from a wide variety of domains (anthropology, forensics, informatics, medicine, demography, dentistry, public health, etc.).3 Data will be gathered from the OMI database as well as next of kin interviews.

In 2016, a National Institute of Justice grant was awarded to create this large-scale database and associated website. It has the potential to answer a large number of previously unanswerable questions that could improve health in the living.

Conclusion

The free-access Decedent Database is currently under development. Researchers will have access to this plethora of data and associated images by the end of 2018.

References

Identification of Patient Subgroups in Metastatic Breast Cancer Patients Based on Somatic Copy Number Alterations: A Bipartite Network Analysis

Suresh K. Bhavnani PhD1, Archana Ayyaswamy MS1, Tianlong Chen BS2, Jeremy L. Warner, MD MS3

1Institute for Translational Sciences, University of Texas Medical Branch, Galveston, TX; 2Dept. of Physics, Univ. of Houston, Houston, TX; 3Dept. of Medicine & BMI, Vanderbilt University, Nashville, TN

Abstract

While somatic copy number alterations (SCNA) in the human genome are known to exhibit complex patterns in cancer patients, less is known about whether they can help to identify patient subgroups critical for the design of targeted interventions in precision medicine. We therefore used a bipartite network analysis and a new network layout method called ExplodeLayout to analyze the heterogeneity within metastatic breast cancer specimens extracted from the GENIE database, with the goal of inferring the underlying mechanisms. The results revealed 5 co-clusters of specimens and SCNAs which were statistically significant, and enabled inference of complex but comprehensible SCNA heterogeneity in metastatic breast cancer.

Introduction

An increasing number of diseases have been shown to have distinct patterns related to copy number variations in the genome. Cancers in particular have a complex combination of somatic gene deletions and amplifications, driving heterogeneity and genomic instability.1 Furthermore, as most amplifications of copy number variation events cause increased gain of function and fitness of the cancer, they result in a poor prognosis. We therefore explored the use of bipartite networks to identify patient subgroups across metastatic breast cancer patients, with the goal of inferring the underlying mechanisms, and designing targeted interventions.

Method and Results

The Genomics, Evidence, Neoplasia, Information, Exchange (GENIE) supported by the AACR is an international data-sharing project that aims to catalyze precision cancer medicine. This database has 18,966 specimens across many cancer types from which we extracted all metastatic breast cancer specimens (n=1,102), of which a subset (n=894) had SCNA data. In order to optimize modularity, we restricted the analysis to the 20 most-commonly amplified genes, resulting in 527 specimens. To analyze heterogeneity in these data, we conducted a bipartite network analysis2 using the following method: 1) conducted a co-cluster modularity2 analysis to identify the number and membership of specimens and genes in co-clusters; 2) calculated the significance of the clustering using 1000 permutations of the network while preserving size and density; 3) used Fruchterman Reingold2 (FR) to lay out the network; and 4) used a new network layout method called the ExplodeLayout3 algorithm designed to separate significant but highly-overlapped clusters to improve comprehension. This algorithm takes as input the network layout from FR, and node cluster membership from the co-cluster modularity, and maps each cluster’s centroid onto equidistant points on an imaginary circle. The algorithm searches for a circle radius such that the ratio of the total non-overlapped area among the clusters, and the total space required for the entire network layout is maximized, resulting in an optimal cluster separation within a compact area. The resulting network was then interpreted by an oncologist to comprehend the patient subgroups and infer their underlying mechanisms.

Before application of the ExplodeLayout algorithm, there was no visually discernible network structure. After the algorithm was applied, as shown in Fig. 1, the analysis revealed 5 clusters with strong clusteredness (co-cluster modularity=0.49), which was significant compared to 1000 permutations of the network (z=37, p<0.001). The results helped to confirm a known breast cancer subtype (ERBB2 a.k.a. HER2-amplified breast cancer) and to identify new molecularly-defined subtypes. Two clusters (pink and gray) are driven by ampiclon events (11q13 and 8q24, respectively). The green cluster is driven by amplification of the 1q chromosome arm, and represents a group of cancers dependent on the anti-apoptotic gene MCL1.1 The yellow cluster contains two genes that are known to be in the stem cell pluripotency pathway and have been described as being co-amplified in brain metastases4; these patients are likely to have particularly aggressive disease. Future research will test the replicability of the results in other datasets, the utility of the above analytical approach in other cancer types.

References

Implementing a Hash-based Privacy-Preserving Entity Resolution Tool in the OneFlorida Clinical Data Research Network

Jiang Bian, PhD1, Andrei Sura, BS1, Gloria Lipori, MT, MBA1, Yi Guo, PhD1, François Modave, PhD1, Zhe He, PhD2, Elizabeth Shenkman, PhD1, William Hogan, MD, MS1

1University of Florida, Gainesville, Florida, USA; 2Florida State University, Tallahassee, Florida, USA;

Introduction: The OneFlorida Clinical Research Consortium is one of the 13 PCORI-funded Clinical Research Data Networks as part of the National Patient-Centered Clinical Research Network (PCORnet) in support of large-scale pragmatic clinical trials and implementation science studies. A key component of OneFlorida is its Data Trust, a secure data repository currently containing 10.6 million patient records, currently including pediatric claims data from Florida Medicaid and electronic health record (EHR) data from 9 unique health systems across the network. However, data records of the same patient can come from different data sources in the OneFlorida network. For example, EHRs and claims have complimentary data records for the same patient. Thus, entity resolution (ER)—linking related data and resolving duplicates—at the individual level in a research network is a significant task in improving the data quality. Nevertheless, linking patient records is not a trivial task balancing between linking efficiency and privacy needs. Record linkage is nonetheless a well explored problem. In this study, we adapted a privacy-preserving record linkage method1 and implemented it as an open source tool (available at: https://github.com/ufbmi/onefl-deduper).

Methods: We used a deterministic (rule-based) ER method. Unlike existing methods, we did not use any direct identifiers, such as social security numbers (SSN). Instead, our linkage rules are combinations of quasi-identifiers such as name, date of birth (dob), race, and sex. We consider two records (A and B) to be about the same patient if one or more combinations of quasi-identifiers have the same corresponding values (e.g., A.name + A.dob + A.race = B.name + B.dob + B.race) and these combinations can uniquely identify the patient in the datasets. Thus, the problem of linking patient data is transformed into equivalence tests based on these linkage rules. We learned these linkage rules—combinations of quasi-identifiers that maximize linking precision (i.e., minimize false positives)—using publically-accessible Florida voter-registration data. We tested the rules using a sample (~1000) of University of Florida (UF) Health and Florida pediatric Medicaid patient data that were linked using Medicaid beneficiary identifiers and manual review.

Results: Using the Florida voter registration data (~13 million), we learned two rules: 1) first name + last name + dob + race; and 2) first name + last name + dob + gender, that can uniquely identify a person (0.03% and 0.04% duplication rates, respectively). On the manually-curated linked sample data, both rules achieved perfect precision (no false positives), and reasonable recall (0.463 and 0.715, respectively). Using both rules, we implemented a Python-based open-source ER tool for linking patient records in the OneFlorida network. We consider two records to be about the same patient if either of the two rules predict a match. The overall recall of the system is at least 0.715, with a perfect precision. To preserve patient privacy, we used a one-way, collision-resistant, cryptographic hash function (seeded SHA-256) to scramble the protected health information (PHI) used in these linkage rules. It is computationally impossible to reverse the hash function to find the original PHI based on only the hash values. Further, to account for common data entry issues, we also incorporated several data normalization procedures.

Conclusion: Using the developed ER tool, we linked the entire UF Health and Florida pediatric Medicaid patients, and found 204,919 matches. A privacy-preserving ER solution is highly desired for clinical research networks like OneFlorida to improve data quality while ensuring the privacy of all whose data are in the Data Trust.

References

Validation of Diabetes Identification within Electronic Health Records

Katie E Bickett, MS₁, Timothy D. Imler, MD, MS₁, ᴲ
¹Regenstrief Institute, Indianapolis, IN, USA; ²Indiana University School of Medicine, Indianapolis, IN, USA

Introduction: Diabetes mellitus (DM) affects 9.3% of the United States population and was estimated to cost $245 billion in 2012.¹ DM is commonly studied in large-scale populations or as a sub-population of clinical trials. However, this is often done utilizing the International Classification of Diseases (ICD) with the large potential for misidentifying patients and classifying them incorrectly due partly to unintentional and intentional coder errors. The identification methods developed over time by individual researchers has never fully been clinically and statistically validated.

Aim: Validate statistical measures of the current standard query for identification of DM patients with clinical review.

Methods: After IRB approval, available clinical data from the Indiana Network for Patient Care (INPC) a large regional health information exchange receiving eMR data in real time, was utilized from 1/2/2014 through 12/31/2014. Patients were identified as having DM if they met at least one of the following conditions: 1) a clinical encounter (e.g., clinic visit) representative of DM (ICD diagnosis codes), 2) pharmacy order/fills for particular medications (e.g., insulin, tobutamid, glipizide, acarbose, etc.), or at least two of one of the following conditions: 3) abnormal fasting and resting glucose values, or 4) hemoglobin A1c values. Patients were excluded if they were < 18 years or > 89 years of age during 2014 (Figure 1).

Patients from the assumed DM and non-DM populations were randomly sampled limiting to one institution to allow complete manual clinical review (83 from DM population, 67 from non-DM). Each patient’s clinical text was reviewed by a clinician (TI) in a blinded manner (random mixing of DM identified and non-DM identified patients). The review served as the gold standard for the study and placed the patients into DM Type I, DM Type II, DM other (e.g., pancreatic type), and no DM categories. Groups DM I, II, and other were considered diabetics and utilized for analysis.

Sensitivity, specificity, and positive predictive value (PPV), and negative predictive value (NPV) along with their exact binomial 95% confidence intervals (CI) calculated based on the results of the standard query and manual review. Statistical analysis was done utilizing SAS version 9.4.

Results: In total, there were 3,742,500 patients seen in 2014 within the INPC and specifically, 148,651 at the institution for manual review. There were 10.5% of patients identified as having diabetes with average age 54.5±13.7 (range) with 58% female (Table 1).

The identification of diabetes achieved a sensitivity of 98% (95% CI (0.91-0.99)), specificity of 94% (95% CI(0.86-0.98)), PPV of 95% (95% CI (0.88-0.98)) and NPV of 97% (95% CI (0.90-0.99)).

Conclusion: Standardization and validation of common clinical diseases for population tracking and clinical trials is a necessary part of the complex and siloed electronic health care system. We demonstrate the ability to clinically and statistically validate within a health information exchange one of the most common diseases utilized for both primary and secondary research analysis. Further discussion is needed to create a standardized repository of validated queries from real-world electronic health data.

Table 1. Demographics of 2014 Population

<table>
<thead>
<tr>
<th></th>
<th>DM (%)</th>
<th>No DM (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>54.5±13.7</td>
<td>39.8±15.1</td>
</tr>
<tr>
<td>Female</td>
<td>9,048 (57.9)</td>
<td>57,598 (59.7)</td>
</tr>
<tr>
<td>Race/Ethnicity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>5,205 (33.4)</td>
<td>36,014 (37.3)</td>
</tr>
<tr>
<td>Black</td>
<td>7,682 (49.3)</td>
<td>38,807 (40.2)</td>
</tr>
<tr>
<td>Hispanic</td>
<td>611 (3.9)</td>
<td>7,224 (7.5)</td>
</tr>
<tr>
<td>Other</td>
<td>2,090 (13.4)</td>
<td>14,496 (15)</td>
</tr>
</tbody>
</table>

Reference

RxNorm Concept History Service

Olivier Bodenreider, M.D., PhD and Lee Peters, M.S.
U.S. National Library of Medicine, National Institutes of Health, Bethesda, Maryland, USA
Contact information: RXNAVINFO@LIST.NIH.GOV

Motivation
RxNorm contains drug concepts which represent the currently marketed drugs in the U.S. With each monthly RxNorm release, new drug products are added and obsolete drug products are removed. While this curation supports e-prescribing and drug information exchange use cases, it is detrimental to analytics use cases. The issue in this case is that some of the RxNorm drug identifiers (RxCUIs) stored in clinical data warehouses have become obsolete and can no longer be interpreted in reference to the current RxNorm dataset. There are over 122,000 obsolete RxCUIs, compared to 114,000 current ones. To address this issue, we started developing an RxNorm concept history service as part of the NLM drug APIs (https://rxnav.nlm.nih.gov/). This service is similar to what we developed for managing obsolete identifiers from the National Drug Code (NDC) Directory. For each obsolete RxCUI, the service returns a canonical representation of the drug concept, making it possible to relate an obsolete drug product to a class through its ingredient or to find a similar active drug product based on ingredient, strength and dose form information.

Creating the Historical Concept Data
To create the historical concept information, the RxNorm monthly releases starting in 2007 were queried to reconstruct the history of every RxNorm concept. Relevant features extracted include:
- The ingredient name(s) and RxCUI(s) related to the concept
- The term type (TTY)
- Start and end dates when the concept was active
- Indication if the concept is currently active
- Indication if the concept represents a brand
- Indication if the concept contains multiple ingredients
- Strength (number and unit) of each ingredient
- Dose form name and RxCUI
- For Packs (TTY=GPCK and BPCK), pack alias name, product component RxCUIs and pack numbers

Table 1 shows an example of the history data returned for the RxCUI =853056

<table>
<thead>
<tr>
<th>ATTRIBUTE</th>
<th>VALUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>RXCUI</td>
<td>853056</td>
</tr>
<tr>
<td>NAME</td>
<td>Amoxicillin 40 MG/ML / Clavulanate 5.7 MG/ML Oral Suspension [Amoclan]</td>
</tr>
<tr>
<td>TTY</td>
<td>SBD</td>
</tr>
<tr>
<td>INGREDIENT NAME(S)</td>
<td>Amoxicillin / Clavulanate</td>
</tr>
<tr>
<td>IN/MIN RXCUI</td>
<td>19711</td>
</tr>
<tr>
<td>MULTIPLE INGREDIENT</td>
<td>Yes</td>
</tr>
<tr>
<td>BRANDED</td>
<td>Yes</td>
</tr>
<tr>
<td>START DATE</td>
<td>072009</td>
</tr>
<tr>
<td>END DATE</td>
<td>122015</td>
</tr>
<tr>
<td>CURRENT</td>
<td>No</td>
</tr>
<tr>
<td>QUANTITY FACTOR</td>
<td></td>
</tr>
<tr>
<td>QUALITY DESCRIPTOR</td>
<td></td>
</tr>
<tr>
<td>COMPONENT 1:</td>
<td></td>
</tr>
<tr>
<td>RXCUI</td>
<td>723</td>
</tr>
<tr>
<td>STRENGTH #</td>
<td>40</td>
</tr>
<tr>
<td>STRENGTH UNIT</td>
<td>MG/ML</td>
</tr>
<tr>
<td>COMPONENT 2:</td>
<td></td>
</tr>
<tr>
<td>RXCUI</td>
<td>48203</td>
</tr>
<tr>
<td>STRENGTH #</td>
<td>5.7</td>
</tr>
<tr>
<td>STRENGTH UNIT</td>
<td>MG/ML</td>
</tr>
<tr>
<td>PACK ALIAS</td>
<td></td>
</tr>
</tbody>
</table>

Table 1. RxNorm History Data returned for RxCUI=853056

Acknowledgments: This work was supported by the Intramural Research Program of the NIH, National Library of Medicine.
Does Semantic Tag Usage in SNOMED CT Match its Concept Hierarchy?
Jonathan P. Bona, PhD
1
1University of Arkansas for Medical Sciences, Little Rock, AR, USA

Introduction

Every SNOMED CT concept is described by a Fully Specified Name (FSN) ending in a semantic tag that disambiguates it from other concepts with similar names and indicates where it fits into the SNOMED concept hierarchy. For example, the concepts [35566002 | Hematoma (morphologic abnormality)] and [385494008 | Hematoma (disorder)] are tagged ‘morphologic abnormality’ and ‘disorder’. In the hierarchy these are subsumed by the highest-level concepts (i.e. closest to the top concept) for morphologic abnormalities and disorders, respectively.

Because semantic tags are substrings within names, and not represented as part of SNOMED’s formal model, it is non-trivial to know whether a concept’s tag should be taken to mean that the concept is necessarily part of the same sub-hierarchy as others with that tag. A concept’s tag would strictly identify its place within the hierarchy if each tag had a single, high-level corresponding concept that used it, and every concept using the tag was below that high-level concept in the hierarchy. For instance, in the clinical finding hierarchy, [404684003 | Clinical finding (finding)] subsumes all other finding concepts. This work examines the January 31, 2017 International Release of SNOMED CT to determine the extent to which semantic tag usage by concepts matches concept placement in the hierarchy.

Methods

We developed computational procedures to answer the question: given a semantic tag S, and a corresponding concept C, is it the case that every concept tagged with S is subsumed by C in SNOMED CT’s concept hierarchy?

For each semantic tag we identified a tag corresponding concept – the concept closest to SNOMED’s root that uses the tag. In the January 31, 2017 release there is one such concept for every tag. To make use of the subsumption reasoning built into semantic web tools, we constructed an RDF/OWL model of SNOMED CT’s concept hierarchy, with each of the 300,000+ SNOMED concepts represented as an OWL class with separate annotations for its FSN and semantic tag. Each SNOMED is-a relation between concepts has a corresponding rdf:subClassOf assertion.

This model was loaded into a triple store database that pre-computed subsumption information, making it accessible for very fast retrieval using simple SPARQL queries. For each semantic tag a query finds every concept with that tag but that is not subsumed by its corresponding concept. Where such concepts exist, they seem to be misplaced or mis-tagged. That is, they are concepts whose semantic tags do not accurately represent their positions in the hierarchy.

Results

While nearly all of SNOMED’s semantic tags (35 out of 42) do have clear corresponding concepts that subsume all other concepts that use the tag, we found exceptions. Two instructive examples are.

- The tag ‘substance’ has a corresponding high-level concept, [105590001 | Substance (substance)] that subsumes 26,000+ concepts also tagged ‘substance’. There is only one single concept tagged ‘substance’ that it does not subsume: [109186003 | Sickle cell test kit (substance)]. This is a direct sub-concept of [385387009 | Test kit (physical object)], as are many other types of test kits. That is, it is situated in the right area of the SNOMED hierarchy, but inconsistently tagged. This is likely an error.

- The ‘disorder’ tag’s corresponding high-level concept, [64572001 | Disease (disorder)] subsumes 72,000+ concepts tagged ‘disorder’. There are 84 ‘disorder’ concepts that it does not subsume, including: [250054005 | Frontal gait disorder (disorder)], [431193003 | Infection of bloodstream (disorder)], and [708484008 | Vesiculoerosive lesion (disorder)]. These require further investigation. No clear pattern is evident at this stage, nor is there an obvious reason that these should not be subsumed by [64572001 | Disease (disorder)].

Whether or not a strict correspondence between tags and concept hierarchy is intended, examining cases where it does not hold is instructive in detecting anomalies that may be errors. A more formal and consistent use of semantic tags would improve their usefulness, both for knowledge representation and reasoning in software systems that use SNOMED CT, and for users working with SNOMED and with data annotated using it.

* The author would like to thank Professor Werner Ceusters, MD for sharing his ideas during discussions that led to this work.
Electronic Health Records and their Modules: Impacts on Clinical Practice and Patient Care

Elizabeth M. Borycki, RN PhD¹, Andre W. Kushniruk, MSc PhD FACMI¹, Yalini Senthirajah PhD²
¹University of Victoria, Victoria, British Columbia, Canada; ²Hofstra Medical School, New York, United States of America²

Abstract

Health information technology (HIT) safety has emerged as an important issue. To date much of the research has focused on researching electronic health record (EHR) usability and safety. Less work has been conducted on the impacts of integrated EHR modules on overall EHR usability and patient safety.

Introduction and Background

HIT can have a significant effect on patient safety. HIT can improve patient safety, but it can also introduce new types of medical errors (i.e. technology-induced errors)¹. Technology-induced errors arise from interactions between health professionals and HIT during patient care. Researchers have studied the usability, workflows and safety of EHRs. Much of this work has focused on a complete EHR as the unit of analysis or individual modules that can be found in an EHR (e.g. e-prescribing, medication administration systems) as the unit of analysis. Less research has focused on the usability and workflow of integrated modules within an EHR and the effects of different user interface designs and workflows between these modules and how this might affect patient safety. In this poster we describe some of our preliminary findings arising from a pilot study of the integration of modules within an EHR.

Methods

Physicians and nurses (n=8) were recruited from medical units, in a fully digitized medical center in the United States that employs a robust EHR that is used by health professionals across the organization. The study took place in a laboratory setting, where participants were asked to review a patients’ EHR as they would in their day to day work. Participants were interviewed using a semi-structured interview approach about the EHR modules and their integration, while reviewing it. The participants’ verbalizations were audio recorded and transcribed. The data were coded using a published, qualitative coding scheme that focuses on usability and safety by three coders and their emerged a high degree of agreement.¹

Results

Findings suggest the EHR and its modules had specific usability and workflow issues that could influence safety. In terms of the overall EHR, it was noted, that activities such as logging on and off the system throughout the day diverted clinicians away from patient care activities and decreased opportunities for face-to-face contact with patients. Participants identified that EHR modules (i.e. clinician order entry, clinical documentation, laboratory information systems and diagnostic imaging systems) had differing usability and workflow issues and their lack of seamless integration affected cognitive and physical work involving patients. These issues included a diminished ability of clinicians to: (1) identify and monitor for patterns of disease and response to treatment, (2) respond expeditiously to new orders, and (3) spend time in direct patient care or face-to-face interactions with the patient.

Conclusion

Our preliminary findings from this pilot study suggest there is a need to analyze the usability and safety of a full EHR, its individual modules and the integration of EHR modules. Such analyses are important to identifying issues that may influence situational awareness of clinicians and diminish clinician ability to spend more time providing patient care. Research is needed to identify ways in which such issues may be overcome (i.e. use of composable EHRs, user interface and workflow guidelines for EHRs and their modules).

References


Acknowledgements: This work was funded by AHRQ #5R01HS023708-02.
Leveraging Large-Scale Computing for Population Information Integration, Analysis, and Modeling

Jessica A. Boten, MPH¹, Donna R. Rivera, PharmD, MS¹, Madhumita Myneni, MBA, MS¹, Georgia Tourassi, PhD², Tanmoy Bhattacharya, PhD³, Ana Paula de Oliveira Sales, PhD⁴, Thomas Brettin, MS⁵, Paul A. Fearn, PhD¹, Lynne T. Penberthy, MD, MPH¹

¹National Cancer Institute, Rockville, MD; ²Oak Ridge National Laboratory, Oak Ridge, TN; ³Los Alamos National Laboratory, Los Alamos, NM; ⁴Lawrence Livermore National Laboratory, Livermore, CA; ⁵Argonne National Laboratory, Chicago, IL

Introduction

The National Cancer Institute’s (NCI) Surveillance, Epidemiology, and End Results (SEER) Program provides cancer data and statistics to support population-based research in an effort to examine and reduce the cancer burden in the United States. As the complexity of cancer care, the dispersion of cancer diagnosis and treatment across multiple healthcare providers and settings, and data generation to characterize each type of cancer increases, the SEER Program seeks methods to address these challenges. NCI is collaborating with four national laboratories of the Department of Energy (DOE) on a pilot project that will leverage the capabilities of high performance computing to support a comprehensive population-level cancer surveillance program and develop an integrated modeling framework.

Goals

The overarching goal of this pilot research program is to deliver the advanced computational and informatics solutions needed to support a comprehensive, scalable, and cost-effective cancer surveillance program and lay the foundation for an integrative data-driven approach to modeling cancer outcomes at scale. This goal has been broken into three aims. Aim 1 will develop scalable machine-learned natural language processing (NLP) tools for deep comprehension of unstructured clinical text to enable automated and accurate capture of reportable cancer surveillance data elements with case-specific uncertainties quantified to guide manual abstraction. Aim 2 will build a working infrastructure that integrates large scale linkage of heterogeneous data sources for research and discovery with a focus on methodology and bias assessment. Scalable graph and visual analytics tools guide exploration of disease specific cancer surveillance algorithms and methods to accelerate understanding of cancer outcomes including progression, recurrence, metastases, and survival. Aim 3 will develop a cancer surveillance data-driven modeling and simulation environment for predicting health trajectories of cancer patients.

Progress

NCI developed an infrastructure for document selection, annotation, algorithm development, and data hosting. This infrastructure has been utilized for human annotation of data elements in electronic pathology reports to facilitate algorithm development. Using deep learning methods, DOE researchers have been developing robust NLP tools that will identify, with known uncertainty, site, histology, laterality and behavior in pathology reports that will be utilized by cancer registries. External heterogeneous data sources have been identified for linkage with SEER data to enable development of longitudinal patient trajectories that will support modeling efforts. Extensive work in variable selection, clinical treatment pathways, and algorithm development is ongoing. Furthermore, clinical experts have joined the collaboration to identify and develop modeling use cases for specific cancer sites.

Conclusions

The NCI-DOE collaboration will enhance the SEER program through automated data abstraction, linkage and visualization of data sources, and predictive modeling. This application of data driven modeling seeks to answer key clinical oncology questions to empower clinician decision making tools which can improve patient treatment selection and health outcomes.
Detection of Opioid-Dependence: Extracting Opioid Use Patterns from Patient Secure Messages

Lina Bouayad, PhD¹ ², Dezon Finch, PhD¹, Stephen L. Luther, PhD¹

¹James Haley Veterans Hospital, Tampa, FL; ²Florida International University, Miami, FL

Introduction

Data about opioid use patterns that may be used to detect and predict the risk of opioid-dependence are stored in the EHR in both coded fields and progress notes. However, research indicates the structured data in EHR are often incomplete and inconsistent. Prior literature has targeted mining of clinical documents written by providers to increase data available for analysis. Yet, data extracted from unstructured progress notes contains provider-reported information only. Recently, secure messaging has been developed as a tool that provides patients with convenient and fast access to healthcare providers regarding non-urgent health concerns. As of October 2015, an exponential growth in secure messaging (SM) adoption was reported among veterans with more than over 1 million secure messages (SMs) sent. SMs represent an important new resource that was previously inaccessible for clinical care or research. Patient-reported use patterns can be extracted from SMs. Data extracted from SMs could be integrated into the EHR (as structured fields) to (1) enhance predictive risk models for opioid-dependence, and (2) improve treatment of patients with pain. To our knowledge, this would be the first study to use analytics to explore the content of SMs. If successful, the method could be applied to other high-frequency, high-risk, and/or high-cost conditions.

Methods

This retrospective study looked at cohort of Veterans who 1) have been diagnosed with pain, 2) have been prescribed opioid therapy during FY12-FY13, and 3) have used SM to communicate with clinicians. Analyzing our sample data, we identified 1,840,747 unique Veterans who were prescribed common opioids (CN101 for opioid analgesics, CN102 for opioid antagonist analgesics) between FY12 and FY13. We then searched for keywords indicating opioid use patterns within the secure messages associated with that cohort of patients - 100,000 randomly selected messages from patient with opioid dependence (ICD 9=304.01), and 100,000 randomly selected messages from patient without opioid dependence. Future text mining work on SMs can therefore be adapted to extract opioid use patterns and aberrant drug-taking behavior¹ (such as early refills, lost prescriptions, use despite harm, undertreated pain, confused and drug-taking behavior, chemical-coping behaviors, and self-medication). Results can then be applied to build models to detect and predict opioid dependence. Additional results will be presented at the conference.

Results

Using SM data (through the VA Informatics and Computing Infrastructure (VINCI)), we were able to complete two preliminary investigations. Results indicating the prevalence of keywords indicating opioid-related use patterns are listed in Table 1. Results also indicate 1) the presence of documentation of use patterns in patient secure messages, and 2) difference in frequency of these terms used by patients with and without opioid dependence. Future text mining work on SMs can therefore be adapted to extract opioid use patterns and aberrant drug-taking behavior¹ (such as early refills, lost prescriptions, use despite harm, undertreated pain, confused and drug-taking behavior, chemical-coping behaviors, and self-medication). Results can then be applied to build models to detect and predict opioid dependence. Additional results will be presented at the conference.

Conclusion

In the present study we assessed the presence of documentation of opioid use patterns in patients’ secure messages. Our future work intends to use text mining and machine learning algorithms to extract these use patterns from SMs which will enable (1) comparison between patient- and provider- reported data and (2) integration of patient-reported symptoms into structured fields for further analysis.

References


<table>
<thead>
<tr>
<th>Table 1: Term SM Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Term</td>
</tr>
<tr>
<td>Refill</td>
</tr>
<tr>
<td>Filled</td>
</tr>
<tr>
<td>Prescription</td>
</tr>
<tr>
<td>Renew</td>
</tr>
<tr>
<td>Lost</td>
</tr>
<tr>
<td>Order</td>
</tr>
<tr>
<td>Reorder</td>
</tr>
<tr>
<td>Forgot</td>
</tr>
<tr>
<td>Pain</td>
</tr>
</tbody>
</table>
Toward a reliable and interoperable public repository for natural product-drug interaction study data

Richard D. Boyce, PhD1, Isabelle Ragueneau-Majlessi, MD, MS2, Jingjing Yu, PhD2, Jessica Sontheimer, PhD2, Chris Kinsella2, Mathias Brochhausen, PhD3, John Judkins, MS3, Bruce Pinkleton, PhD4, Rebecca Cooney, MA4, Mary F. Paine, RPh, PhD5, Jeannine S. McCune, PharmD2

1 University of Pittsburgh, PA; 2 University of Washington, Seattle, WA; 3 University of Arkansas for Medical Sciences; Washington State University, 4 Pullman and Spokane, WA

Introduction

Concomitant use of prescription drugs and natural products, including vitamin, mineral, or herbal supplements, is a frequent occurrence. Approximately 50% of Americans in the Midlife in the United States Study were exposed to prescription drugs at the same time as a natural product1, raising concerns for adverse natural product-drug interactions (NPDIs). In a different study, approximately 34% of patients prescribed warfarin were exposed to natural products that conferred a significant risk for an adverse interaction2. Pharmacokinetic-based NPDIs (PK-NPDIs) are of particular concern because their potential impact on drug effectiveness or toxicity is often unknown. To provide evidence-based information regarding purported PK-NPDIs, a new Center of Excellence for PK-NPDI Research was established (U54 AT008909). The Center is creating a publicly accessible database where researchers can access scientific results, raw data, and recommended approaches to optimally assess the clinical significance of PK-NPDIs.

Methods

One critical component of the Center is the Informatics Core (IC). The IC is leveraging prior experience with an existing proprietary repository of preclinical and clinical drug-drug interaction data3 to build the new publicly accessible PK-NPDI database. The IC works with the Center’s Analytical and Pharmacology Cores to gather, organize, and archive the various types of data that the teams are generating prospectively (Figure 1). In parallel, the IC is extending the Potential Drug-drug Interaction and Potential Drug-drug Interaction Evidence Ontology (DIDEO) ontology4 to provide a common set of data elements for the database. Finally, the IC is engaging PK-NPDI researchers in a user-centered design process to create a public facing portal to the PK-NPDI database that meets their information needs.

Conclusion

The IC has been operating since Fall 2015 and has 1) implemented a beta cloud-based data repository with a web-based data entry application, 2) made significant progress on the use of standardized terms and operating procedures for data entry, and 3) initiated a user-focused information needs questionnaire that will result in user scenarios for PK-NPDI researchers that will be useful for designing a public portal to the repository.

References

Leveraging Electronic Medical Record Data for Research with Consistency and Efficiency: A Use Case in an Ongoing Obstetrics Clinical Study

Robert L Bradford, BSBC, CCIS¹, Karen Dorman, RN, MS¹, Alison Stuebe, MD, MSc¹, Wanda Nicholson, MD, MPH, MBA¹, Kiira Lyons, MA¹, Anna Maria Siega-Riz, PhD²
¹University of North Carolina at Chapel Hill, Chapel Hill, NC; ²University of Virginia, Charlottesville, VA

Introduction:
Many epidemiological and clinical studies have proposed using the electronic medical record (EMR) for abstracting pertinent patient information for their research questions of interests. Additionally, by utilizing EMRs to obtain patient data, researchers strive to improve efficiency, consistency and reduce costs of clinical studies. However, inherent complexities within clinical operational data in EMRs present unique challenges for researchers in completing these tasks for clinical studies¹. For an on-going pregnancy cohort study (n=458), our goal was to identify information in the EMR (Epic 2014) in a consistent manner relating to pregnancy observations including: previous medical history, health habits, prenatal vitals that exists in flow sheets, medication use, clinical laboratory results, problem lists and details related to labor, delivery and the postpartum visits.

Methods:
Multiple efforts were necessary to complete the initial report and develop a re-usable algorithm for this study. Efforts included meetings of the clinical team and analyst, as well as continued communication throughout. Screenshots were used to develop the overall data flow, and to locate critical workflow items to be traced within the EMR database. Data tools including data dictionaries, vendor reporting documentation, and others were necessary to locate the required data. Two rounds of data extraction for women participating in the cohort allowed us to examine reproducibility of our methods.

Results:
The clinical objectives of the study required reconciling various clinical views and pathways in the Epic EMR with the data available in the Epic Clarity reporting database. Achieving this goal required deciphering and translating data in the database to a clinically relevant format, rebuilding relationships (i.e. connecting mothers with their infants), and addressing obstacles with patient identity tracking through medical record number changes over the course of the study. Addressing each of these obstacles required significant collaboration between the technical and clinical teams. The upfront investment of the clinical team and analyst totaled approximately 80 FTE hours, which were spread across meetings, programming, and testing to develop the initial data pull for 100 patients. In contrast to a second data pull (n=174), time required dropped significantly to 5 FTE hours.

Conclusions:
In this pregnancy cohort study, we addressed multiple challenges in extracting data to prepare a reconstruction of clinical views in the Epic EMR. While exploring these pathways, collaboration between clinicians and analysts was paramount to reconciling data discrepancies, and in some cases, modifying clinical workflows to bring consistency to data collection. Additionally, we developed functionality to handle medical record number changes and merges that can occur within the medical record system. These procedures can be universally employed by other EPIC users and be adapted for other EHR systems for the purpose of extracting data. Future considerations include potential changes in data flow resulting from system updates to Epic EMR requiring re-validation of the data extraction methods for accuracy and efficiency.

Enabling Collaboration for Building High Quality, Sustainable and Scalable National Health Information Systems in Resource-Limited Settings

Pascal Brandt, MSc1, William B. Lober, M.D., M.S., Jan Flowers, M.S.1
1School of Nursing, University of Washington

In resource-limited settings, health information systems are often developed by donor funded organizations. These organizations frequently work in silos to ensure they meet their specific grant requirements. As a result, efforts in the country are duplicated and developed solutions lack oversight and are underutilized. With development expertise lacking, best practices and standards are often not followed. Furthermore, knowledge gained is siloed within a single organization. While resulting systems may meet the functional requirements at the time of release, scalability and maintainability is challenging due to ad hoc development processes. As organizations scale these systems, they often end up with multiple software versions and configurations across facilities that are difficult to support and manage. Systems built with little coordination between organizations and a lack of interoperability results in a fragmented national healthcare information system. Ministries of Health faced with this fragmented system may struggle to effectively manage and use all of these systems to manage and improve care and outcomes in their country.

In order to address these issues and enable collaboration, shared technical infrastructure and processes need to be in place to support health information system development. Processes need to be agreed upon by all partners and well documented to enable onboarding of new members and organizations. Functionality should be designed collectively to enable broad applicability, final products should be well documented and demonstrated to allow knowledge transfer between individuals and organizations. Mechanisms should be in place to ensure the delivery of high quality products, and reliable automated delivery pipelines should ensure predictable release timelines and enable easy upgrading. Finally, systems developed should work together to enable a health information exchange.

In partnership with local stakeholders supporting the Ministry of Health, we implemented a solution in Mozambique that addresses some of these quality issues. The initial step was the establishment of a cross-organizational software development community to collaborate on OpenMRS, called eSaude1. The community has adopted a number of open sources tools for project management, including Trello for task scheduling, a Google Groups mailing list, and Slack for real-time chat between community members. For source code collaboration and version control, eSaude uses GitHub. The Travis continuous integration platform, which is free for open source projects, automatically runs tests against all newly contributed code and publishes a packaged deployable end product. The tests programmatically validate all new code against community agreed functional and style guidelines, ensuring every new feature is high quality and error free. This gives junior contributors the confidence to add their contributions, and as a result gain experience as developers. The community uses container technology by the open source Docker project to generate predictable and sharable deployable packages that are easy to install and upgrade across sites and ensure consistent versioning. An implementation packet, and related process and technical documentation was published to support consistent implementation process in the MoH facilities.

As a result of establishing the eSaude community in Mozambique, the 5 HIV care & treatment implementing partners work together to build shared health information systems, using software development best processes collectively agreed upon. The structured setup of standard tools and processes enables the effective contribution of developers with a range of experience, while deduplicating efforts across organizations, and creating a scalable and manageable national platform of HIS products. This process is ongoing, and we’ve successfully released numerous versions of a growing set of facility-level clinical applications.

References


1 http://www.esaude.org/
Modular Health Informatics Architecture for a Patient Centered mHealth Diabetes Intervention
Ian M. Brooks, Ph.D.,1 Justin Gatwood, Ph.D., MPH,2 Bonnie Binkley, MA,3 Emanuel Villa, MS,4 Jennifer Featherston, MS,4 James E. Bailey, MD, MPH,3
1-Dept. Biomedical Informatics, MedStar Health Research Institute, Hyattsville, MD; 2-Col. of Pharmacy, UTHSC, Nashville, TN; 3-Center for Health Systems Improvement, UTHSC, Memphis, TN; 4-Center for Biomedical Informatics, UTHSC, Memphis

Introduction and Background
Safety-net primary care practices throughout the country are seeking to develop capacity to inform and activate their most vulnerable patients. Poor self-care behaviors are common, but behavioral interventions are intensive and sustaining health benefits over time requires individualized approaches to address patient-specific health beliefs. Text messaging (TM) is an effective, low-cost approach for engaging patients with chronic disease in better self-care, and research suggests that health-related, tailored TM effectively influences specific behaviors. Our previous research showed that diabetes patients living in medically underserved areas (MUA) were very interested in receiving TM from their physician’s office; however best approaches for effective delivery to people with chronic disease is unknown. Our PCORI-funded Management Of Diabetes in Everyday Life (MODEL) study examines the comparative effectiveness of patient-driven text messaging (TM) and health coaching for African-American adults with uncontrolled diabetes and multiple chronic conditions in MUA in the MidSouth. Here we present the innovative, modular, biomedical informatics architecture we developed to manage the data of our vulnerable patients in a chronic disease registry, with links to the research data from patients in the clinical trial, while also serving as a platform to implement our tailored mHealth diabetes intervention.

Methods and Tools
Three main components form the informatics stack: (1) the Diabetes Wellness and Prevention Coalition registry (DWPC-R); (2) a REDCap clinical trial data center; and (3) Mosio/Twilio SMS application. The DWPC-R hosts all patient medical data from a hospital EMR, augmented by data of uncontrolled diabetics retrieved from nine minority-serving community medical practices in the Memphis, TN area. Data are collected under a HIPAA waiver using encrypted file transfer, harmonized via n-gram matching using Talend DI software, and transformed to the OHDSI common data model. The DWPC-R is used to recruit for our clinical trial. Patients randomized to the TM arm undergo a structured interview to tailor their TM program for content and frequency. All clinical trial data are hosted in a REDCap database synched with the DWPC-R. Demographic information is harvested to build patient profiles in the TM messaging application. Twilio is a cloud-communications platform for en masse text messaging. Interfacing with Twilio requires computer programming, thus we used the Twilio extension Mosio. In Mosio we create batch messaging lists from our REDCap clinical trial demographics forms. These allow us to add/remove patients from messaging groups based on their responses to the TM process. Messages exist as “stems” which Mosio concatenates (greeting + tailored-content), adding personal salutations to create coherent and tailored messages (samples to be displayed in our poster). The message libraries are stored as .csv files and hosted on a server behind our firewall. Custom Python scripts shuttle these files to the Mosio messaging API.

Discussion and Conclusion
MODEL launched in summer 2016. Given the complexity of the technologies underlying the process, it is no surprise that development of these tools has not been without difficulties. Human factors have often proven the most challenging. Working with so many clinical partners highlighted issues in synchronizing data transfer and ensuring repeatability of data feeds. This is essential if we are to automate Extract/Transform/Load (ETL) without throwing errors each time the scripts run. Our patients use TM to report their health status. To ensure the internal consistency and repeatability of our experiment we must ensure patients correctly use the response process (1/good to 5/poor on a Likert scale). Additionally, some patients use a voice-to-text messaging app and some of these responses register as text, not integers, in our database forcing us to manually curate data and increasing the risk of coding errors. We believe we can overcome this issue programmatically. The MODEL study demonstrates how a tailored mHealth intervention can be empowered by a modular informatics architecture. We hope to bundle this application stack to allow others to use our experience to better serve their own at-risk patients.
TNF-alpha Use as a Risk Factor for Lymphoma in Rheumatoid Arthritis Patients Using the Virtual Research Data Center

Gregory T. Brown, MD, PhD, Fabricio S. Kury, MD, Clement J. McDonald, MD

1National Library of Medicine, National Institutes of Health, Bethesda, MD

Introduction

Lymphomas arising in immunocompromised patients are well-documented. Patients with rheumatoid arthritis (RA) are often treated with immunomodulators such as methotrexate, and failing that, TNF-alpha inhibitors (TNFαIs) such as infliximab. Several studies have examined the association between lymphoma and TNFAI use, but with no consensus. We address this by utilizing the Virtual Research Data Center (VRDC) which, collects and de-identifies vast amounts of data generated by Medicare.

Methodology

We reviewed beneficiary data from the VRDC covering the years 2007-2014, comprising 52,876,553 patients. We used two ICD-9 codes to identify RA patients, in addition to patients with lymphoma. We also included “other” inflammatory disorders that are treated with TNFαIs, such as psoriasis, psoriatic arthritis, and ankylosing spondylitis. The VRDC contains Medicare Part D information, and from that we identified patients billed for TNFαIs as a surrogate for TNFαI use. We cross-tabulated the populations to infer associations.

Results

With unadjusted preliminary data we identified 1,610,699 patients with RA, 605,403 with lymphoma, and 983,822 with “other” inflammatory disorders. Patients with both RA and lymphoma represent 1.75% of the total RA population. Those taking TNFαIs comprised 1.4%.

Table 1. Cross tabulation of VRDC data with RA, lymphoma, and TNFαI.

Conclusion

With large numbers of patients we can explore associations that might otherwise be difficult with clinical trials. To our knowledge this is the first study using big data to examine the question of lymphoma and TNFαI use. This simple analysis does not show increased risk of lymphoma while taking TNFαIs, and our findings are bolstered by the lack of definitive evidence in the literature. However care must be taken, as always, not to mistake association with causality. Further analyses, such as Cox regression and propensity scoring is required, as well as comparison to other diseases and conditions.

References

Knowledge-Enriched Data: Realizing the Riches of Our Data

Matthew Burton, M.D. 1,2, Amy Knopp, M.S. 1, Terri Glaze1, Meghan Sandberg1, Davide Sottara, Ph.D. 1,2, Scott Finnesgard1, James Buntrock1, Daryl J. Kor, M.D. 1
1Mayo Clinic, Rochester, MN; 2Arizona State University, Scottsdale, AZ

Introduction

Data management is an end-to-end and often cyclical process, covering every step from the original capture of data at its operational source to the final presentation or delivery of that data for use by each data consumer (i.e. other operational units, research, predictive analytics, population health) and all the storage, movement, and any manipulation of the data between. Much of the data manipulations in the data management process are data enrichments that not only derive new information from raw source system data, but also includes other (recursive) processes such as the integration, correlation, normalization, standardization, and interpretation of the data 1. Moreover, many enrichments processes are driven by clinical knowledge as much as the processes that leverage the enriched data, e.g. for decision-making.

Mayo Clinic has made strategic investments in a Unified Data Platform (UDP) that brings together diverse clinical and administrative data from multiple historical and operational databases to provide data access for all non-EHR applications and capabilities (e.g. advanced analytics, research and discovery, informatics innovations). The requisite enrichments are variable across numerous different teams working with the same data, with variation impacting the ability to capitalize on vast data and knowledge assets. Therefore, an initiative was launched with the purpose to advance data enrichment processes, capabilities, and quality assurance to achieve trustworthy data with increased value and utility. The initial phase of this initiative was a qualitative analysis of current and future state through the engagement of numerous internal stakeholders and a team of acknowledged experts.

Methods

We used a mixed-methods approach including conducting semi-structured interviews designed specifically for key cohorts (front-line managers, staff, and practice and IT leadership) with thematic analysis, surveys based on key themes, focus group discussions that built on key issues identified in interviews and surveys, and a modified Delphi approach to present the collective findings to a group of relevant experts and elicit feedback and recommendations. Questions posed were designed around broad assessment of current and desired future state for data enrichment including personnel, capabilities, tooling and technology, methods and processes. Over seventy semi-structured interviews were conducted along with relevant surveys for key cohorts. Three focus groups were conducted with prior findings from inquiry presented using a modified Delphi approach. All descriptive findings were analyzed using iterative thematic analysis and multiple abstractors. Detailed findings were documented in an internal report.

Results

Results included substantial data and insights from diverse organizational perspectives that are being used to establish new organizational alignments, standardize best practices, converge, enhance, and develop new technologies, and support workforce development. Key areas highlighted by staff and leadership alike included: data quality, practice engagement, documentation and knowledge representation, knowledge elicitation and collaborative knowledge engineering. Our panel of experts proposed a roadmap and framework for knowledge-enriched data addressing best practices across technological, process, and informatics perspectives. These recommendations are being implemented as part of the UDP business plan including: a functional Clinical Concept Glossary (for indexing and resolving high value enriched data); detailed descriptions of data enrichment processes, capabilities, and data categories; a pilot for and documentation of best practices for knowledge engineering in the context of enriched data, principles and best practices for metadata and persistence of enriched data; and implementation and documentation of best practice processes and tooling perspectives for a high performing, value-enhancing 'data enrichment factory'.

References

Physicians Behavior with a Non – Overriding Alert Prescriptions CDS Dosing System Developed in Argentina

María S. Hernández Caffot, MD, PhD*1,2, Carolina Vogt, Pharm.1, Solana Muñoz Fernandez, Pharm.3, Francisco Bernabeu, MD1, Bayan V. Khorsandnia, MD*1,2
1Departamento de Informática en Salud Hospital Privado Universitario de Córdoba; 2Instituto Universitario para las Ciencias Biomédicas de Córdoba (IUCBC), – Argentina.

Introduction
Clinical decision support system (CDSS) is an important tool to reduce the medication errors and to improve the quality and safety patient care [1]. There is a wide variability in their use and in how physicians respond to provider recommendations. Investigators have reported wide variations in over-ride rates, ranging from 25% to 96% depending on the site, settings, and alert type [2]. The objectives of this work were (i) to analyze the physician behavior with a hard stop alert with non-overriding option; (ii) explore adherence versus workaround with the dose red alert.

Material and Methods
The study took place at the Hospital Privado Universitario de Córdoba, Hospital Raúl Angel Ferreyra and Clínica Privada Richieri, from Córdoba, central Argentina. The data set corresponds to a period of 5 month: August to December 2016. We develop a complex CDS dosing system of two category alert (orange and red) for the most frequently prescribed drugs (248 in total). These drugs have adjusted doses, and we take into account the age group (pediatrics and adults), and patient/drug renal function.

We analyzed all the electronic prescription during this period in the three institutions and we focus in maximum doses red alert CDS category (non - overriding option): doses higher than maximum effective dose. Then we observe the behavior of the doctor with the alert, if they accept the suggestion by modifying the dose or canceling the prescription, in this case re-indicating it by an alternative mode (free text or oral request). In addition, we divided the dose in excess of the maximum allowable dose into three categories: a- up to 100%; b- between 100 and 500%; and c- over 500%.

Results
We registered 5157 CDS alerts (3.79%) on 135,873 medication orders during the study period. Among the pediatric patients the drugs that generate the highest numbers of hard stop alerts were the antifebrile, opioids and glucocorticoid; and among adults patients the highest number of hard stop alerts were generated by antibiotics, glucocorticoid and benzodiazepines. The 19.19% (990 alerts) corresponds to red alerts. The 87% (861) of the red alerts was accepted, from which the 83% (821) of physicians decided to modify the dose and the 17% (168) decided to canceled the prescription and do not prescribe it again. The remaining 13% (129) did not accept the alert; deciding to cancel the alerts and to realize the same prescriptions in free text or oral prescription. Physicians who accepted the alert were exceeded as follows: the 47.7% (473) in up to 100%, the 23.1% (228) exceeds between 100–500% and the 29.2% (289) of them more than 500%. Finally, from those who did not accept the hard stop alert, a 60.8% (78) of physicians exceeded the maximum allowable dose up to 100%, the 19.3% (25) between 100–500%, and the 19.9% (26) was in more than the 500%. Only 0.8% had false positive results.

Conclusions
There were 990 alerts for maximum dose, and we prevented 859 medication errors. Our CDSS prevent an 87% of serious to moderate adverse effects (respiratory and Central Nervous System depressions, myelotoxicity, renal failure, and cardiovascular disorders). Nevertheless, there were a 13% of the hard stop that physician managed to avoid in which we believed they correspond to serious or moderate adverse effects. Taking into account this system failure, we are currently working to prevent physicians to realize prescriptions by an alternative way. However, our CDS dosing system show to have high acceptance within physicians from the three institutions, and low level of workaround.

References
Improving EHR Chart Review Efficiency via Semantic Relevance Assessment

Tianrun Cai, MD1,2, Andy Beam, PhD2, Stephanie Chan3, Yuk-Lam Ho, MS4, Jacqueline Honerlaw, MS4, Yichi Zhang, PhD3, David Gagnon, PhD4, Kelly Cho, PhD, MPH4, Michael Gaziano, MD, MPH, Katherine Liao, MD, MPH1,2,4, Jacqueline Honerlaw, MS4, Michael Gaziano, MD, MPH, Katherine Liao, MD, MPH1,2,4, Tianxi Cai, ScD2,3,1
1Brigham and Women's Hospital, Boston, MA; 2Harvard Medical School, Boston MA; 3Harvard University, Boston MA; 4VA Boston Healthcare System, Boston MA;

Introduction

Electronic medical records (EMR) have expanded opportunities for precision medicine research. A major barrier to fully realize the translational potential of EMR is the lack of tools for efficiently and accurately defining phenotypes. Developing phenotyping algorithms often requires labor intensive manual chart review, which is particularly challenging for healthcare systems such as the Veteran Affairs (VA) with large volumes of notes per patient. We introduce a novel unsupervised algorithm to rank the relevance of the notes to reduce the chart review burden.

Methods

Algorithm: We identified a set of concept unique identifiers (CUIs) related to the phenotype by extracting clinical terms from 5 relevant online knowledge sources including Wikipedia via the algorithm in Yu et al1. We mapped each CUI to a 300-length semantic vector (SV) obtained by performing a word-to-vec analysis of all PubMed Central articles via the “skip-gram” algorithm2. Using these CUIs as a dictionary, we performed NLP with NILE3 on each clinical note and counted the number of mentions for each CUI. We summarize the phenotype-specific semantic content of a note as a weighted average of the CUI-specific SVs with weights the standard TF-IDF weight multiplied by a prior weight \( w_{\text{prior}} \), calculated as the log-frequency of each CUI appearing in the Wikipedia article. We also obtained a reference SV as a TF-IDF weighted SV calculated for the Wikipedia article. The relevance of a note to the phenotype, denoted by \( R \), is calculated as the cosine-similarity between the reference SV and the summary SV for the note. Notes without any mention of any candidate CUI are assigned as uninformative. For remaining notes, we performed clustering on their \( R \) measures and classify them into uninformative (U), somewhat informative (S) and informative (I). Informative notes are prioritized for chart review.

Data and Metrics for Evaluation: We use the algorithm to rank 60128 notes from 130 VA patients sampled to annotate their ischemic stroke (IS) status. Chart review was performed on a random subset of 200 notes to annotate the relevance of each note to IS as informative, somewhat informative and on 20 patients to determine whether reviewing the top ranked notes was sufficient to determine the patient level IS status. We compared our method to the Latent Dirichlet Allocation (LDA) model and the Correlated Topics Model (CTM)4.

Results

The algorithm, classifying 27% notes as I and 26% as U, has a rank correlation of 64% with the gold standard annotation, while LDA and CTM yielded correlations between 18% and 32%. At the patient level, the chart review determined that the decision on the IS status can be achieved by only reviewing the top ranked notes in \( I \), which yielded about 50% reduction in review time needed per patient. Incorporating prior weights \( w_{\text{prior}} \) from Wikipedia results in better performance with higher sensitivity (95% w/ \( w_{\text{prior}} \) versus 90% w/o \( w_{\text{prior}} \)) while at the same time selecting fewer notes as informative (27% w/ \( w_{\text{prior}} \) versus 31% w/o \( w_{\text{prior}} \)).

Conclusion

Note ranking can greatly improve both the efficiency and accuracy of chart review. Incorporating prior knowledge on the disease based on semantic analysis can improve the selection performance.

References

Exploring Novel Graphical Representations of Clinical Data in a Learning EMR

Luca Calzoni, MD1, Gilles Clermont, MD MSc2, Gregory F. Cooper, MD PhD1,3, Shyam Visweswaran, MD PhD1,3, Harry Hochheiser, PhD1,3
1Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA, USA; 2Department of Critical Care Medicine, University of Pittsburgh, Pittsburgh, PA, USA; 3Intelligent Systems Program, University of Pittsburgh, Pittsburgh, PA, USA

To reduce the risks of cognitive overload associated with large amounts of data in Electronic Medical Records (EMRs), we are designing a Learning EMR that is able to draw a physician’s attention to the right patient data at the right time (1). The system, developed for use in Intensive Care Units (ICUs), uses a data-driven approach to analyze patterns of past EMR usage by clinicians and highlight the most relevant data for each patient.

In order to facilitate the development of an EMR user interface that prioritizes the presentation of high-value data, it is necessary to acquire a thorough understanding of the information practices of ICU providers. Most prior work in the literature focuses on investigating ICU clinicians’ data needs, reasoning strategies and use of information sources. A literature review provided us with useful insights on design aspects that we found applicable to the design of a Learning EMR. We identified four main themes:

1. EMRs should convey and summarize clinical information effectively, by encoding health parameters to visual attributes, and by utilizing information views that prioritize the display of high-value data, organize information into clinically relevant concepts, and provide quick overviews of the patient’s conditions.
2. EMRs should highlight changes in clinical outcomes across different data points or from an expected course: physicians prioritize cases based on level of concern. They are reassured by patients who are progressing as expected, while patients showing unexpected changes escalate on their prioritization list.
3. EMRs should offer features that facilitate comparisons across patients for clinical prioritization purposes, by helping physicians to quickly identify appropriate groupings for their patients, and by showing how a patient’s affinity towards different groupings changes over time.
4. EMRs should provide support for analytical reasoning, by visually arranging related data together, and by providing highly configurable user interfaces, smart search tools, and workspaces that make it easy to manipulate information at the level of entities and their relationships.

To inform the design of our Learning EMR user interface, we intend to combine the design principles identified above with observations and design activities. To gain a deeper understanding of information practices in ICUs, we will shadow ICU physicians and collect observational data on their information seeking activities and interactions with the EMR. Interviews conducted outside of care settings will be used to help characterize such activities and organize them into general process models, representing how ICU physicians interact with the EMR to perform care activities. We will use these models to inform the creation of a series of wireframes, identifying possible ways to highlight high-value data in our Learning EMR.

In a subsequent focus group session, we plan to actively involve ICU physicians in the user interface design process, using techniques inspired by Participatory Design. After learning about our goal to develop an EMR designed to reduce information overload in the ICU, clinicians will be asked to provide feedback about the concept. Physicians will then generate and discuss design ideas for the Learning EMR user interface, using craft materials and low-fidelity sketches. Participants will then be asked to provide their feedback on the user interface strategies for highlighting data of interest explored in our wireframes.

Study findings will help us gain a deeper understanding of ICU clinicians’ information practices and desiderata regarding EMR user interface design. Feedback on the Learning EMR concept and suggested design ideas will lead to the creation of a new series of wireframes reflecting participants’ work and preferences. These revised wireframes will inform future design choices for our Learning EMR, and eventually the implementation of a fully functional EMR user interface.

This work is supported by the NLM Institutional Training Grant in Biomedical Informatics 5T15LM007059-29.

References
A Natural Language Processing System for Biomedical Dataset Retrieval
Xiaoling Chen, PhD1, Jun Xu, PhD1, Jinqqi Wang, MS1, Anupama Gururaj, PhD1, Lucila Ohno-Machado, MD, PhD2, Hua Xu, PhD1

1University of Texas Health Science Center at Houston, Houston, TX; 2University of California, San Diego, CA

Introduction
DataMed (https://datamed.org), a data discovery index prototype, is developed by bioCADDIE for users to find datasets efficiently in the biomedical domain. To improve the search performance in DataMed via query interpretation and resource retrieval, we developed a high-performance NLP system, which can recognize several important biomedical concepts (gene, disease, drug, biological process, cell line, MeSH terms) and integrated it into the DataMed prototype. We are currently working on performing direct comparison of search results with and without the NLP module. We will report our findings on the same at the symposium in our poster.

System Development
We adopted training datasets from different resources, using the bioCreative challenge1,2 corpora, with annotations for diseases, drugs and genes. We used our challenge-proved conditional random field (CRF) algorithm3 for recognizing gene, disease, drug and cell line. We developed a dictionary-based model for recognizing the biological process with a list from the QuickGo database2 (with 109550 entities) while the corpus provided by University of Turku1 was adopted for cell lines. For evaluation, we generated a corpus of randomly selected 700 test documents from 21 repositories included in DataMed. For each selected dataset, only the dataset title and description from the metadata were extracted. Two annotators with biomedical background worked separately to annotate these 700 documents based on the annotation guideline provided to them. The inter-rater reliability was 0.4475. The observed proportions of agreement were calculated for each entity and the final number is the average of all the entities. The adjudicator reviewed annotation disagreements and a final decision was made on the discrepancy after discussion.

To remediate the discrepancy between our training and test documents issue, post-processing rules were added to the algorithms to improve the performance of the system. The entities were also encoded to the UMLS Concept Unique Identifiers (CUIs) for normalization. The encoder is built based on Vector Space Model (VSM). Of the 131 search queries from DataMed, 104 (79.39%) of them mapped to UMLS CUIs. We also integrated MetaMapLite4 into our system to recognize MeSH terms.

Table 1 shows the performance of our NLP system. Due to low number of cell line entities, we have excluded it from the results. The NLP system is implemented in two ways, as a Java program for batch processing and a web service for real-time query processing. We have integrated the system into DataMed search process, in both metadata enrichment and query expansion. The challenges and possible solutions to address the low recall will be discussed in the poster.

Conclusion
We have developed a NLP system that recognizes the biomedical entities: diseases, drugs, genes, biological processes, cell lines and MeSH terms with high precision from the metadata of biomedical datasets. The system is integrated into the DataMed search process to improve dataset retrieval.

Acknowledgements. bioCADDIE is funded by NIH U24AI117966

Table 1. NER results of the NLP system

<table>
<thead>
<tr>
<th>Gene</th>
<th>Disease</th>
<th>Drug</th>
<th>Biological Process</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>89.95%</td>
<td>92.54%</td>
<td>91.08%</td>
</tr>
<tr>
<td>Recall</td>
<td>50.75%</td>
<td>88.89%</td>
<td>69.06%</td>
</tr>
<tr>
<td>F-score</td>
<td>64.89%</td>
<td>90.68%</td>
<td>78.55%</td>
</tr>
<tr>
<td>Entities No.</td>
<td>670</td>
<td>684</td>
<td>462</td>
</tr>
</tbody>
</table>

References
4. QuickGo, https://www.ebi.ac.uk/QuickGO

1969
The Access of Bone Marrow Transplant Patients to Online Support

Ming-Yuan Chih, PhD, Jennifer Christian, BS, Kang Namkoong, PhD, Miranda Hatfield, MSW, Meghan Johnson, MPH, Rancie Hannah, BS, Jamie Studts, PhD, Robin Vanderpool, PhD, Gerhard Hildebrandt, MD, University of Kentucky, Lexington, KY, US

Abstract

An online support usage survey was completed by 30 adult blood and marrow transplant (BMT) patients at the University of Kentucky Markey Cancer Center (MCC). Results show that most patients, even with low socio-economic status, use advanced communication technologies regularly. Furthermore, these patients do not access cancer-related online support, and do want to seek health information and communicate with providers/peers online. Cancer center-sponsored online support may overcome patients’ barriers to accessing necessary support.

Introduction

BMT patients are often physically and mentally secluded during the transplant process. This isolation is an obstacle to seeking social support and clinical advice. One way to assist the patients through this process is to allow them to connect with other patients and professionals by way of an online or smartphone-based application. Existing literature in this area is mostly in pediatric BMT programs, not for adult patients. Smartphone and online support seeking among MCC patients, in particular those from rural Kentucky, is also unclear. To fill this gap, a pilot patient survey was designed to assess the adult BMT patient’s access to the Internet and online support resources.

Methods

Between 1 Sep 2016 and 6 Mar 2017, 30 pre-BMT patients at the MCC completed an online support usage survey. Survey questions include demographics and online support usage. These questions were developed based on Health Information National Trends Survey. Descriptive statistics were reported.

Results

Most survey patients are male (20/30), white (28/30) and without a college degree (19/30). Ages ranged from 23 to 71 (median: 59). Among the 30 patients, 17 were unemployed, 16 married, and 15 from rural Kentucky. The median annual household income was lower than $35,000. All but one reported having Internet access at home or elsewhere and 27 (90%) had faster Internet (cable and/or 3G/4G broadband). Twenty-seven (90%) patients had mobile devices, such as smartphones or tablets. All patients used text messaging and 80–90% of them used emails or mobile apps over the past 12 months. Twenty-three (70%) patients used social media, such as Facebook in the last 12 months while only 7 (23.3%) participated in cancer specific online support groups, such as American Cancer Society online forum. The main reasons that patients did not access cancer-related online support were because: 1) they did not know how to find an online support group (20%); 2) they did not know how to participate in an online support group (24%); or 3) they did not want to share personal information (20%). All but one expressed the desire to access health-related information and to communicate with their providers and peer patients online.

Conclusion

Patients referred for BMT at MCC have adequate personal access to information technology and are familiar with emails, text messages, and social networking sites. Their access to cancer related online support could be limited due to their unfamiliarity and discomfort of using these resources. A cancer center-sponsored, provider-facilitated application may be able to ease the discomfort and to improve the motivation to access available cancer information and social support online. Future efforts to connect patients with their healthcare providers and peer patients may enhance ongoing care and support to these patients.

Acknowledgment: The authors appreciate the support from NCI Cancer Center Support Grant (P30 CA177558).

References

Usability Evaluation of a Prototype mHealth App for Symptom Self-Management in Underserved Persons Living with HIV

Hwayoung Cho, PhDc, MSN, RN, Hestia Rojas, MPH, Casey Fulmer, MPH, Rebecca Schnall, PhD, MPH, RN-BC
School of Nursing, Columbia University, New York, NY, USA

Abstract Symptom management is essential for PLWH. mHealth apps are a potentially useful delivery mode of symptom self-management strategies. We developed a prototype mHealth app for symptom self-management for PLWH and tested its usability in a laboratory setting via end-user testing and a heuristic evaluation with HCI experts. The prototype was iteratively refined; further evaluation will be conducted in a 3-month feasibility trial.

Introduction
Persons living with HIV (PLWH) are confronted with persistent symptoms related to the disease, medication side effects and comorbidities making symptom management essential for improving quality of life. A team at UCSF developed a paper-based symptom management manual with self-care strategies for 21 common HIV symptoms. As mHealth apps can reduce challenges associated with racial/ethnic disparities by increasing access to health information, apps have the potential to be an effective delivery mode of symptom self-management strategies in underserved PLWH, yet no apps for symptom management currently exist. We developed a prototype app (mVIP) for symptom self-management in PLWH by incorporating these strategies into a mHealth tool. The purpose of this work was to evaluate the usability of the prototype app from both expert and end-user perspectives in a laboratory setting.

Methods
We conducted a usability evaluation of the prototype mVIP through end-user testing with 20 PLWH (10 using Android and 10 using iPhone) and a heuristic evaluation with 5 experts in human-computer interaction (using Morare software). The usability evaluation incorporated a think-aloud protocol of use case scenarios and the process was audio-recorded. The app’s usability was rated by end-users via a post-study system usability questionnaire (PSSUQ) and by experts via a heuristic evaluation checklist.

Results
The app’s content, functionality and interface were refined. Based on end-users’ recommendations, an error message ‘Check your credentials’ was changed to ‘Username or password you entered is not valid. Try again’ since the term of credentials was unfamiliar to several participants. We added instructions for ‘How our app works’ on the home page to help participants more easily use the app. The username was simplified from the user’s email address to a short code for easier text entry. The overall PSSUQ scores ranged from 1.00 to 2.88 (from 1-best to 7-worst), reflecting strong user acceptance of the app. Based on experts’ recommendations, the term ‘Dashboard’ was changed to ‘VIP Home,’ which was a more familiar term to end-users. To make the app’s functionality more generalizable, we added a response option of ‘Didn’t try’ in addition to ‘Yes/No’ options for the helpfulness assessment question for each of strategies suggested: if users select ‘Didn’t try’ of the strategy previously suggested, one more chance to try would be provided. The mean scores and sample comments resulting from the heuristic evaluation are listed in Table 1.

Conclusion
Based on the findings from this usability evaluation in a laboratory setting, a prototype mVIP was iteratively refined. Further evaluation will be conducted in a real world setting during a 3-month feasibility trial (n=80). Findings from this work have the potential to reduce health disparities and improve outcomes.

Acknowledgements This study was supported by the Agency for Health Research and Quality (R21HS023963; PI: R. Schnall).

Table 1. Mean Scores and Sample Comments from Heuristic Evaluation

<table>
<thead>
<tr>
<th>Usability Factor</th>
<th>Mean (SD)</th>
<th>Sample Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Visibility of System Status</td>
<td>1.60 (1.14)</td>
<td>Headers not very noticeable</td>
</tr>
<tr>
<td>Match between System and the Real World</td>
<td>0.80 (1.10)</td>
<td>Collapse strategies into the relevant symptoms and organize the symptoms depending on the severity</td>
</tr>
<tr>
<td>User Control and Freedom</td>
<td>2.20 (1.10)</td>
<td>Should have a home button on every single page so that users have the ability to quit/log off/keep the questions</td>
</tr>
<tr>
<td>Consistency and Standards</td>
<td>0.80 (1.10)</td>
<td>Add the function of avatar selection and keep showing it consistently on every page</td>
</tr>
<tr>
<td>Help users Recognize, Diagnose, and Recover From Errors</td>
<td>0.40 (0.89)</td>
<td>Some error messages repeated; put the red color for ‘Password updated failed’</td>
</tr>
<tr>
<td>Error Prevention</td>
<td>0.40 (0.89)</td>
<td>Not sure how to go back</td>
</tr>
<tr>
<td>Recognition Rather Than Recall</td>
<td>0.80 (0.84)</td>
<td>No instructions on how the app works</td>
</tr>
<tr>
<td>Flexibility and Efficiency of Use</td>
<td>1.40 (1.34)</td>
<td>Have an option of download/email on the ‘Your history’</td>
</tr>
<tr>
<td>Aesthetic and Minimalist Design</td>
<td>0.60 (0.89)</td>
<td>Avatars too big, rather strategies with larger font size</td>
</tr>
<tr>
<td>Help and Documentation</td>
<td>1.20 (0.84)</td>
<td>No help function apart from the log-in page</td>
</tr>
</tbody>
</table>

Reference
Evaluation of OMOP Vocabulary on Transplant Registry Data

Sylvia Cho, MHS\(^1\), Margaret Sin, BS\(^1\), Karthik Natarajan, PhD\(^1\)
\(^1\)Department of Biomedical Informatics, Columbia University, New York, NY

Introduction

Organ transplant is increasingly becoming a treatment option to improve the quality of life and survival rate for patients with organ failure. Thus, continuous efforts to improve the outcomes of transplant recipients within and across transplant centers is significantly important. The Center for Medicare and Medicaid Services (CMS) currently uses data from the United Network for Organ Sharing (UNOS) to track and monitor transplant outcomes. While UNOS registry data captures a wide range of baseline data, it lacks longitudinal patient-level data and important determinants of patient outcome\(^1\). Thus, augmenting the UNOS data with electronic health record (EHR) data can substantiate the outcomes analysis. However, barriers to interoperability exist due to disparate data models and concept representations among different transplant centers. While the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) is considered as an optimal infrastructure to integrate disparate observational databases, limitations such as potential data loss may exist.\(^2,3\) To our knowledge, there has been no assessment on whether the concepts in the UNOS registry database can be mapped to the OMOP standard concepts. Thus, in this preliminary study, we aim to evaluate the coverage of the OMOP vocabulary on UNOS data elements in order to build a gold standard mapping.

Methods

The UNOS data entry system collects transplant-related data through forms submitted by transplant centers. In this study, we used the Transplant Candidate Registration (TCR), Transplant Recipient Registration (TRR), and the Transplant Recipient Follow-up (TRF) forms which include the major transplant-related data for all organs. We extracted unique concepts from these forms and its corresponding look-up table concepts. We mapped the UNOS source concepts to OMOP standard concepts using the USAGI tool, an open-source application developed by the Observational Health Data Sciences and Informatics (OHDSI) community.\(^4\) Two researchers with biomedical informatics domain knowledge respectively mapped all concepts. We measured the content coverage of OMOP standard concepts on the UNOS source concepts by examining the percentage of UNOS source concepts which could not be mapped to the OMOP standard vocabulary coding system. The reliability of concept mapping was measured through the Cohen’s Kappa score which shows the degree of agreement between investigators.

Results

A total of 1,203 unique source concepts from UNOS were reviewed. The first and second annotator successfully mapped 554 (46\%) and 508 (42\%) source concepts to OMOP standard codes, respectively. A Kappa of 0.79 was calculated based on whether a UNOS element was mapped to an OMOP concept or not. Details of our results are presented in Table 1.

<table>
<thead>
<tr>
<th></th>
<th># of source concepts</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Both annotators could not map the source concept to an OMOP source code</td>
<td>609</td>
<td>50.62%</td>
</tr>
<tr>
<td>Both annotators mapped the source concepts to the same OMOP source code</td>
<td>342</td>
<td>28.43%</td>
</tr>
<tr>
<td>Both annotators mapped the source concepts, but to different OMOP source codes</td>
<td>126</td>
<td>10.47%</td>
</tr>
<tr>
<td>Annotator 1 mapped the concept, but Annotator 2 did not</td>
<td>86</td>
<td>7.15%</td>
</tr>
<tr>
<td>Annotator 1 did not map the concept, but Annotator 2 did</td>
<td>40</td>
<td>3.33%</td>
</tr>
</tbody>
</table>

Table 1. Results on mapping UNOS registry source concepts to OMOP standard concept codes

Conclusion

The results of this study show that approximately half of the UNOS source concepts could not be mapped to OMOP standard concepts. Most of the unmapped concepts are related to transplant-specific clinical data and patient-reported outcomes which shows the need to extend OMOP vocabularies in these areas. We plan to involve clinicians to validate our results in order to develop a gold standard mapping between OMOP and UNOS data. This mapping will enable cross-institutional, transplant-specific outcomes analyses.

References

Use of Storyboarding in the Development of a Pain Management Mobile App: A Case Study

Jeeyae Choi, PhD, RN

1University of North Carolina Wilmington, Wilmington, NC, USA

Abstract
New mobile technologies are enabling care providers to deliver more effective and convenient care to patients. In order to develop a science-based mobile app, it is essential to use a guideline and communicate knowledge from that guideline accurately to a developer. During the development of a pain assessment mobile app, storyboarding was used to communicate knowledge from the guideline among research team members and its usefulness and benefits were explored.

Introduction
Mobile-health-based technology has grown exponentially due to its ubiquitous computing capability. Patients and care providers are increasingly using health-related apps to educate patients, to prevent and/or treat illness\(^1\). Pain management apps are no exception; however, studies have found a gap between the science of pain and commercially developed apps. In order to develop a science-based mobile app, it is necessary to translate and then communicate pain management guidelines accurately to the developer. Storyboarding is a popular technique used in filmmaking production\(^2\). Pictorial representations of story components are prepared and organized sequentially; each picture shows graphic representation as well as explanatory texts. However, the storyboarding technique hasn’t been used in health informatics research. In this project, the storyboarding technique was used to communicate translated knowledge from pain management guidelines. The purpose of this study was to explore the usefulness and benefits of storyboarding as a communication tool when developing a pain management mobile app.

Steps of knowledge translation and communication

**Step I:** Concepts related to pain management tasks were extracted from the National Cancer Institute recommendations for health professionals (Cancer Pain PDQ\(^5\)-health professional version). These concepts were then converted into activity diagrams in the Unified Modeling Language, representing the logics of the app. An expert oncology nurse was recruited to validate and modify the process of assessing pain management in the activity diagrams.

**Step II:** Using the National Cancer Institute recommendations for patients (Cancer Pain PDQ\(^5\)-Patient version), pain assessments and associated sub-assessments were categorized into general pain information, pain level, and past and current pain treatments, and patient’s goals and choices. The expert oncology nurse validated a table that categorized these pain assessments.

**Step III:** Two artifacts, activity diagrams and categorized pain assessments, were shared with the developer. The developmental team produced storyboards and the PI accepted or requested modification of storyboards. These iterative process was done three times and 17 storyboards were subsequently produced. Figure 1 is an example of a storyboard.

**Discussion**
Seventeen storyboards successfully visualized knowledge translated from pain management guidelines. Each storyboard is a pictorial representation of knowledge with explanatory text. With this visual representation, the researcher and the app developer were able to organize pain assessment knowledge from the guidelines without ambiguity into a pain management mobile app. The storyboarding process proved beneficial before production began because the 17 storyboards provided the structure and flow of the design of the mobile app. This is the first step in an ongoing study. After production and usability testing is complete, the true usefulness and benefits of storyboarding will be seen.

References
Validating a Single-Item Stress Scale for mHealth Interventions

Wonchan Choi, PhD\textsuperscript{1}, Sherry L. Pagoto, PhD\textsuperscript{2}, Edwin D. Boudreaux, PhD\textsuperscript{2}, Bengisu Tulu, PhD\textsuperscript{1}

\textsuperscript{1}Worcester Polytechnic Institute, Worcester, MA; \textsuperscript{2}University of Massachusetts Medical School, Worcester, MA

Introduction

This poster presents the results of an online survey that aims to validate a single-item stress scale that can be used in health care interventions targeting stress management, especially in the mHealth context. Our scale has six anchors with numerical values and visual indicators—faces corresponding to each of the six levels of stress in the scale. We adopted the definition of stress and question wording used in a validated single-item measure of stress symptoms: “Stress means a situation in which a person feels tense, restless, nervous or anxious or is unable to sleep at night because his/her mind is troubled all the time. Are you experiencing this kind of stress currently?”

Methods

Based on a pilot study we conducted in 2016\textsuperscript{2}, we made minor modifications to some of the visual indicators that seemed relatively less intuitive for users to distinguish the indented levels of stress—we adjusted the angles of eyebrows in the faces for “A Little Stress” and “A Whole Lot of Stress” (Table 1). We used an online crowdsourcing site, Amazon Mechanical Turk, to recruit large samples of participants. Of the 315 participants, 56.8% were males. In terms of racial background, a majority of them (76.8%) were White Caucasians. Each participant answered an online survey where they matched six different visual indicators (faces) we created with six anchors of stress. Participants were then asked to rank seven stressful events, which were selected from a validated stressor scale\textsuperscript{3}, in the order of intensity of stress and rate each of them using our stress scale.

Results

Overall, 80.3% (253 out of 315) of the participants matched all the six faces to the corresponding levels of stress as intended. In particular, as presented in Table 1, nearly the 90% or greater proportion of the participants successfully matched each of the faces to the intended levels of stress. A Spearman’s rank order correlation between the rank orders of and the ratings on the seven stressful events ($r_s = - .773, n = 2205, p < .001$) indicated a good concurrent validity of our single-item stress scale—the higher the stressful events were ranked, the more likely they received higher ratings on the stress scale.

<table>
<thead>
<tr>
<th></th>
<th>Visual Indicators (Faces)</th>
<th>Corresponding Labels for Stress Anchors</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>No Stress</td>
<td>295 (93.7%)</td>
</tr>
<tr>
<td>1</td>
<td>A Little Stress</td>
<td>288 (91.4%)</td>
</tr>
<tr>
<td>2</td>
<td>A Little More Stress</td>
<td>296 (94.0%)</td>
</tr>
<tr>
<td>3</td>
<td>Even More Stress</td>
<td>280 (88.9%)</td>
</tr>
<tr>
<td>4</td>
<td>A Whole Lot of Stress</td>
<td>283 (89.8%)</td>
</tr>
<tr>
<td>5</td>
<td>Worst Stress</td>
<td>301 (95.6%)</td>
</tr>
</tbody>
</table>

Conclusion

We will further validate our stress scale, especially in the mobile context. In a following study, we will examine the visual indicators whether they can capture users’ stress levels consistently and accurately without the numerical values or text labels (or both), which will be particularly useful for space-efficient mobile platforms.

References

Creating a Framework to Standardize Data Extraction from Electronic Health Record Systems for Researchers and to Support Distributed Queries

Mera Choi, JD, MPP, MPM1, Nagesh (Dragon) Bashyam2, Daniella Meeker, Ph.D3, Eliel Oliveira MS, MBA4

1Office of the National Coordinator for Health Information Technology, Washington, DC; 2Drajer LLC, Boyds, MD; 3pSCANNER, San Diego, CA; 4REACHnet, New Orleans, LA

Problem
Researchers and research networks face time-consuming challenges extracting and transforming patient data from multiple data sources, including electronic health records and data warehouses.

Objective
The Office of the National Coordinator for Health Information Technology (ONC) developed a Data Access Framework (DAF), a three phase effort to identify, test, and validate standards necessary to access and extract data from within an organization’s health IT systems, external organization’s health IT systems, and health IT systems across multiple organizations. The third phase, DAF for Research (DAF-R), is designed to increase access to patient data using standards for data queries, application programming interfaces (APIs), or services. Three organizations, REACHnet, Lincoln Peak, and pSCANNER, piloted DAF-R to enable researchers to create value out of complex data from multiple sources, without having to rely on existing access paths.

Methods
(1) Clinical Data Research Networks (CDRNs) use standardized data extraction mechanisms from multiple clinical data sources to populate data marts; (2) CDRNs publish standardized data mart metadata about source data, useful for researchers to determine the available data; (3) researchers use standardized query distribution mechanisms to query multiple data marts; (4) CDRNs use standardized aggregate query results to return data from data marts back to researchers in response to a query. The DAF-R pilots successfully: (1) implemented an API on an example FHIR® framework and demonstrated conformance; (2) mapped data elements from a data model; (3) aligned metadata management standards to be consistent with national and international standards; (4) automated queries at regular intervals.

Results
By standardizing access to patients’ data, DAF-R: helps researchers reduce the burden of customary data mapping for each data source; helps researchers standardize the metadata collected from the data source; enables the use of common security and privacy controls across networks; and allows researchers the ability to consume meaningful results from various research sites. DAF-R advances research efforts to develop an interoperable data network infrastructure to maximize efficiency, advances research opportunities, and improves future health policies. This project was funded by Patient-Centered Outcomes Research (PCOR) Trust Fund, managed by the Office of Assistant Secretary for Planning and Evaluation (ASPE) in the U.S. Department of Health and Human Services.

References
OMOP on FHIR as an Enabler for Analytics-As-A-Service

Myung Choi, PhD, Richard Starr, and Jon Duke, MD
Georgia Institute of Technology, Atlanta, GA, USA

Introduction

While advances in machine learning have greatly improved the performance of predictive models in healthcare, less progress has been made in facilitating the delivery of analytics at the point of care. As part of Georgia Tech’s Health Data Analytics Platform, we developed an interoperable framework to connect OMOP-based predictive models to clinical decision support systems via FHIR. This builds on our previous OMOP on FHIR project, which defined mappings between the Observational Medical Outcomes Partnership Common Data Model (OMOP CDM) and FHIR resources. In this poster, we will describe the new OMOP on FHIR (www.omoponfhir.org) architecture and provide example use cases for deployment.

OMOP on FHIR Architecture

The OMOP on FHIR architecture (Figure 1) provides mappings and workflow for reading and writing between FHIR and OMOP CDM. We decoupled reading and writing transaction flows to improve the data integrity and quality. For writing, we introduce staging tables for FHIR contents to be natively mapped before the ETL processes are triggered for transforming these staging tables to OMOP CDM. Several sub-operations can take place during these processes based on application needs. The ETL processes provide the auditing and quality control steps to ensure data are persisted into the proper locations and mapped appropriately. Once data are persisted into the OMOP CDM, we gain access to the Observational Health Data Sciences and Informatics (OHDSI) resources that can perform data aggregations and packages for cohort creation and various population level data analytics. As we are mapping the FHIR contents to OMOP CDM, any EHRs that can understand this data model can take advantage of our platform.

As some of data elements in FHIR cannot be one-to-one mapped from OMOP CDM, SQL views are used to map the OMOP CDM tables to the appropriate FHIR resources for read requests. By constructing FHIR-friendly views, the FHIR API layer can be implemented relatively easily and is more robust to changes in the OMOP model.

Analytics-As-A-Service

The OHDSI software provides various analytic tools that minimize the overhead for data scientists. These tools provide robust analytics capabilities but are limited in their real-world impact on clinical care. With OMOP on FHIR, any data analytics process can be turned into service (‘analytics-as-a-service’) for delivery at the point-of-care. Figure 2 depicts this process.

As a test case, we deployed a predictive model for assessing risk of C. Difficile colitis infection based on individual patient characteristics. We set up an API gateway that received test patient data via FHIR, transformed to OMOP, called the C. Diff predictive model, and returned the personalized risk prediction. Total transit time for ingestion, mapping, and result generation averages 5-6 seconds with mid-end virtual machines.

Conclusion

OMOP on FHIR is an open-source platform that provides bidirectional mapping processes between OMOP and FHIR as well as a framework for API-based analytics as a service using OMOP-based predictive models. Our future work includes expanding mappings and evaluating several OMOP-based predictive models for impact on clinical decision-making. We hope that tools such as OMOP on FHIR, which connect research analytics to the point-of-care, will advance the goals of precision medicine while allowing innovation in analytics.

References

Reproducibility in Biomedical Natural Language Processing

K Bretonnel Cohen, PhD1,2, Aurélie Névél, PhD2, Jingbo Xia, PhD3, Negacy Hailu, MSc1, Cyril Grouin, PhD2, Lawrence Hunter, PhD1, Pierre Zweigenbaum, PhD2

1 CNRS, LIMSI, Université Paris-Saclay, F-91405 Orsay; 2University of Colorado School of Medicine, Denver, USA; 3Huazhong Agricultural University, Wuhan, China

Introduction

There is a growing concern that the reproducibility of much work in science in general and in the biomedical field in particular is questionable. In order to improve the robustness of methods in biomedical Natural Language Processing (NLP), we need to assess the current situation in the field, after recent experiments on a shared task suggest reproducibility cannot be taken for granted even under favorable conditions where data, code and evaluation toolkits are available1. Nonetheless, data and code availability is the first requirement to enable reproducibility.

Methods

Herein, we apply a protocol previously put forth to assess reproducibility in computer science2: all 29 articles in the proceedings of the 2016 BioNLP workshop are reviewed by two analysts (one expert in the field of biomedical NLP and one computer scientist) to extract information on code and data availability. The article passages supporting the evidence of availability are marked. However, the retrieval of the material and actual reproduction of experiments are not attempted at this stage, but would be impossible without data and code. The annotation guidelines are available at https://github.com/KevinBretonnelCohen/InterRaterAgreementReproducibility. They were created by writing an initial draft, doing a pilot annotation project, modifying the guidelines based on findings of the pilot project, and then doing the full annotation as described above. Inter analyst agreement is computed in terms of Cohen’s kappa3 in order to assess the reliability of the annotation process as well as the understanding of the task by the analysts.

Results and Conclusion

Although 48% of papers provided pointers to data and 61% provided pointers to code, only 21% reported both. Inter-analyst agreement was 0.57 (moderate) for identifying data and 0.63 (substantial) for identifying code. This suggests that reports of data availability might require more domain knowledge to be identified, e.g. a link to github might be easier to follow as evidence of code availability than the name of a dataset possibly unknown to the reader. In addition, some disagreement on data availability occurred for subsets of MEDLINE where the expert considered data as not available if the specific subset and annotations were not explicitly linked. Although the kappa scores that we are reporting for the inter-analyst agreement are not high, those numbers badly underestimate the reliability of the analysis, as the calculation sets the value for expected agreement too high, resulting in an inappropriately low kappa value. The situation with respect to reproducibility in clinical natural language processing is arguably better than the situation in computer science, as Collberg et al. found that only 13% of papers in their study reported both code and data availability. However, we believe reproducibility in biomedical NLP should receive increased attention, with the introduction of guidelines for reporting work in a reproducible way. While the availability of code and data do not necessarily guarantee that work can be reproduced, it is a necessary condition and as a community we should strive to facilitate access to code, data and experimental set-up as a way to address reproducibility issues.

Acknowledgements

This project was funded in part by the EU H2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 676207 and by the National Natural Science Foundation of China (Grant No. 61202305).

References

The Impact of Health IT Adoption: Are We Measuring the Right Outcomes?

Tiago K Colicchio¹, MBA, MS, Guilherme Del Fiol¹, MD, PhD, Watson A Bowes III¹,², MD, MS, Julio C Facelli¹, PhD, Debra L Scammon³, PhD, Scott P Narus¹,², PhD

¹Biomedical Informatics, University of Utah, Salt Lake City, UT; ²Medical Informatics, Intermountain Healthcare, Salt Lake City, UT; ³David Eccles School of Business, University of Utah, Salt Lake City, UT

Introduction

Electronic Health Record (EHR) systems adoption has rapidly increased in the U.S. partially due to financial incentives provided by the Meaningful Use program, and, as a result, the number of studies on the impact of EHR implementations on health care quality, productivity and safety, has also increased. Yet, previous studies have produced mixed-results, failing to increase our understanding of the full impact of health IT adoption. A contributing factor to this gap is the use of simple, non-standardized measurements in evaluations of the impact on health care from health IT adoption. In this study, we compare the most common measures in previous studies assessing the impact of health IT with measures suggested by experienced health care leaders and informaticists. Our goal was to determine if measures in prior studies provide a comprehensive coverage of processes likely impacted by EHR implementations.

Methods

We used a multi-method approach to identify relevant measures for monitoring health IT interventions, including implementations of new EHRs or new versions of current EHRs. We conducted semi-structured interviews with clinical and administrative leaders from Intermountain Healthcare. The leaders were identified via referrals from a convenience sample of medical informaticists. We combined their suggestions with the most commonly used measures from the literature¹, and classified them into the categories quality, productivity and safety using our previously published taxonomy⁷. The measures were included in two online surveys that were sent to informaticists nationwide. One survey included measures of ambulatory outcomes, and the other of hospital outcomes. Respondents were asked to provide suggestions of additional measures not included in our original lists. Survey invitations were sent to members of the AMIA Implementation Forum and HIMSS Nursing Informatics Alliance. We also invited research faculty with interest in EHRs or health IT adoption from several U.S. informatics academic programs.

Results

Thirty Intermountain Healthcare leaders suggested 63 measures. One-hundred twelve informaticists participated in the surveys and suggested 11 additional measures. We combined measures identified in previous studies with experts’ suggestions, producing an inventory of 102 measures. The experts suggested several measures not reported in the literature, including 13 of quality, 32 of productivity, and 16 of safety care processes (Figure 1).

Discussion

Experts suggested additional measures that have not been reported in the literature, especially for productivity and safety. These care processes deserve more attention from the informatics community. To increase our understanding of the impact of health IT adoption, we recommend future studies covering a wider range of quality, productivity and safety outcomes. Two separate studies are in development to (1) assess informatics experts’ perceptions of the relevance of these measures, and (2) develop a methodology to detect the impact of EHR implementations over time.

Acknowledgement

Intermountain Healthcare, Salt Lake City, UT, supported this project. JCF has been partially funded by National Center for Advancing Translational Sciences under Award Number 1ULTR001067.

References

ProjectFlow: Configurable Clinical Trial Management with Enterprise Data Integration and Point-of-Care Study Support

Ryan C. Cornia, BS¹², Nilla Majahalme, BSE³, Danne C. Elbers, MS³, Svitlana Dipietro, BS³, Brian Ivie, BS¹², Brad Adams, MS¹², Ramana Seerapu, BS¹², Valmeek Kudesia, MD, MS¹³, Scott L. DuVall, PhD¹²

¹VA Salt Lake City Health Care System; ²University of Utah, Salt Lake City, UT; ³VA Boston Healthcare System, Boston, MA, ⁴Boston University, MA

Abstract
Clinical trials offer tremendous value to medical research, but traditionally require extensive manual tracking or expensive and rigid IT investment for integrated electronic record tracking. We present the architecture that supports complete clinical trial tracking according to trial-defined requirements while allowing for easy integration with clinical data and services. The core of the system is ProjectFlow, a web application that allows users to define trial workflows, incorporate clinically relevant data, and integrate enterprise services.

Introduction
Tracking and managing clinical trial data has typically been an expensive and time-consuming task. While both open- and closed-source systems exist, they focus on electronic data capture (EDC) and clinical trial management (CTM), functionality is limited, and integration points are difficult. This forces study designers to fit their study to the features of the system by reducing integration with other enterprise systems and forcing study workflows to meet only those supported by third party EDC/CTM solutions.

Flexible data collection and workflow configuration can efficiently support a broad range of trials, including novel trial designs like point-of-care clinical trials.[1] Users should see only the patient data pertinent to the task they are completing, and workflows should be flexible enough to accommodate different paths through the clinical trial based on enterprise service input along the way. Clinical trial design needs to be flexible and user-adaptable to match varying requirements of clinical trials. ProjectFlow builds upon best-of-breed open source projects with a novel approach to accessing clinical data that streamlines clinical trial development and implementation in the Department of Veterans Affairs (VA).

System Description
ProjectFlow solves the complexity of clinical trials differently than existing EDC/CTM systems by:

• Allowing graphical workflow definition by clinical trial staff using the standards-based Business Process Model and Notation. No IT support is required to design, modify, or launch a trial and workflows are decoupled from data.
• Data is pulled directly from the existing enterprise relational database with no ETL or modification needed.
• Clinical trial designers create views of the data pertinent to each specific task in the workflow. Staff see a view of the data for a task that efficiently communicates the information they need to complete the task.
• Workflows can easily integrate with enterprise services to make decisions about how to route the workflow or to show specific selectable data customized to the participant.
• Patient paths can support a high level of complexity if needed, including decision points based on enterprise services or upstream data.

Conclusion
By allowing for flexible clinical trial design, investigators and trial staff have been able to implement complex clinical trial workflows that integrate with enterprise services at many key points along the way. Because of the flexibility of the system, trial staff are able to create and deploy trials, and modify existing trials as needed, all without developer support. The self-service nature has allowed VA to meet the changing requirements of existing trials and add new clinical trials in an efficient manner not previously possible.

Acknowledgements
This work was supported using resources and facilities at the VA Salt Lake City Health Care System with funding from VA Informatics and Computing Infrastructure (VINCI), VA HSR HIR 13-457.

References
An International Comparison of High-priority and Low-priority Drug-drug Interactions in Different Electronic Health Record Systems

Pieter Cornu, MPharm, PhD\(^1\), Shobha Phansalkar, RPh, PhD\(^2,3\), Diane L. Seger, RPh\(^2,4\), Insook Cho, RN, PhD\(^5\), Sarah Pontefract, MPharm\(^6\), Alexandra Robertson\(^7\), David W. Bates, MD, MSc\(^2,3,8\), Sarah P. Slight, MPharm, PhD, PGDip\(^2,9,10\)

\(^1\) Clinical Pharmacology & Clinical Pharmacy, UZ Brussel, Brussels, Belgium, \(^2\) The Centre for Patient Safety Research and Practice, Division of General Internal Medicine and Primary Care, Partners Healthcare, Boston, MA, USA, \(^3\) Harvard Medical School, Boston, MA, USA, \(^4\) Partners Healthcare, Wellesley, MA, USA, \(^5\) Department of Nursing, Inha University, Incheon, Republic of Korea, \(^6\) School of Pharmacy, Institute of Clinical Sciences, University of Birmingham, Edgbaston, Birmingham, UK, \(^7\) Tufts University School of Medicine, Boston, MA, USA, \(^8\) Harvard School of Public Health, Boston, MA, USA, \(^9\) Division of Pharmacy, School of Medicine Pharmacy and Health, Durham University, UK, \(^10\) Newcastle upon Tyne Hospitals NHS Foundation Trust, Newcastle, UK

Introduction

Drug-drug interactions (DDIs) have the potential to cause serious harm. Electronic health record (EHR) systems with integrated decision support can alert physicians to potentially dangerous DDIs during the prescribing process. However, these alert warnings are often overridden by prescribers, with override rates in excess of 80%. Previous studies have identified 15 high-priority DDIs which should always generate warnings in all EHRs\(^6\). 33 low-priority DDIs have also been identified that do not merit interruptive alerting\(^6\). The objective of this study was to investigate whether alert warnings for the high-priority and low-priority DDIs existed in five different international EHRs, compare the severity level assigned to them, and establish the rate at which they were overridden.

Methods

This was a comparative, retrospective, multinational study of five EHRs from the U.S., U.K., Republic of Korea and Belgium.

Results

Of those previously defined 15 high-priority DDIs, alert warnings were found to exist for 11 in the Korean and UK systems, 9 in the Belgian system, and all 15 in the US systems. Of 742 individual drug-drug combinations included in those 15 class-interactions, 619 (83.4%) were included in the Belgian system, 462 (62.3%) in the US inpatient and 441 (59.4%) in the US outpatient system, 212 (28.6%) in the UK system, and 53 (7.1%) in the Korean system. Alert warnings were active for all DDI combinations contained in the US and Korean EHRs, but only 8.4% and 52.4% in the Belgian and UK EHRs, respectively. The override rates for high-priority alerts requiring provider action ranged from 56.6% to 85.6%. Of those 33 previously defined low-priority DDIs, alerts existed only in the US systems for three class-based DDIs and the majority were non-interruptive (no provider action required). The inpatient and outpatient US systems generated 38 and 237 interruptive alerts, with override rates of 57.9% and 66.7%, respectively. The UK and Korean systems alerted on none of the 33 low-priority DDIs, the Belgian system included 24 inactive low-priority DDI alerts.

Conclusions: Alert warnings existed for the majority of high-priority DDIs in the different EHRs although it was easy to override some of the most severe interactions in most systems. Alerts were delivered for only a few low-priority DDIs. This study highlights a lack of consistency in how DDIs are configured in different knowledge bases. Future efforts should concentrate on enforcing an international standard for high risk DDIs and implementing strategies to minimize overriding of severe drug interaction alerts while balancing over- and under-alerting.

References


1980
Automated Image Quality Assessment for Fundus Images in Retinopathy of Prematurity

Aaron S Coyner1,2, Ryan Swan1,2, Jayashree Kalpathy-Cramer1,3, Sang Jin Kim1,4, J. Peter Campbell1, Karyn E. Jonas5, Susan Ostmo6, R.V. Paul Chan6, Michael F. Chiang1

1. Ophthalmology, Oregon Health & Science University, Portland, OR, United States.
2. Department of Medical Informatics and Clinical Epidemiology, Oregon Health & Science University, Portland, OR, United States.
3. MGH/Harvard Medical School, Boston, MA, United States.
4. Ophthalmology, Sungkyunkwan University School of Medicine, Seoul, Korea.
5. University of Illinois at Chicago, Chicago, IL, United States.
6. Ophthalmology, Illinois Eye and Ear Infirmary, Chicago, IL, United States.

Accurate image-based ophthalmic diagnosis relies on clarity of fundus images. This has important implications for the quality of ophthalmic diagnosis, and for emerging methods such as telemedicine and computer-based image analysis. The purpose of this study is to implement an algorithm for automatically assessing the quality of fundus images, and to evaluate its performance on a set of Retinopathy of Prematurity (ROP) images compared to expert assessment.

A data set of 30 wide-angle ROP images (RetCam; Natus, Pleasanton, CA) was collected from infants undergoing routine screening examinations. A previously published method by Bartling et. al for automated assessment of digital fundus images was implemented. Briefly, 30 images were resized and converted to grayscale. Each image was broken into 16x16 pixel blocks, and blocks were assessed using metrics for brightness and sharpness. Scores for each block were used to compute an overall image quality score for each image. Images were ranked by quality score from 1 (lowest quality score) to 30 (highest quality score). Six independent experts were provided the same set of 30 images and, using a web-based interface, were prompted to "select the higher quality image for diagnosis of plus disease." Elo rating values for each image were determined for each rater and for the consensus of all raters. Spearman rank correlation was used to assess the similarity of the algorithm’s performance to human raters when determining image quality.

Individual expert ranks correlated highly with one another (correlation coefficient [CC] 0.89-0.94) and with the consensus rank (CC 0.94-0.98). The algorithm rank correlated with individual expert ranks (CC 0.81-0.87) and the consensus rank (CC 0.86), but not as highly as individual experts (Table 1).

Experts in this study had high correlation to one another in assessing quality of ROP images. The computer-based algorithm assessed image quality with strong correlation to the expert consensus rank, although not as high as individual expert correlation with the consensus rank. Future studies will investigate discrepancies between the algorithm and the consensus, using the established consensus rank as a reliable training set.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1.00</td>
<td>0.86</td>
<td>0.94</td>
<td>0.94</td>
<td>0.94</td>
<td>0.96</td>
<td>0.96</td>
<td>0.96</td>
</tr>
<tr>
<td>0.86</td>
<td>1.00</td>
<td>0.98</td>
<td>0.94</td>
<td>0.94</td>
<td>0.96</td>
<td>0.96</td>
<td>0.96</td>
</tr>
<tr>
<td>0.98</td>
<td>0.92</td>
<td>1.00</td>
<td>0.90</td>
<td>0.94</td>
<td>0.93</td>
<td>0.93</td>
<td>0.93</td>
</tr>
<tr>
<td>0.94</td>
<td>0.81</td>
<td>0.90</td>
<td>1.00</td>
<td>0.94</td>
<td>0.96</td>
<td>0.92</td>
<td>0.95</td>
</tr>
<tr>
<td>0.94</td>
<td>0.82</td>
<td>0.89</td>
<td>0.94</td>
<td>1.00</td>
<td>0.91</td>
<td>0.91</td>
<td>0.92</td>
</tr>
<tr>
<td>0.94</td>
<td>0.94</td>
<td>0.96</td>
<td>0.91</td>
<td>1.00</td>
<td>0.94</td>
<td>0.97</td>
<td>0.97</td>
</tr>
<tr>
<td>0.96</td>
<td>0.86</td>
<td>0.93</td>
<td>0.95</td>
<td>0.92</td>
<td>0.97</td>
<td>1.00</td>
<td></td>
</tr>
</tbody>
</table>

Table 1:
Spearman’s rank test correlation matrix of Consensus Rank, Algorithm Rank, and ranks from six individual experts.

1981
Where’s the Data and Where Do I Start? Assessing Stakeholder Needs for Effective Use of Institutional Data

Chandrea Culbreath, MPH, BSN1; Maria Souden, PhD1; Melissa Nemon, PhD1,2; Holli Kreinbring12; Margaret Gonsoulin, PhD1; Denise M. Hynes, MPH, PhD, RN 1,3,4

1 US Department of Veterans Affairs (VA) Information Resource Center (VIReC), Edward Hines, Jr. VA Hospital, Hines, IL; 2 A-Team Solutions, LLC, Fulton, MD; 3 Center for Innovation in Complex Chronic Healthcare (CINCCH), Edward Hines, Jr. VA Hospital, Hines, IL; 4 College of Medicine and School of Public Health, University of Illinois, Chicago, IL

Introduction
The Veterans Health Administration (VHA) is the largest integrated healthcare system in the United States, providing care to almost nine million Veterans a year at over 1700 sites of care nationwide. Administrative and clinical data generated in the course of care are made available to VA clinical, operations, and research employees in order to inform healthcare decisions, improve the health system, and contribute to generalizable knowledge. The VA Information Resource Center (VIReC) is funded by VA Health Services Research & Development (HSR&D) to advance VA capacity to use data effectively for research and quality improvement (QI) projects. VIReC develops knowledge about data contents, structures, and methods for use and disseminates this knowledge through its communications and education programs. Recently the Center undertook a large-scale needs assessment to examine the challenges faced by some 5,000 users and potential users of VA data, identifying unmet needs for effective use of VA data and generating stakeholder recommendations for addressing those needs.

Methods & Sample
VIReC’s Needs Assessment is a multi-level study that includes three distinct phases of data collection in an iterative, mixed methods design. Participants for the first two phases were recruited from 4,675 VA staff using VIReC products and services or otherwise affiliated with the VA research and analytic enterprise, including health service and QI researchers, program office analysts, and clinical staff. In Phase 1 six focus groups were conducted with VA data users (n=44) in order to understand their data use concerns and challenges within the context of daily work activities. Phase 2 consisted of a survey of 829 data users, aimed at quantifying themes identified in Phase 1 and anticipating topical needs for knowledge resources and support. Phase 3 funneled data from the first two phases to six panels of leadership-level organizational stakeholders (n=50), including research centers, QI and evaluation programs, data offices, and advisory authorities. Stakeholder focus groups generated ideas for meeting data user needs and participants completed a follow-up survey (n=41) in order to develop consensus on the priority and feasibility of possible activities.

Results
Data collected from the first two phases of the assessment identified four major areas of concern or unmet needs for data users: 1) navigating the VA data landscape; 2) complicated data access processes and access delays; 3) data quality concerns; and 4) working effectively with VA data. In the final phase of the assessment, leadership-level panels considered these results and identified four high-priority and high-feasibility focal areas for VIReC: 1) Increase education about VIReC products and services, especially for new researchers; 2) Work towards affecting system change for improved data access and data access processes; 3) Facilitate community contributions around data quality, effective use of data, and code sharing; and 4) Facilitate acquisition of non-VA data.

Conclusions
Findings from the VIReC Needs Assessment provide insight into the challenges faced in supporting the use of institutional data resources for a broad range of users and use purposes. Effectively supporting the use of rich data resources within the context of a learning healthcare system requires supporting data knowledge and access throughout all phases of project activity and being attentive to the range of experience and perspectives brought to the data. Stakeholders envisioned solutions to user challenges that capitalized on the Center’s core strengths of data expertise, communication, collaboration, and facilitation. The needs assessment process presents a valuable framework for addressing user perspectives while accounting for the realities of leadership priorities and organizational constraints.
The Capture of Social and Behavioral Determinants of Health in Electronic Health Records
Theresa Cullen, MD, MS\(^2\), Katie Allen, BS\(^3\), Sara Armson MS\(^3\), Anna Roberts, MS, MS, Daniel J. Vreeman, PT, DPT, MS\(^2\)

- Regenstrief Institute, Indianapolis, IN; - Indiana University School of Medicine, Indianapolis, IN

Introduction

Increasingly, the clinical community is becoming cognizant that health status is dependent upon numerous external influences often referred to as social determinants of health (SDH). Precision health and personal health initiatives ask physicians to look at the whole person, including SDH, in addition to clinical factors. There have traditionally been limited ways to enter and retrieve SDH data within the EHR. In 2014, the Institute of Medicine released *Capturing Social and Behavioral Domains and Measures in Electronic Health Records: Phase 2*. The IOM identified core domains in Phase 1 of this work that were recommended for inclusion in patients health histories; Phase 2 endorsed specific measures for each of these core domains. In December 2015, the Regenstrief Institute, Inc, working with ONC, published a SDH panel in LOINC that contained SDH variables representing a minimum set of measures to be included in electronic health records. These LOINC terms were referenced as an optional component of the Meaningful Use incentive program’s 2015 Certification Criteria. This panel (80216-5), included social, psychological, and behavioral observations consistent with the IOM recommendations. We sought to understand how the availability of a standard, coded collection of SDH variables (endorsed by ONC within the MU program) would permeate into electronic record systems where these data were previously captured as free text elements (if at all). Specifically, we wanted to examine the impact of this new standard in supporting the capture of SDH elements in a large, regional, health information exchange (HIE) before and after the release of the Meaningful Use certification panel.

Methods

We sought to characterize the representation of SDH elements in the Indiana Network for Patient Care, one of the most comprehensive and longest tenured HIEs in the U.S. In this retrospective study, we leveraged existing clinical data from regional health systems. First, we created a cohort of patients aged 18 and older with at least two clinical visits (outpatient or emergency department) during the study period, to use as the denominator. Utilizing the clinical message flow within the HIE, the team examined Continuity of Care Documents (CCDs) for patients within the identified cohort. The study period was January 1, 2014 through December 31, 2016, corresponding with the IOM report and release of the LOINC panel. Presented data will include prevalence of the LOINC SDH panel codes in the periods pre/post the release of the Meaningful Use panel and a more detailed analysis for a subset of codes that were available prior to the IOM report. The subset includes content from three domains: exercise in days (68515-6); alcohol consumption (68518-0); and marital status (63503-7). Basic demographics and analysis will also be shared.

Evaluation Results

Our preliminary findings indicate minimal, if any, uptake of this LOINC panel as recorded in the HIE.

Conclusions

Identifying SDH elements from clinical notes requires providers to wade through a cacophony of unstructured data. Use of structured data elements, like the LOINC panel, should facilitate collection and longitudinal analysis of these important clinical factors. However, primary findings indicate minimal, if any, uptake of this LOINC panel as recorded in the INPC. Future work includes discussions with local CMIOs to identify how SDH data may be recorded and/or stored within the certified EMR and which data elements are selected for inclusion in CCDs. Future work may also include analysis in other health systems and a human factors usability evaluation.
Using Environmental Sensors to Estimate Users Activity

Dorothy W. Curtis, M.Sc., Mike Wang
Computer Science and Artificial Intelligence Laboratory, MIT, Cambridge, MA, USA

Overview
This project analyzes data from a prototype monitoring system designed for wheelchair users. This monitoring system has several sensors and a small computer mounted on the wheelchair. The computer collects the sensor data and transmits it to a central system for further analysis. The sensor data includes temperature and humidity data that can be used to generate alerts indicating excessive heat exposure. This project analyzes whether the temperature data can also be used to infer whether a user is indoors or outdoors, which could be helpful in determining the activity levels of wheelchair users. Our poster will show the results of our algorithm and evaluate its performance.

Methodology
Our program first extracts the raw temperature and humidity data from a Postgres database and writes the temperature and humidity values into a CSV file as a list of time-value pairs. Then it filters out any outliers and windows the data into 15 second windows. Based on estimated outdoors and indoors values for temperature, the program estimates whether a person is inside or outside at a particular time. However, in order to reduce the effect of noise when a person transitions from outside to inside or from outside to inside, the program requires that these transitions are at least 5 minutes apart. The program uses historical weather data to establish the baseline reference for the outdoor temperature. The initial inside temperature estimate is based on the initial readings from the temperature sensor because the wheelchair starts inside the facility every morning. Even though the facility is air conditioned, its ambient temperature increases during the day, and so the program updates its reference value for the indoor temperature to the median value of the most recent block of time when the person was inferred to be inside.

Evaluation

Figures 1 and 2 show some preliminary results from the algorithm. The y-axis is temperature in Fahrenheit and the x-axis is time of day in hours. The thin red line represents the reference indoor temperature. The green line represents the reference outdoor temperature. The strong red points in the graph are the raw temperature measurements. The thin blue line indicates whether the algorithm infers the person is outside or inside: when the blue line takes on the max value, the person is inferred to be outside and when the blue lines take the min value, the person is inferred to be inside. These plots illustrate some issues with the algorithm: for example, in the first graph, the algorithm indicates that the person went outside at 12.5 hours, though this is likely not the case and we expect that the algorithm became confused because the outdoor (green) and indoor (red) temperature reference lines crossed, i.e., it became cooler outside than inside, which is possible but unlikely in the summer in the daytime in Boston, MA, USA. In the second graph, the person is likely to have gone outside at around 16.5 hours even though the algorithms thinks the person stayed inside. This may be due to the algorithm requiring any transition to last at least five minutes.

Conclusion
We have developed an algorithm to infer whether a wheelchair user is inside or outside based on ambient temperature data. This experiment has shown that when we deploy this system we should also deploy stationary indoor and outdoor temperature sensors and improve our algorithm to handle more situations.
Mixed Methods Approach for Understanding Clinical Workflow
Ioana Danciu, MS\textsuperscript{a,b}, Stuart T. Weinberg, MD, FAAP\textsuperscript{b}, Daniel Fabbri, PhD\textsuperscript{a,b}, Kim M. Unertl, PhD\textsuperscript{b}

\textsuperscript{a}Vanderbilt University School of Engineering, 2301 Vanderbilt Pl., Nashville, TN 37235
\textsuperscript{b}Vanderbilt University School of Medicine, 1161 21st Ave S., Nashville, TN 37232

Introduction
Although the wealth of electronic patient record data is extremely promising for large-scale computational approaches, making causal inferences through purely computational methods can lead to potentially incomplete or misleading false conclusions. Qualitative methods can assist with understanding clinical processes through the perspectives of the participants, and generate a wealth of information about human behavior, work practices of healthcare teams, and technologies that are available to them. Qualitative methods however face challenges of scale across space, time and human participants. We propose a mixed methods approach (Figure 1) that links qualitative methods with data analytics to model clinical workflow and make causal inferences.

Methods
The Vanderbilt Ingram Cancer Center (VICC), one of 41 National Cancer Institute (NCI)-designated Comprehensive Cancer Centers, was the setting for our study. The data in our analyses were produced by a workflow and communication tool called the Outpatient Whiteboard which aggregates appointments, laboratory results, pharmacy medication dispenses, and other information, into a user-friendly web interface. In the last year on average 1900 Whiteboard transactions have been recorded on 225 patients every clinic day at VICC. We analyzed 25487 appointments for 6344 patients.

The metrics we selected were the total number of patients in a location at a given time aggregated over the two months study period and the average duration in minutes of each transition through the different areas. Over the course of two months during the summer of 2014, one researcher (ID) observed healthcare team members for 31 hours at the infusion center, the infusion center front desk, the VICC registration desk, the laboratory front desk, the laboratory, the pharmacy, and cancer clinic exam rooms. We used information collected during the observations about the workflows in the different areas, the performers and the technologies available to them to explain trends in the room movement quantitative data from the outpatient whiteboard.

Results
A trend in our observation data revolved around events being sequential and in a predefined order. When interacting with the VICC areas patients usually visit the registration area, the lab, the doctor’s office and the infusion area in this particular order. In addition, there are only two entry points into the system: the VICC front desk, and the infusion center front desk. The Pearson correlation between the number of patients and the time spent receiving care or waiting calculated over our 2 months study period (Table 1) shows some interesting trends. The time spent in the lab wait room is directly proportional to the number of patients. The time spent in the Lab correlates moderately with the number of patients because as we observed during the qualitative phase, the limiting factor is the staffing. The clinic wait room shows no correlation between our two metrics. The Infusion clinic time spent in the wait room time does not correlate with the number of patients in the area. Again, this could be due to patients arriving earlier for their scheduled appointments. The time spent in the infusion area is the longest for all areas, because as we observed during the qualitative phase, chemotherapy administration takes three hours on average. The infusion nurses we observed were extremely busy during the infusion start and the infusion end. Staggering patients when initially put in infusion rooms could lead to less busy discharges.

Discussion and Conclusion
In this work, we presented a mixed methods approach to understand and represent clinical workflow. Using data analytics, we were able to quantify workflow patterns in clinics. Qualitative observation assisted us with understanding trends in the data. Next steps include targeting the wait areas and understanding the factors that impact the time dependence of the delays. Also, incorporating other data sources such as the Laboratory Information System or the Outpatient Scheduler can provide vital insight into the trends observed in the Whiteboard data.
Structured Data Capture (SDC)

Farrah Darbouze¹, William E. Hammond, PhD², Richard Moldwin, MD, PhD³, Sue Dubman, PhD, MA⁴, Matthew Chudy⁵

¹Office of the National Coordinator for Health Information Technology, Washington, DC; ²Duke University, Durham, NC; ³College of American Pathologists, Northfield, IL; ⁴University of California, San Francisco, CA; ⁵Data Consulting Group, Detroit, MI

Problem

The use of electronic health record (EHR) data for secondary purposes is limited due to a lack of uniformity in data structure, transport, and definitions across EHR systems. Streamlining EHR data in a structured way can accelerate determination of coverage, quality and safety improvements, improve population health, and foster scientific research.

Project Objective

The Structured Data Capture (SDC) Initiative provides a standards-based infrastructure for the capture and use of patient-level data in EHR systems for adverse event reporting, public health reporting, and scientific research. The initiative adopted and enhanced four national standards specific to common data elements (CDEs) used in filling, structuring, or designing electronic forms/templates; standardizing EHR interaction functions; auto-populating forms/templates; and transporting structured data between systems.

Project Description

The SDC Workflow steps include: (1) a Form Filler requests a form/template from a Form Manager; (2) a Form Manager sends the requested form/template to the Form Filler; (3) the Form Filler converts, populates, and displays the form; (4) the Form Filler stores/transmits the structured data to an external data repository; (5) upon request, the external data repository extracts, transforms, and sends the completed structured data to the end user. Three pilot teams successfully completed end-to-end testing of multiple forms using the SDC architecture. The three forms are as follows: FormDesign, RequestForm and SubmitForm.

Conclusion

SDC forms are an efficient mechanism to achieve interoperability with standardized structured clinical content. This project was funded by the Patient-Centered Outcomes Research (PCOR) Trust Fund, managed by the Office of the Assistant Secretary for Planning and Evaluation (ASPE) in the U.S. Department of Health and Human Services.

References

Evaluating Health Kiosk Technology as an Enabler of Shared Decision Making

Selena Davis a, Karen L. Courtney a, and Judith T. Matthews b
a University of Victoria, Victoria, BC Canada; b University of Pittsburgh, Pittsburgh, PA US

Introduction

Today there is increased interest in health information technologies (HIT) that engage patients in decision making as part of self-management. Shared decision making (SDM) is a collaborative process that allows patients and their providers to make decisions together by considering the best available evidence as well as patient values and preferences, to identify the best strategy at a specific point in time. SDM involves choice, option, decision, and action to shift the patient’s initial preferences to informed preferences and culminates in a shared decision expressed in a care plan with explicit follow-up that respects and tracks outcomes related to those preferences. Effective SDM requires patients to have access to their health information, health management and decision support tools, and the means to communicate with their providers. Multi-user health kiosks provide an ideal platform for optimizing SDM.

Health Kiosk Application

Initially designed to assist older, community-dwelling adults in monitoring selected health parameters and communicating with their care team, our multi-user health kiosk has expanded to include screening and interactive intervention aimed at improving self-management of sleep, mobility, lifestyle, bladder control, and emotional health. Data are collected via several integrated devices as well as self-report. Recommendations are tailored to the patient.

Methods

By adapting a conceptual framework, which maps characteristics and elements of the SDM process to enabling personal health record system functions by patient activity, this descriptive qualitative study evaluates the enablement of SDM in context of our health kiosk. Using the adapted framework (SDM-HIT) and a 3-point scale, the researcher performed an onsite system walkthrough to simulate the user experience followed by individual semi-structured interviews with members of the development team to evaluate the availability of system functions for the user. Directed content analysis was used to reveal optimization of our Kiosk for SDM – e.g. SDM element “Receive decision support – Elicit personal preference in context of decision” was assessed as: (i) currently implemented; (ii) available but not implemented to enable SDM; and (iii) not currently implemented but recommended for future development.

Results

An assessment summary of our health kiosk functions that could be leveraged in support of SDM identified many areas for optimization, including diverse functionalities which enable patients to access health information, record health information, receive decision support and communicate with others – e.g. our Kiosk clinical profile comprises data from mostly self-report and a few integrated devices, a simple preference profile is created, unsophisticated decision support is used, and communications with others are limited including care providers.

Discussion

To engage patients in health self-management and decision making, systems must be designed for that purpose. The opportunities and identified gaps from this study may be translated into implementation and operational requirements for our health kiosk and other community-based, multi-user, self-management technologies. Using the SDM-HIT framework as an evaluation tool offers a novel approach to assessing HIT systems for the enablement of SDM.

Acknowledgement

The project was supported by grant #R01HS022889 from the Agency for Healthcare Research and Quality (AHRQ). The content is solely the responsibility of the authors and does not necessarily represent official views of the AHRQ.

References


**Corresponding Author: daviss@uvic.ca**
Testing an Explainer Animation for Public Health Education

Jeffrey Day, MD, MA1, Anne Altemus, MA1
1Lister Hill National Center for Biomedical Communications, NLM, Bethesda, Maryland, USA

Abstract

We are experimenting with online explainer animations to improve the outreach mission of the National Library of Medicine. A short animation on histamine will accompany a food allergy article in NIH MedlinePlus magazine and we will see if it can help enrich the magazine articles, attract new audiences, and expand awareness of the NIH. If successful, this project can inspire more multimedia communication strategies to convey public health messages.

Overview

The National Library of Medicine (NLM) serves as a national health information source for the public, patients, and families.1 As part of this mission, NLM produces NIH MedlinePlus magazine to deliver popular, consumer-friendly content. To improve the magazine’s online presence, we will test the use of explainer animations to: 1. enrich the magazine’s written stories, 2. attract new audiences to magazine content, and 3. expand awareness of NIH.

Online video is a promising strategy to serve each of the above goals. Eighty-six percent of Americans now use the internet,2 over 70% of those users share content via social media,3 and social actions for video are growing faster than for any other shared content.4 Explainer animations in particular have been shown to quickly convey complex ideas to target groups in a short amount of time,5 making them ideal for delivering health information.

We will create a four-minute explainer animation on histamine to be released in conjunction with a food allergy article in the Summer 2017 issue of NIH MedlinePlus the Magazine. The animation will follow Section 508 standards of accessibility including audio described narration, closed captioning, and WCAG 2.0 AA contrast standards.

To gauge enrichment and user interest, a crossover study will compare the food allergy article and animation. Subjects will view the article and animation in two separate sessions, with half the subjects reading the article first, and the other half watching the animation first. Attitudes about the NIH and the magazine’s content will be surveyed before the sessions, after the first session, and after the second session, concluding with a debriefing interview. The crossover study will allow for A/B comparison, as well as user insights after subjects see both the article and video.

Measurements of a video’s ability to draw new audiences will be based on analytics from the magazine website, social actions, and YouTube views. Social actions will also be compared between a Spanish and English version of the explainer animation, which can provide insights into how the two different language demographics interact with the NLM’s online content.

This will be the first of several explainer animations created in conjunction with the magazine website, and can inform multimedia communication strategies to share public health messages at the NLM.

References

CTB: A Custom Taxonomy Builder for Named Entity Extraction
Dina Demner-Fushman, M.D., Ph.D. and Willie Rogers
US National Library of Medicine, Bethesda, MD

Abstract
Clinical researchers are often interested in finding specific disease phenotypes in clinical text. When using dictionary-based named entity recognition tools, they might need to add an important local terminology or limit the scope of named entities provided in the publicly available resources. In this poster, we present an online, as well as downloadable, open-source tool that allows constructing custom-built subsets of the UMLS Metathesaurus.

Introduction
Processing of biomedical or clinical text for a specific task often does not require the full contents of the Unified Medical Language System® (UMLS®) Metathesaurus®. Several tools for customizing the Metathesaurus are available: MetamorphoSys - the UMLS installation and customization program – allows excluding vocabularies, suppressible terms, languages, and semantic types, among other filters. The MetaMap Data File builder allows creating UMLS-like data models from a provided source. These tools, however, require download, installation, and potentially some knowledge of command line procedures. For users interested in creating their vocabulary in an online Graphical User Interface (GUI), we developed an open-source Custom Taxonomy Builder (CTB) that allows submitting a set of seed terms for a given task or phenotype, edit the automatically-generated taxonomy, and download the custom-built UMLS-like files for use with MetaMap Lite or MetaMap.

Methods
CTB is implemented in Clojure and Java and uses the UMLS MRCONSO.RRF files restricted to level 0 for public use, as well as the lexical tools (LVG) to generate variants and synonyms for the user-submitted seed terms. For each seed term, the system extracts the sub-terms using a MetaMap-like tokenizer and searches its local UMLS index files for synonyms of each sub-term. Candidate terms are then constructed from recombination of the synonyms using one synonym from each sub-term and ordering sub-terms in the original seed-term order. Any UMLS strings matching the candidate terms are added to the candidate synonym set with associated synonyms grouped by concept. If no UMLS strings are found, the LVG’s fruitful variants are applied to each of the sub-terms and the intersection of the UMLS and LVG synonyms are then added to the candidate synonym set using a synthetic concept id.

Results
The online CTB version can be used to generate vocabularies in real time. The downloadable version can be installed on any operating system that supports Java. The GUI allows selecting or de-selecting whole or partial synonymy sets for a given term. The resulting files can be downloaded and used with MetaMap Lite or processed using the MetaMap Data File builder.

Conclusions
We present a publicly available tool that allows to easily build a UMLS-based custom vocabulary for downstream use with MetaMap Lite, MetaMap or similar tools. In the future, we plan to add an option for including WordNet synonymy in the custom-built vocabulary. CTB can be accessed at https://ii-public1.nlm.nih.gov/ctb

Acknowledgements
This work was supported by the intramural research program at the U.S. National Library of Medicine, National Institutes of Health. We thank Clem McDonald and Francois Lang for inspiring this work.

References
5. https://wordnet.princeton.edu/
Improving the Quality of Clinical Data Extracted from Text

Daniel Denhalter, BS\textsuperscript{1,2}, Olga V Patterson, PhD\textsuperscript{1,2}, Brian Robison, MPH\textsuperscript{1}, Scott L DuVall, PhD\textsuperscript{1,2}

\textsuperscript{1}VA Salt Lake City Health Care System; \textsuperscript{2}University of Utah, Salt Lake City, UT

Abstract

Data quality has an impact on all potential uses of the data. Natural language processing (NLP) can be used to complement structured data and increase both completeness and correctness by adding data extracted from clinical notes and reports. However, data quality errors can be amplified when data entry errors are combined with data extraction errors. This poster reports on data quality improvement steps implemented to detect and filter out incorrect entries resulting from NLP extraction of variables from large note data sets.

Introduction

The Department of Veterans Affairs (VA) Informatics and Computing Infrastructure (VINCI) aims to improve access to VA data and to facilitate the analysis of that data. One of the most requested variables for research in heart disease is left ventricular ejection fraction (LVEF). LVEF is a measure of heart function and is expressed as the percentage of total blood volume pumped out of the heart with each beat. Only a small fraction of LVEF values are available as structured data in the VA electronic medical record system; the majority are recorded within notes and reports. To improve availability of LVEF data, VINCI has developed an NLP system to extract LVEF values from these clinical documents. The system was validated with average precision of over 98% across different document types.[1] While it is a respectable performance for an NLP tool, additional data cleaning was required.

Methods

LVEF values extracted from a comprehensive set of VA clinical documents were combined with available structured data to create a combined set of 49 million records for 3.4 million patients. The NLP output included LVEF values (which could include single measures or two measures defining a range) and the corresponding phrase from the source document. The initial step in the analysis was plotting the values as a normalized histogram binned into intervals based on the American College of Cardiology recommendations for heart failure diagnosis.[2] Values outside the range were flagged as potential data entry or data extraction errors.

A disproportionally large number of records (209 thousand) were found to be between 0 and 1, which is not a viable value for LVEF. Manual review of the source phrases for 1000 random instances indicated that in most cases, the LVEF value was represented as a decimal number rather than percent. In a small number of instances, the source phrases revealed extraction errors. These errors could be identified by mismatching units of measure mentioned in the source phrase, such as “cm” or “ms”, as well as the use of double zeroes, which often appear in eye exam results. For example, the term “EF” in a phase “EF -0.50-1.00” stands for “Eye Field”. Similar issues were found for values that were above 1, therefore, all values with mismatching unit of measure keywords or double zeroes in the source phrase were marked as erroneous. The rest of the values between 0 and 1 were accepted as true LVEF values and were multiplied by 100 to be consistent with percent values.

To inform the planned improvements of the NLP system, a review of 800 randomly selected instances greater than 100 was conducted. The most frequently occurring error was mistaking date for the upper boundary of a range. For example, “EF 55%-2000” should be interpreted as the value measured in 2000.

After all marked values were removed from the dataset and duplicate instances of the same values on the same day for the same patients were filtered out, the resulting dataset is made available to any VA affiliated researchers with appropriate access authorizations. While the erroneous values comprised less than 0.2% of all instances, they accounted for a large number of records (~78 thousand). The final set contains 35 million LVEF values for 3.4 million patients with the date when the value was recorded.

Conclusion

Ensuring data quality is a vital step in any study. By providing NLP extracted values and performing data cleaning on a common data set, VINCI has made a major step from facilitating access to data to providing previously inaccessible high quality variables to research groups.

References

Blockchain for Privacy and Security: The Case of Health Informatics

Gurpreet Dhillon, PhD1, Kane J. Smith, MSIS1, Monica C. Tremblay, PhD2

1Virginia Commonwealth University, Richmond, VA; 2Florida International University, Miami, FL

Abstract

Blockchain implementation in Health Informatics is a significant challenge in a rapidly changing era of privacy and security concerns. When dealing with such concerns, healthcare institutions are presented with a serious problem in how to manage it and where to allocate resources to maximize value. It is important to understand how organizations address these concerns by exploring blockchain implementation so it can be managed in the context of cybersecurity. To do this the problem question is twofold: First, what objectives are important based on the strategic values of an organization with regard to the implementation of blockchain technology. Second, how can these objectives be used to evaluate proposed solutions for blockchain implementation in electronic medical record systems in healthcare organizations. In this research we utilize Keeney’s value focused thinking to demonstrate how the process can occur to maximize value-add within healthcare organizations implementing blockchain technology to address security and privacy concerns with respect to patient data.

Introduction and Background

To handle patient concerns healthcare institutions are presented with a serious problem; how to manage security and privacy and how to maximize value in the application of technological solutions. Hence, it is important to understand how organizations address these concerns by exploring blockchain implementation so it can be managed in the context of cybersecurity. Numerous academic contributions exist regarding the application of blockchain in health informatics for solving privacy and security concerns in healthcare organizations. However, there is currently no research with respect to the actual decision making process such that an organization might best select an implementation strategy for blockchain to maximize its benefits. Hence, our research proposes such a model using Keeney’s Value Focused Thinking (VFT) to create a decision framework model that organizations can adapt for these own use-case and we demonstrate its application on the decision context.

Methodology

To ensure greater measures of protection from digital crime using blockchain technology, it is necessary to identify the fundamental objectives for implementing the technology in Health Informatics. This is where Keeney’s VFT for decision analysis can help create an objective-based framework for healthcare and related organizations to model the necessary objectives for preventing digital crime using blockchain technology. This model can then be used to assess alternatives for implementation and demonstrate gaps in all proposed solution’s performance relative to the objective’s goals which acts as quantitative metrics for comparison. This process occurs in five basic steps adapted from Keeney and Merrick et al. and derives its objectives either directly from literature and expert users of blockchain technology and health informatics, which are interviewed to elicit the objectives as well as the necessary importance and swing weights to fully build out the proposed model.

Conclusion

The investigation mixes qualitative and quantitative methods, value-focused thinking and multi-attribute utility modeling to reveal that an objective-based framework can be utilized to create the necessary model by which an organization can evaluate how to secure patient records using blockchain technology. This will provide a significant contribution as previous research in this area is under-developed and as such falls short of evaluating alternative solutions in the context of maximizing application value within an organization. Preliminary results clearly indicate that a feasible model can be built using decision analytic techniques like VFT to create useful models for healthcare organizations faced with the difficult task of integrating solutions to manage privacy and security concerns.

References


1991
The Development of the Research Precision Oncology Program Data Repository (PODR) in the Veterans Affairs Healthcare System

MAVERIC, VA Boston Healthcare System, Boston, MA

Abstract

The objective of the VA’s Precision Oncology Program Data Repository is to collect genomic, proteomic, and phenotypic data from multiple sources such as the EHR, Corporate Data Warehouse, patient reported outcomes, and medical imaging archives from consented participants. PODR is an essential part of the VA’s effort towards a Rapid Learning Health-care System. PODR’s data will be accessible to researchers external to the VA through the National Cancer Institute’s (NCI) Genomic Data Commons.

Introduction

Through the Precision Oncology Program, the VA started offering genomic analysis to patients with advanced stage solid tumors in 2015 as part of routine clinical care. Recognizing that the promise of genomic-driven cancer medicine remains limited given the many challenges such as unexplained drug resistance and diversity of tumor mutations; the Massachusetts Veterans Epidemiology Research and Information Center (MAVERIC) started the Research for the Precision Oncology Program (RePOP) in 2016 to acquire the knowledge necessary to improve the care for Veterans with cancer. Since genes with driver mutations tend to be rare and more likely to be discovered through large numbers of patients, MAVERIC established PODR with an additional goal of data sharing with the larger research community outside of the VA.

Overview of System

Eligible patients for RePOP are identified once tissue biopsy and genomic analysis have been performed successfully. Patients are screened further for eligibility with their physicians and then contacted by mail with information and an opt-out option. A research coordinator then contacts the patient by phone to provide additional scripted information about the research program and the different options of informed consent. Once consented for enrollment in RePOP, the participant’s longitudinal data are collected from various data sources and stored in PODR. A combination of natural language processing (NLP) and chart review are used to extract and validate unstructured data. For data sharing, data is assembled, deidentified, and mapped to the NCI’s Genomic Data Commons (GDC) data model and Common Data Elements. If a VA clinical data element is not currently defined in the GDC’s data dictionary, then it will be made available to external researchers as part of the GDC’s clinical supplemental file submission.

Results

There are currently 34 Veterans enrolled in RePOP with a 71% successful recruitment rate. 100% of recruited Veterans consented to sharing their data for research within the VA, 91% consented to sharing de-identified data outside of the VA, 97% consented to future contact, and 88% consented to all three options. There are 28 patients with lung cancer and 6 with other cancers. Of those enrolled, 94% had mutations annotated as “Genetic Alterations of Medical Importance”. Initial analysis of the GDC data dictionary against the clinical data of four enrolled patients estimates that 38% (14/37) of GDC clinical data elements are available in the VA as structured data (e.g. diagnosis), and 46% (17/37) of GDC data can be found in the VA as unstructured clinical text (e.g. tumor staging) which necessitate the development of NLP tools to improve the efficiency of the curation process.

Conclusion

The development of the VA’s PODR is an essential part of the VA’s long-term strategies toward a Rapid Learning Health-care System. PODR is being built to collect all longitudinal data relevant to the patient’s cancer, and various approaches are being explored to accommodate the growing clinical, genomic, proteomic, medical images, and patient generated data streams. A partnership has been established between NCI and VA to use the NCI’s Genomic Data Commons to break through the barriers of data sharing and accelerate the pace of new discoveries.
Reasons for Rework in Clinical Data Provisioning: A Root-Cause Analysis

Amanda L. Do, MPH\textsuperscript{1,2}, Ruby Y. Wan, MS\textsuperscript{1,2}, Robert W. Follett\textsuperscript{1,2}, Marianne Zachariah\textsuperscript{1,2}, and Douglas S. Bell MD, PhD\textsuperscript{1,2}

\textsuperscript{1}Department of Medicine, UCLA David Geffen School of Medicine, Los Angeles, CA, and \textsuperscript{2}UCLA Clinical & Translational Science Institute, Los Angeles, CA

Introduction

Secondary use of electronic health record (EHR) data in clinical research has become more prevalent with the deployment of large-scale EHR systems. Challenges frequently arise in repurposing clinical data for research, and many investigators are unable to relate their needs to the existing clinical data. The UCLA Clinical & Translational Science Institute’s (CTSI) Informatics Program (IP) has implemented, and continues to refine, a data provisioning service model, which includes consulting and programming services, to assist investigators in defining EHR data requests that address their research questions. However, investigators sometimes find that a dataset does not meet their needs only after it has been provided, which costs the IP additional time and resources to remedy. The IP seeks to understand the reasons for this kind of rework within the context of the honest broker program at the UCLA CTSI.

Methods

IP data consultants analyzed requests for data from UCLA’s Integrated Clinical and Research Data Repository (xDR), a clinical data warehouse of UCLA’s CareConnect (Epic) EHR data, linked with data from older legacy systems and other sources. Data requests were specified via a standardized template, with two components: 1) definition of patient selection criteria and 2) selection of demographics and output variables e.g. diagnoses, lab results, etc. Data requests were considered complete once all template variables were queried, exported, and delivered to investigators. Rework was defined as any instance in which the original query code was revisited and modified. To identify and evaluate rework, we reviewed all records from our project tracking system, which included consulting and programming notes, email correspondence, or additional files sent by the investigator. Based on our consulting experience, we generated 6 classifications for rework: data variables needed, selection criteria adjustment, data processing errors, newly available data, formatting (data analysis optimization), or miscommunication. We assigned any one or combination of the 6 reasons to each rework request.

Results

From March 2014 to February 2017, 33 requests (16.4%) for rework have occurred out of a total 201 completed data requests. 19 out of the 33 requests (57.6%) had only 1 cause for rework. The remaining 14 requests (42.4%) had more than 1 rework reason. Among the 14 requests were: 11 requests with two reasons, 2 with three reasons, and 1 with four reasons. Adjustment of patient selection criteria was the most common reason given by investigators for rework, accounting for 21 out of 51 reasons. Requests for additional data variables comprised 14 reasons.

Conclusion

Overall, most investigators were able to use the data provisioned through the UCLA CTSI IP service model without needing to request revisions. Most instances of rework were related to inadequate specification of the original data request and few reworks were due to external factors such as data processing errors. The root cause of most rework may be insufficient understanding of the nature of EHRs by the investigator. EHR data is not collected for research purposes and may be incomplete or inaccurate. In order to conserve IP resources, further efforts are needed to address these limitations and to aid in the planning of well-specified data requests that would not require subsequent adjustments in selection criteria or output variables.

\begin{table}[h]
\centering
\begin{tabular}{|l|l|l|}
\hline
Rework Reason & Example & Frequency* (%) \\
\hline
Adjustment to patient selection criteria & Implement less restrictive criteria to identify more patients for recruitment & 21 (41.18) \\
\hline
Additional data variables & Request for additional laboratory results or medication data & 14 (24.45) \\
\hline
Data processing errors & Omissions due to technical issues in extract, transfer, load (ETL) process & 6 (11.76) \\
\hline
Formatting & Request for procedure ID to be added to every file for easier merging across files & 6 (11.76) \\
\hline
Newly Available Data & Transplant Registry data became available in xDR & 3 (5.88) \\
\hline
Miscommunication & Variables were discussed but were never captured on template & 1 (1.96) \\
\hline
\end{tabular}
\caption{Rework Reason Examples and Frequencies}
\end{table}

*Frequency: counts are not distinct. Rework requests could be assigned more than 1 reason.
Physician Conception of Patient Frailty in Cardiac Care Decisions

Kristina Doing-Harris, PhD1; Rashmee U Shah, MD2; Yijun Shao, PhD4; Bruce E Bray, MD2,3; Qing Zeng-Treitler, PhD4; Jennifer H. Garvin, PhD, MBA3; Charlene Weir, PhD3,5

1 University Buffalo, Department of Biomedical Informatics; 2Division of Cardiovascular Medicine, 3Department of Biomedical Informatics University of Utah; 4Medical Informatics Center at George Washington University; 5VA Healthcare System, Salt Lake City, UT

Introduction: Patient frailty is an important indicator of patient outcomes.[1] Frailty commonly represents a complex clinical syndrome of increased vulnerability to stressors, which results from multiple impairments across different systems.[2] However, there are large variations in relevant impairments and idiosyncratic applications between physicians. Despite these differences, frailty judgments are especially important in making decisions about cardiac interventions such as coronary revascularization or aortic valve replacement by surgical vs transcatheter methods.[1] We have found in prior work that clinicians document functional status and frailty extensively in regular progress notes and clinical documents. Our project’s overall purpose is to transform the descriptive narrative found in clinical notes into a structured, scaled, inclusive variable. Generating this variable requires an operational definition of frailty related to cardiac care decisions. Once we have narrowed the definition, we will score the relevant indications, find them in patient records, and calculate patient frailty.

In this study, we use physician interviews to generate a conception of frailty as it relates to cardiac care decisions. We investigate physician perception and application of frailty measures in their decision-making.

Method: The physicians on the research team developed vignettes describing 7 complex cardiac cases. Each case included a mix of indications of patient frailty, e.g., difficulties performing their Activities of Daily Living (ADLs), older age, and a variety of chronic co-morbid conditions (cancer, diabetes, renal failure). Treatment options were coronary artery bypass surgery (CABG), coronary stenting (PCI), or medical management. We interviewed 5 physicians (4 cardiologists, 1 CT surgeon). In order to elicit the interviewee’s conception and use of frailty indicators in decision-making, we generated open-ended questions and encouraged as much elaboration as possible.

We asked: How frail is the patient? What indicates patient frailty? What is a successful outcome? What role frailty plays in that outcome? What decisions need to be made? What is your treatment decision? Is frailty a factor? Two investigators extracted indications about physician conceptualizations of frailty and its impact on treatment decision.

Results: Common frailty indicators included age, social support, lower body strength, hospitalization-related deconditioning, and mental status. A key decision point was whether, or not, a frailty indication could be relieved by revascularization or medical management; taking into account time since symptom onset, patient’s baseline functionality, and symptom’s relationship to cardiac impairment. A key decision point, unrelated to frailty, was tenability of CABG/PCI indicated by coronary anatomy and blockage location and chronic co-morbid conditions (cancer, diabetes, renal failure). Participants varied in their final decision (there was no right or wrong answer).

Table 1. Key quotes from 5 physician interviews.

<table>
<thead>
<tr>
<th>Quote</th>
</tr>
</thead>
<tbody>
<tr>
<td>“He could get through surgery. So I would offer it… would only mean 3-5 extra days in nursing home.”</td>
</tr>
<tr>
<td>“I would want… her co-morbidities, possible malignancy causing night sweats, before I made a decision”</td>
</tr>
<tr>
<td>“…would live longer with by-pass (traditionally)… multiple reasons [to] consider staged PCI”</td>
</tr>
<tr>
<td>“CABG produces the best result. So surgery would decide. Him recovering well is for them to figure out.”</td>
</tr>
</tbody>
</table>

Discussion: We found some agreement about relevant impairments overall. Our results indicate that the frailty indicators used in decision-making fell into 2 groups symptoms that can be fixed by surgery or cardiac procedures and indicators of patient’s ability to thrive post-surgery. Most seemed to take the opinion that surgery provided the best symptom resolution. Thus, they applied frailty assessment as it indicated surgical survival.

References

Heuristics for Evaluation of Dashboard Visualizations

Dawn W. Dowding, PhD, RN, FAAN\textsuperscript{1,2}, Jacqueline A. Merrill, PhD, MPH, RN, FAAN, FACMI\textsuperscript{1}

\textsuperscript{1}Columbia University School of Nursing, New York, NY; \textsuperscript{2}Visiting Nurse Service of New York, New York

Introduction

Heuristic evaluation is a common method used in human computer interaction (HCI) to assess the usability of a system.\textsuperscript{1} The heuristics commonly used in HCI design are based on usability principles originally identified by Nielsen\textsuperscript{2}, and expanded by Pierotti.\textsuperscript{3} Whilst these cover general design principles their utility for evaluating systems that provide visualized data is not well established. In order to conduct a heuristic evaluation of a point of care dashboard for home care nurses, which uses data visualization to summarize information alongside more traditional design principles, we developed and refined a tailored set of heuristics for visualized data.

Methods

Nominal group technique using online survey methods. We constructed a list of heuristic principles with associated usability factors combining general principles\textsuperscript{3} with principles and usability factors pertinent to information visualization systems.\textsuperscript{1} Specifically, the heuristics and associated criteria identified by Forsell and Johnson\textsuperscript{1} were extracted from the original papers. An online survey was used to distribute the list of principles and factors to 12 experts selected because they have published in the field of nursing or biomedical informatics and information visualization. They were asked to rate each factor on its importance as an evaluation heuristic for visualization systems on a scale from 1 (definitely don’t include) to 10 (definitely include). After reviewing the distribution of scores the researchers established a median score \( \geq 8 \) to represent consensus for an item to be included in the final checklist.

Results

10 individuals responded to the online survey. The initial checklist had 7 general usability principles with 36 usability factors (visibility of system status, match between system and the real world, user control and freedom, consistency and standards, recognition rather than recall, flexibility and efficiency of use, aesthetic and minimalist design) and 7 information visualization specific principles with 21 usability factors (spatial organization, information coding, orientation and help, data set reduction, flexibility, consistency, remove the extraneous (ink)). After analysis of consensus responses, a total of 49 usability factors had a median score of \( \geq 8 \). These were organized into the final version of the checklist consisting of 10 principles and 49 usability factors. Individual factors from the visualization specific principles concerned with flexibility (n=4) and removing the extraneous (n=2) were combined with general usability principles (40 factors). In addition 3 visualization specific principles, with 9 factors remained (spatial organization, information coding and orientation).

Conclusion

To heuristically evaluate the usability of information systems that use visualization strategies to summarize data, it is necessary to incorporate general heuristic principles and information visualization design principles. We have developed a checklist that can be used by experts to conduct heuristic evaluation of visualization systems. The checklist subsequently has been used to evaluate a dashboard for home care nurses to use at the point of care.

References

Curbside Consultation in the Virtual Healthcare Environment: Mixed Methods Analysis Occupational Medicine Teleconsultation in the Veterans Health Administration

Jennifer Lipkowitz Eaton, MD MPH FACOEM, Amir Mohammad, MD MPH FACOEM

1Occupational Health Services, Office of Patient Care Services, Veterans Health Administration, Washington DC; 2Department of Internal Medicine, VA Connecticut Healthcare System/Yale Occupational & Environmental Medicine, New Haven CT

Background:
In an increasingly virtual healthcare environment, the role of informal provider-to-provider consultation remains unclear. Electronic curbside consultation is an evolving mechanism to answer clinical questions and build provider relationships.

Objective/Purpose:
This quality improvement study describes and compares the characteristics of complex cases referred for two types of electronic curbside consultation in Occupational and Environmental Medicine (OEM). We explore variation in cases evaluated through (1) direct, informal email consults versus (2) a formal specialty tele-consultation program.

Methods:
This analysis uses mixed methods techniques to describe and analyze cases submitted for two types of virtual curbside consultation. The VHA OEMedicine tele-consult initiative is a National provider-to-provider specialty practice group. Web-based intake and tracking processes generate data for clinical quality improvement efforts. Fifteen months of consult requests submitted to a formal, curbside tele-consult program were screened, coded and analyzed (n = 114). The comparison sample was comprised direct email or telephone consults to individual specialty physicians who participate in an online OEM community of practice (N=43). These informal email consults were forwarded directly to the QI team. Two OEM specialty physicians coded significant co-morbid conditions based on free text comments. Coder inter-rater reliability was 83.6%.

Results:
Professional comparisons revealed that all independent provider types were more likely to consult the formal tele-medicine group than choose an informal electronic curbside. Physicians were most likely to contact physician colleagues for informal curbsides (35.2% of cases) compared to Nurse Practitioners (26.2%) and Physician Assistants (7.4%). In chi-square tests, the distribution of consults did not vary significantly by specific medical condition (e.g. Endocrine, Cardiovascular, Infectious). Multi-morbidity did appear to be associated with greater frequency of formal tele-consultation (p<0.005). Average consult duration was significantly longer for direct consults to individual providers (31.51 vs 43.75 minutes, p<0.05). This did not vary significantly by profession.

Conclusion:
Findings suggest that electronic curbside consultation preferences and behaviors likely vary by case complexity. Compared to informal contacts, formal tele-consultation may create more opportunities for case collaboration between physicians and non-physicians. The results of this study can be used to integrate opportunities for curbside consultation into clinical process and informatics supports.

References
3. Worth ER, Patrick TB. Do electronic mail discussion lists act as virtual colleagues? AMIA 1996;8280:325-339

1996
Implementation of an Informatics Solution to Improve Management of Pathology Specimens for the VA Precision Oncology Program

Danne C. Elbers MS, Robert B. Hall MS, Beth J. Katcher BS, Svitlana diPietro BS, Sung Feng-Chi MS, Sergey D. Goryachev MS, Karen E. Pierce-Murray RN BA, Corri L. DeDomenico BS, Lauren MacMullen BS, Helene Garcon MD, Nhan V Do MD, Louis D. Fiore MD

MAVERIC, VA Boston Healthcare System, Boston, MA

Introduction
The Veteran Affairs (VA) Precision Oncology Program (POP) is a clinical care effort intended to bring personalized health care and cutting edge cancer treatment to Veterans by sequencing tumor samples. The program includes a research component (RePOP), which consents patients to share their clinical, genetic and imaging data. Program staff initially relied on MS Excel and InfoPath to handle the activities to identify patients diagnosed with cancer, track pathology samples with vendors, and enroll Veterans into the RePOP cohort. As the program expanded across the VA a more scalable informatics approach was required to better manage current and future functions of the POP program. To meet these requirements we developed a workflow software application in-house called ProjectFlow with the VA Informatics and Computing Infrastructure (VINCI) group to manage this and other research activities. For the implementation we created an integrated system consisting of a database model, web services, stored procedures, ETL’s and workflows adhering to the BPMN20 standard.

Business Requirements and Technical Development
A key requirement was to provide a solution for POP/RePOP activities that could manage the high volume of specimens and patients. Captured data needed to be available for reporting and analysis purposes. The program needed the ability to grow and evolve with new workflows, auditing and reporting needs that would develop as operations expanded. BPMN20 workflows were designed for the ProjectFlow application to capture the attributes associated with the clinical and research operations. To assist these workflows in their execution a relational database model was developed as well as stored procedures and a web service to validate ProjectFlow’s input before storing data through stored procedures in the database schema. We created a scheduled ETL that queries pathology electronic health records for VA Region 4 based on SNOMED and ICD codes and incorporated them into the POP database as a starting point for specimen identification.

System Deployment
The POP application was deployed in August of 2016. To date 1085 pathology specimens have been managed through the application among 55 sites. The approach incorporated structured data items and a SQL relational database schema to capture business processes which improved efficiency by reducing input errors and improving data management efforts. A wider array of reports and graphs has been generated from the POP data repository utilizing reporting services rather than relying on InfoPath features. These reports have been informative in directing operational improvements such as enhancements to specimen selection. Future plans involve the deployment of a RePOP workflow version (in development) and expansion of the ETL to query specimens beyond VA Region 4.

Conclusion
The development of the ProjectFlow application demonstrates how a structured Informatics solution can facilitate and improve business operations in the precision medicine arena. The application allowed our developers to streamline business operations for the POP program with less data errors and improved reporting and querying capabilities. Our workflow application provides a scalable approach that can handle a large number of pathology records and consents, and is extendible for modifications and additions. The team overcame challenges with incorporating VA pathology data and CDW patient records into the application. Our team continues to integrate with the POP operations staff to improve and develop novel ways to implement new workflows, integrate more VA healthcare data, and enhance system usability.
Authors: Kyle Erickson, BS; Jessica M. Malenfant, MPH; Kimberly Barrett, MPH; Adam Paczuski, BS; Chayim Herzig-Marx, PhD; Jeffrey S. Brown, PhD

Harvard Pilgrim Health Care Institute, Boston, MA

Title: Distributed Health Data Networks: A Rapid, Systematic Approach to Verifying Machine-Generated Queries Within Different Systems

Background: Distributed health data networks allow for participants to maintain ownership over their data and to run queries/programs in the privacy of their own data environments. Some networks, such as the PCORnet Distributed Research Network, utilize machine-generated queries powered by PopMedNet that may be difficult for a human to parse. PCORnet's queries are automatically translated and directed to various relational database management systems (RDBMSs) in order to accommodate a larger variety of participants in a more rapid fashion. This presents various challenges when testing queries to ensure they perform as expected within each unique data environment. We describe a rapid, systematic approach to verify the accuracy of queries within different environments.

Objective: Create and maintain a process that continuously verifies the accuracy of machine-generated SQL queries directed to different RDBMSs/IT environments within a distributed health data network.

Methods: The approach consists of five steps: (1) understanding the intent/expected behavior of a query based on the human-supplied query input information and the researcher's intent, (2) creating a variety of fake patient records that fall within and outside the query parameters, (3) use an open-source Test Case Inserter tool (TCI) to enter patients into six test databases across three different RDBMS platforms (SQL Server, Oracle, and PostgreSQL), (4) run the given query against all six data sources, (5) review results to ensure the query runs and returns the expected fake patient records.

Using the TCI allows for (2) and (3) to be performed in a reliable fashion by any person as long as they are familiar with the data model used by a given network. Fake patient records are entered by hand within Microsoft Excel sheets, which are consumed by the TCI. The TCI also consumes a separate Excel sheet defining the structure of the target data model. When records are entered into databases, the TCI will roll back all changes if any single insert fails. This provides reliability of data without the need for a SQL expert's direct involvement.

To simulate the variety of RDBMSs, tests are performed in two different versions of SQL Server, Oracle, and PostgreSQL each. This testing can be started and concluded in as quickly as one day, allowing for minimal delay between query design and distribution.

Findings: This testing process has successfully identified unexpected query behavior in various conditions, most of which are fairly complex involving joins with patient-record information across several tables, or utilizing different encounter time windows with different associated patient information within a single query. Unexpected query input information (or unexpected amounts of query information) can be a major contributor to errors in resulting query behavior, although this is partially avoided in PopMedNet due to standardization of query input information.

Conclusion: Rapid, accurate query verification is vital to accommodate the diverse, fast-paced nature of distributed health data networks. This novel testing approach is designed specifically for this fast-paced and occasionally unexpected nature. It is extensible and flexible so that it can be applied to any data model and any type of health data network.

1998
Leveraging an Enterprise Data Warehouse to Support Clinical Workflow: Targeted Notification of Pathology Results

Michael Eubank¹, Janice Sung, MD¹, Krishna Juluru, MD¹
¹ Memorial Sloan Kettering Cancer Center, New York, NY

Introduction
Radiologists perform biopsies that guide patient care. The resulting pathology report (PR) may take several days to complete, requiring frequent checks of the electronic medical record (EMR), a time consuming process prone to lapses. Timely communication of results is necessary to minimize patient anxiety and guide treatment decisions. Institutional Enterprise Data Warehouses (EDW) are typically used for operational data analytics, but some recent applications have shown value in clinical and research support (1,2). We hypothesized that our EDW infrastructure and underlying data could be used to develop an application that would deliver automated notifications of PRs to a radiologist who generated the radiology procedure report (RPR).

Methods
The application was built using IBM DB2, Java, and open source Jquery JavaScript library. In this retrospective study, we assessed application activity from December 1, 2015 to January 31, 2017. Radiologists and pathologists generated and signed RPRs and PRs, respectively, per routine. The EDW received date of procedure (from Radiology systems), date of specimen (from Pathology systems, distinct from date of PR), text of all reports, and names of all radiologist signatories (residents, fellows, and faculty) in the RPR. When date of procedure in the RPR equaled date of biopsy specimen in the PR, a match was identified, and the application sent an email notification to the radiologist(s), with link to PR. Usage data was obtained from the application. Clinical impact was assessed through an anonymous, online survey distributed in March, 2017 to all participating radiologists.

Evaluation
During the study period, 52 radiologists participated in the service. The application tracked 6,487 patients and delivered notifications involving 12,047 pathology specimens. The average time from radiology procedure to PR was 5.5 days. 100% of notifications were sent within 1 day of PR creation. 30 radiologists completed the online survey (30/52 = 58%). Radiologist subspecialties were breast imaging (67%), interventional radiology (17%), body imaging (16%). Survey respondents included radiology faculty (85%) and fellows (17%). To the question, “Automated pathology results helped me deliver better patient care,” 97% of respondents answered ‘Strongly agree’ or ‘Agree’, with an average time savings of 3 hours per week. To the question, “I plan to continue using the automated pathology results system,” 97% of respondents answered ‘Strongly agree’.

Conclusion
We demonstrate in this study one application of EDWs in near-real time clinical support. The system saves the Radiologist time by eliminating the need to regularly check for pathology results, and also enables Radiologists to act on and communicate results more quickly, thereby staying integral to patient care.

References
Visual Query Combinators for BioMedical Query Mediation
Clark Evans, Charles Tirrell, Oleksiy Golovko, Andrey Popp, Jason Simeone,
Kyrylo Simonov PhD, Leon Rozenblit JD, PhD
Prometheus Research, LLC, New Haven, CT

Introduction. Even as informaticists continue to deliver better and larger central data stores to support research, three important goals have remained out of reach for biomedical research teams: (1) the ability for all team members, including ones who are not database specialists, to proficiently explore data stores for information relevant to their research questions; (2) frictionless sharing of resulting queries to serve as documentation for further refinement and discussion; (3) deep integration of meaningful statistical analysis with the queries. Lacking these capabilities, research teams must involve database specialists to mediate even the most basic database queries and to periodically refresh results. We evaluate whether all three goals can be achieved, and the need for mediation reduced, through an innovative query system.

Innovation. Prometheus Research has designed a combinator-based database query model (“Rabbit”) [1] and released an open source implementation [2]. Rabbit is a practical, declarative query language designed for domain experts. Unlike SQL and LINQ, Rabbit’s model lends itself to dynamic, incremental, visual query construction. Moreover, the Rabbit query model flexibly accommodates all sorts of statistical and domain specific operators. Queries in Rabbit notation can be understood and shared among all research team members, reducing the need for upstream query mediation. We include a visual query tool based on Rabbit in our open source RexDB system [3].

Methods. To validate our claims about the completeness and learnability of this visual query system, we conducted usability tests on a convenience sample of participants that included experienced data analysts and novice users. Each subject was given an unfamiliar data model and 3 minutes to explore the query interface. Subjects then answered a series of challenging timed questions about the data. For example, subjects were asked to select a table, add a column from a related table, perform a maximum value aggregation on the related table, and filter the primary table records based on this aggregate value.

Results. Novice query users with no SQL experience were able to generate complex queries using our interface, including multi-table joins with aggregation, advanced filtering, and distinct operations. Novice query users were able to complete an average of 71% of the 15 assigned tasks in under three minutes. We found that moderately technical query users with minimal SQL experience were able to complete 93.3% of the assigned tasks in under three minutes. Exceeding our expectations, we found that novice users were only 53.5% slower, on average, than more experienced data analysts.

Conclusion. The preliminary results suggest that a combinator-based visual query builder can empower a broader range of users to retrieve complex data from rich data stores. We expect to demonstrate in future research that this approach also facilitates sharing and refinement of queries, as well as the integration of statistical procedures into the queries themselves. Because our implementation uses a statistical processing language, Julia, deep integration of querying with with statistical analysis is technically straightforward.

References
3. RexDB Database Platform http://rexdb.org
Clinician, EHR Alerts, and Patient Outcome in Primary Care Setting
Mohammad A. Faysel, PhD¹, Tameca Miller, MS¹
¹SUNY Downstate Medical Center, Brooklyn, New York

Introduction

The use of Clinical Decision Support (CDS) tools in primary care settings has been increasing dramatically in the recent years, especially in managing patients with chronic conditions. However, their impact on patient outcome has not been widely studied. The purpose of this retrospective study is to evaluate whether clinician’s acknowledgement and adherence to alerts generated by CDS tools that are embedded within the electronic health record (EHR) system improve the outcome of patients with chronic diabetes.

Methods

De-identified one month clinical data of elderly (65 or older) diabetes patients with hemoglobin A₁C (HbA₁C) ≥ 6.5 were extracted from the clinical database of a multispecialty outpatient clinic that also provides primary care services. For each of these patients, a CDS alert indicating elevated HbA₁C was generated that required primary care clinician’s attention and further clinical intervention(s). To comply with the widely used National Institute of Diabetes and Digestive and Kidney Disease (NIDDK) guidelines, the organization’s protocol requires repeating HbA₁C test every 3 months for diabetes patients and every 6 months for patients whose blood sugar is well controlled. We used patient records whose alerts were acknowledged, and the records with ignored alerts to verify whether the clinician’s acknowledgement to CDS generated alerts improves chronic diabetes patient outcome.

For the group whose alerts were acknowledged, we claimed that due to clinician’s acknowledgement of the alerts, average HbA₁C value after alerts would be improved (lowered) compared to the average of the last HbA₁C values prior to the alerts. Our null hypothesis for this case was that the average HbA₁C would remain unchanged or worsened (elevated) after clinicians’ acknowledgement of the alerts. A paired t-test was performed to verify our claim. For the records whose alerts were ignored, we also claimed that the average HbA₁C value after the alerts would be worsened (elevated) compared to the average of their last HbA₁C values prior to the alerts due to clinicians’ ignoring of the alerts. Our null hypothesis for this case was that the average HbA₁C would remain unchanged or improved (lowered) before and after the alerts. We performed a paired t-test to test to verify our claim for this group. All the data analyses were performed, and statistical values and 95% Confidence Intervals (CI) were generated using SAS 9.4 software.

Results

Out of total 264 alerts, 191 (72.3%) were acknowledged by clinicians and 73 (27.7%) were ignored. For the group whose alerts were acknowledged, the paired t test produced a t value of 2.9 with a probability value of 0.004 (p-value < 0.05) which is sufficient to reject the null hypothesis and supported our claim. It also generated an HbA₁C mean difference of 0.24 (95% CI, 0.08-0.41) between before and after acknowledgement of the alerts. For the group whose alerts were ignored, the paired t test produced a t value of 1.87 with a probability value of 0.07 (p-value > 0.05) which supported the null hypothesis for this group and rejected our claim that the average HbA₁C value of this group after the alert would be worsen (elevated) compared to the average of their last HbA₁C values prior to the alerts. This test also showed that the average HbA₁C improved (lowered by 0.17) after alerts compared to the average of the last HbA₁C values prior to alert (95% CI, -0.01-0.36).

Conclusion

The study showed improved average HbA₁C values for patients whose CDS generated alerts were acknowledged by their clinicians and the group of patients whose CDS generated alerts were ignored by their clinicians thereby, suggesting that clinicians’ acknowledgement and adherence to embedded CDS generated alerts for chronic diabetes patient management do not have any effect on improvement in patient outcome. This could be due to the study dataset’s lack of detail information on clinicians’ adherence to organizational protocols. Future steps of the study will address this issue and verify the study claims using larger datasets from multiple sources.

References

1. Epic Systems Corporation EpicCare Ambulatory. IU 2 edn, 2006

2001
Using EHR Clinical Decision Support to Improve Emergency Department Care for Children with High Risk Asthma

Daria F. Ferro, MD¹, Levon Utidjian, MD¹, Joseph J. Zorc, MD¹
¹Children’s Hospital of Philadelphia, Philadelphia, PA

Introduction
Asthma is the most common chronic childhood illness and a leading cause of pediatric hospitalization. Although evidence suggests that intervention strategies such as preventive controller medications, asthma action plans, and community-based programs can improve outcomes, delivery in busy clinical settings has been challenging, and frequent asthma emergency department (ED) visits and hospitalizations remain a challenge. We sought to bundle clinical decision support (CDS) tools within our electronic health record (EHR) to identify at risk patients and provide clinical decision support to the healthcare team to deliver a multidisciplinary intervention in the ED setting.

Methods
This project was developed as part of a multi-disciplinary Asthma Population Health Initiative that has focused its initial efforts on identifying high risk patients in a variety of settings across a large pediatric care network. It has also included the development of a pediatric asthma registry using EHR data.

We aimed to improve care for pediatric patients with asthma who were seen in the ED and subsequently discharged home. A multidisciplinary team developed a group of CDS tools driven by criteria from the pediatric asthma registry. High risk patients were defined as having at least 2 prior admissions for asthma in the past year and presenting to the ED with a chief complaint of respiratory distress. Primary outcomes include: (1) increasing healthcare team awareness and education in the ED; (2) providing an asthma care plan in all patients with persistent asthma; (3) initiating inhaled corticosteroids (ICS) in patients with persistent asthma symptoms without a prior ICS prescription; and (4) referral to case management for assessment of need for community asthma resources.

The CDS bundle consisted of vendor native EHR tools such as pop-up alerts and dynamic order sets that encouraged providers to alert other members of the healthcare team (respiratory therapy, nursing, case management) about the patient’s high risk status. The CDS tools went through an iterative design process with clinical domain experts in multiple roles (physicians, nurses, respiratory therapists) and the CDS design committee. We adopted low cost techniques to accelerate the iterative process including close monitoring of CDS behavior and user actions, as well as user directed retrospective aided recall. With the latter, we showed users screenshots of CDS as they would have seen at the time of their interaction and conducted semi-structured interviews. The CDS was modified accordingly in response to feedback.

Results
Preliminary data of the ED asthma CDS bundle shows increases in education ordering suggesting increased team awareness (2% baseline to 20%), increased asthma care plan filing (4% baseline to 11%), and increased case management referrals (3% baseline to 80%). Prescriptions of inhaled corticosteroid has remained high. Qualitatively we have identified several aspects of CDS development and implementation in need of improvement such as the need for high fidelity testing environments and the importance of user feedback after initial implementation.

Discussion
Through this work we have learned that close monitoring of CDS performance and low cost solicitation of user feedback can lead to rapid iteration and optimization of CDS tools. These tools can be combined in effective ways to address complex pediatric patients, such as those with asthma, with complex care challenges in an ED setting. Further data will be presented as data collection is ongoing.
Honest Broker Process Using System-Wide Permanent Research Identifiers

Robert W. Follett\textsuperscript{1,2}, Theona T. Tacorda\textsuperscript{1,2}, Javier Sanz\textsuperscript{1,2}, and Douglas S. Bell MD, PhD\textsuperscript{1,2}

\textsuperscript{1}Department of Medicine, UCLA David Geffen School of Medicine, Los Angeles, CA, and \textsuperscript{2}UCLA Clinical & Translational Science Institute, Los Angeles, CA

Introduction

Healthcare institutions are increasingly providing electronic health record (EHR) data for research purposes. Data is typically provided to researchers through an honest broker process, which is intended to protect patient privacy. The honest broker process often involves generation of a study id, which enables linkage of patient records without using protected health information (PHI).

In many cases, researchers need only de-identified information, but they often also need to associate data from the same patients across multiple files as well as from subsequent data extracts for the same project. In other cases, researchers may need identified information in order to perform chart review but PHI must be kept to a minimum during the provisioning process. Conversely, across separate projects, patient identities must be protected such that records cannot be linked without specific IRB approval. However, following approval, it may be necessary for investigators to collaborate across different projects with little effort.

We designed an honest broker process that is flexible enough to allow for all of these scenarios. Investigators do not have to create their own study id, and can easily separate PHI from non-PHI.

Methods

We assign a permanent global research identifier for every UCLA patient during the extract, transform, and load (ETL) process into our data warehouse. We also maintain a table used to track each research project or data requisition request. In addition to a project id, brief description of the project, and key project personnel, the table contains an encryption key used to generate a study-specific patient ID (study id) by encrypting the global research ID as well as the encryption method used.

Using our system of encryption, we can generate the same study ids for the same patients in subsequent or supplemental data pulls for a given project, without revealing the underlying global research ID. This system also enables us to unencrypt a study id if re-identification of patients becomes necessary and is IRB-approved. Separately tracking the encryption method for each research project gives us the flexibility to use different encryption functions and methods as the technology available in our ecosystem evolves.

When provisioning identified information, we supply one table or file which contains all PHI and the study id. Each additional file contains the study id with non-identifiable data. The study id is the key to linking the files together. This isolation of PHI makes it easy for an investigator to de-identify their data after PHI is no longer needed. When provisioning de-identified information, we simply use the study id in all files.

Results

We began using this system of encrypted global unique identifiers in January 2016. We have used the ID encryption process to provision more than 307 million records across 38 separate projects to date. For three projects, we have pulled daily or weekly incremental data an additional 23 times. We have pulled supplemental data for an additional two projects. No projects to date have needed to re-identify the encrypted global unique patient IDs or to link them across different studies.

Conclusion

We have implemented a system of creating permanent global research identifiers for the UCLA patient population and for providing studies with encrypted versions of this identifier. We are using this system successfully to link patient identities across sequential data sets within studies. It will also enable patient linkage across studies, if needed.
Crowdsourcing Evidence-Based Medicine (EBM)

Paul Fontelo, MD, MPH; Fang Liu, MS
National Library of Medicine, Bethesda, MD

Abstract

We developed a new crowdsourcing method to search, review, and rate recent evidence from PubMed citations. It provides an open environment for users to not only discover and rate the usefulness of a citation, but also read relevant citations previously rated by other clinicians. This tool harnesses the wisdom of the crowd and may help healthcare professionals find high quality evidence from PubMed, more efficiently.

Introduction

Searching high quality relevant citations in PubMed sometimes is a challenge for healthcare professionals because of the huge number of citations retrieved for a specific topic and lack of enough time to review them. Previously, we developed Web tools and Apps, like PICO and askMEDLINE, for helping users build the search query. A rating citation tool has been implemented in several of our existing utilities to evaluate the efficiency of the search tool and the usefulness of the retrieved citation. An open rating system through crowdsourcing may further help healthcare professionals to obtain the high-quality citations from PubMed, more quickly.

Methods

The crowdsourcing evidence project is designed as a Web tool\(^1\). The search and review interface (Figure 1) allows the user to search using PICO or simply posing a question and limited to publication type (randomized controlled trials, systematic review, meta-analysis, etc.) through PubMed’s E-Utilities API. After reviewing individual citations, the user can rate the usefulness and rationale for the score using a 7-point Likert scale. Previous crowdsourced evidence may be reviewed by the Archives with a topic filter or a score filter (Figure 2). The most recent reviewed citations are also available for browsing.

![Figure 1.](image)

![Figure 2.](image)

Result and Conclusion

We built a crowdsourcing evidence system and invited several clinicians to test the tool. Early feedback is positive. Users have the option of finding and rating as many as they find useful or only those they deem as useful clinically. The tool may provide some practical clinical information for decision-making powered by the crowd.

References

Medication Errors Generated When Using Computerized Provider Order Entry Systems in Pediatrics: A Systematic Review

Niamh E. Forde\(^1\), Clare L. Tolley MP\(\text{Pharm}\)\(^{1,2}\), Katherine L. Coffey\(^1\), Dean F. Sittig PhD\(^3\), Joan S Ash PhD\(^5\), Andrew K. Husband PhD\(^1\), David W. Bates MD MSc\(^5,6\), Sarah P. Slight MP\(\text{Pharm}\) PhD PGDip\(^{1,2}\)

\(^1\)School of Medicine, Pharmacy and Health, Durham University, U.K.; \(^2\)Newcastle upon Tyne Hospitals NHS Foundation Trust, U.K.; \(^3\)School of Biomedical Informatics, The University of Texas Health Science Center at Houston, TX, USA; \(^4\)Oregon Health & Science University, Portland, OR, USA; \(^5\)The Center for Patient Safety Research and Practice, Division of General Internal Medicine, Brigham and Women’s Hospital, Boston, MA, USA; \(^6\)Harvard Medical School, Boston, MA, USA.

Introduction: Pediatrics are particularly vulnerable to medication errors and incidents are all too common.\(^1\) Computerized Provider Order Entry (CPOE) with Clinical Decision Support (CDS) has been associated with a reduction in medication errors.\(^2\) However, some new types of errors have also emerged with the use of these systems. We performed a systematic literature review to identify and understand the factors that contribute to medication errors associated with the use of CPOE in pediatrics and provide recommendations for improvement.

Methods: Our review was conducted according to the PRISMA guidelines and was registered with PROSPERO (CRD42016039984). We included primary research studies published in English that discussed the types and/or causes of errors generated when using CPOE systems. The timeframe was not restricted. Non-peer-reviewed publications, editorials and commentaries were excluded. We searched three large databases, the Cumulative Index Nursing and Allied Health Literature, Embase and Medline using broad search terms related to ‘CPOE’, ‘Errors’ and ‘Pediatrics’. Three authors independently reviewed the titles; abstracts and full texts were then reviewed independently by two authors, with one acting as a constant. Data were extracted onto a data extraction sheet. We then developed a narrative synthesis of all eligible studies.

Results: We identified 414 papers of which 47 (44 full texts and 3 conference abstracts) published between 2004 and 2015 met our inclusion criteria. Table 1 outlines five key factors that contributed to the occurrence of medication errors made when using CPOE systems and potential recommendations.

<table>
<thead>
<tr>
<th>Key Factor</th>
<th>Example</th>
<th>Recommendations</th>
</tr>
</thead>
</table>
| Lack of drug dosing alerts          | • Failure to warn clinicians about errors because dosing support was not active.  
• Absence of patient weight in the system prevented safety checks. | • Utilize dosing support         
• Up-to-date weight should be obtained before alerting. |
| Generation of inappropriate dosing alerts | • Presentation of the maximum dosing alert only if the highest possible dose on the system was exceeded, regardless of the drug indication | • Indication based dosing decision support or order sentences |
| Inappropriate drug duplication alerts | • False-positive warnings due a lack of consideration of the route, drug name, strength or time of administration | • More sophisticated context-specific CDS. |
| Drop-down menu selection            | • 1000X overdose as a result of mis-selection from a drop-down menu       | • Indication included next to specific drug dose |
| Inappropriate system design         | • Lack of available dosing options on the system                         | • Evaluation of workflow         
• ‘Select all’ function in order sets | • Limited use of select all functionality                                 |

Table 1: summary of results and recommendations

Conclusions: This review outlines five key factors which contributed to the occurrence of medication errors associated with use of CPOE in pediatrics. The use of CDS, and design safeguards, are critically important.

References:
Improving Ambulatory Care Patient Safety for Community Acquired Pneumonia through a New Electronic Clinical Decision Support

Maysoun Freij, PhD, MPH,1 Prashila Dullabh, MD,1 Richard Shiffman, MD, MCIS,2 Nitu Kashyap, MD,3 Rajdeep Brar, MD,2 Katherine Donaldson, MHI,1 Sarah Islam, MSPH,1 Richard Ricciardi, PHD, NP, FAANP, FAAN4

1NORC at the University of Chicago, Bethesda, MD, 2 Yale University, New Haven, CT, 3Yale New Haven Health System, New Haven, CT,
4Agency for Healthcare Research and Quality, Rockville, MD

Overview of Methodology

Inspired by the Agency for Healthcare Research and Quality’s (AHRQ) technical brief “Patient Safety in Ambulatory Settings”, and funding from AHRQ (contract HHSP233201500023I), the project specifically addresses the safety concern area of diagnostic errors in ambulatory settings through the use of health information technology (HIT). HIT provides several tools to improve patient safety, yet challenges to adoption in ambulatory settings prevent more widespread use. Therefore, investigators developed an electronic clinical decision support (CDS) tool for providers in the process of diagnosing the highly prevalent and sometimes catastrophic condition of community acquired pneumonia (CAP), and evaluate its implementation in two ambulatory care settings.

Investigators adapted the CURB-65 instrument for determining the severity of CAP into an electronic CDS based upon an analysis of the workflow in two ambulatory care practices: an emergency department (ED) and a primary care practice (PCP). The validated CURB-65 provides guidance in making an accurate CAP diagnosis and determining an appropriate treatment in outpatient or inpatient settings using five indicators: confusion, uremia, respiratory rate, blood pressure, and age 65 years or older. The workflow analysis in each setting led to two CDS designs: an interruptive alert for the ED, and a passive alert in the PCP. Prior to its implementation in each practice, the usability of the CDS was tested in a laboratory setting with nine providers; refinements were made to the design to clarify the necessary actions required of the user.

The CDS was implemented in the two settings from March-July 2017, and an evaluation was designed to assess successes and challenges with its implementation in the two real-world settings. Site visits were conducted in June 2017 to conduct interviews with providers about the design and utility; and the frequency of usage of the CDS was monitored throughout the pilot period. Investigators explored the extent to which the CDS was integrated into the practice’s workflow, and how usage varied across the two sites.

Evaluation Results

Based upon the qualitative interviews conducted four months post-implementation and data on CDS usage, the ED setting had much higher usage of the CDS than the primary care practice, and providers in this setting were much more accustomed to engaging with alerts in their practice than the primary care setting. Interviews suggest that changes to the sensitivity and contextual relevance of when and why the CDS is activated would improve its usability and ultimate utility to the providers. In addition, providers recommended removing the hard-stop interruptive alert, and making the valuable tool simply available on-demand.

Conclusion

While there may be validated diagnostic tools to improve patient safety, integrating them into EHRs in different types of ambulatory settings must take into account contextual differences in order to be effective. This project informs how diagnostic decision support tools can be leveraged as part of EHRs to make decisions that impact the disposition of patients with CAP.

References

Using Systems Engineering Methods to Identify, Assess, and Mitigate Preventable Harm as Part of a Patient Safety Learning Laboratory

Theresa E. Fuller¹, Anuj K. Dalal, MD¹,², Dominic J. Breuer, MS³, Pamela M. Neri, MS⁴, David W. Bates MD, MSc¹,², James C. Benneyan, PhDb

¹Brigham and Women’s Hospital, Boston, MA; ²Harvard Medical School, Boston, MA; ³Northeastern University’s Healthcare Systems Engineering Institute, Boston, MA; ⁴Partners Healthcare Information Systems, Boston, MA

Introduction: Adverse events are a significant burden on patients, families, and health systems¹. As part of a Patient Safety Learning Laboratory (PSLL), we are implementing an integrated suite of electronic patient- and provider-facing tools for the inpatient setting designed to identify, assess, and mitigate preventable harms in the acute care setting. This large and dynamic system combines myriad users, technologies, and environments, spanning multiple clinical services, and units. In this study, we describe the systems engineering methods employed during implementation of our tools that have allowed us to prospectively identify potential failures, and retrospectively learn from actual safety events.

Methods: We assembled a team of systems engineers, human factors experts, clinicians, and data analysts, who recommended two risk management methods: 1) prospective Failure Modes Effects Analysis (FMEA), and 2) retrospective Root Cause Analysis (RCA) assessments. In FMEA, potential failures were brainstormed according to the Systems Engineering Initiative for Patient Safety (SEIPS) domains² (users, technology, organizational, and environmental) with input from researchers, hospital administrators, clinical leadership, and unit staff in structured sessions. Failure modes were graded with Risk Priority Numbers (RPN), and then reviewed with the main research team. Event-based “RCA huddles” were structured to align with the operational work of the hospital. The team analyzed actual adverse events (Falls, Central-line Associate Blood Stream Infections) and risky states (e.g. “Code Status presumed”) through the lens of the PSLL tools. The information gathered and lessons learned from these risk assessments were used to inform how the tools could be refined to prevent failures.

Results: The primary results of the failure analysis are shown in Table One. From the initial FMEA, education and communication failures were identified as the primary targets for high impact improvement. From an event-based RCA (CLABSI), we determined that we would need to adjust dashboard logic to trigger a high-risk state for central lines in place for 7 days or more. Most importantly, risk assessment system allowed us to further engage end-users.

Conclusion: Structured risk assessments are proving an effective way to engage clinical and administrative staff while implementing our tools. Discussion with staff has promoted awareness, helping to resolve a potential failure mode (education/communication). Our experience to date underscores the value of using system engineering methods pragmatically during the implementation of new technology to improve patient safety in a hospital setting.

Table One: Results of Two Risk Management Methods in a Patient Safety Learning Laboratory

<table>
<thead>
<tr>
<th>Domain</th>
<th>Result: Failure</th>
<th>Recommended Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>Technology</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Inaccurate content displayed (RPN: 512)</td>
<td>Empower users, conduct interviews/focus groups</td>
</tr>
<tr>
<td></td>
<td>Load time of tools (RPN: 204)</td>
<td>Prioritize loading time during development</td>
</tr>
<tr>
<td>Organizational</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Education/Communication failures (RPN: 441)</td>
<td>Engage champions, communicate vital information via “competency checklists”</td>
</tr>
<tr>
<td>Users</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Inconsistent interpretation of data reported from the PSLL team (RPN: 336)</td>
<td>Provide context for how data should be interpreted, perform teachback</td>
</tr>
<tr>
<td>Environmental</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Physical access to tools (RPN: 256)</td>
<td>Educate clinical staff, engage clinical operations</td>
</tr>
</tbody>
</table>

RCA: Why did this event occur, and how can we prevent it from happening again?

<table>
<thead>
<tr>
<th>Event</th>
<th>Root Cause</th>
<th>Recommended Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>Central Line-associated Blood Stream Infection (CLABSI)</td>
<td>Lack of tool use and awareness; Data within tool inaccurate/imprecise; Logic of the safety dashboard may not provide actionable enough information in this context</td>
<td>Perform ongoing user coaching; Engage with clinical staff about technical issues; Align dashboard logic with CLABSI prevention literature (i.e. adjust logic to alert when central lines that have been in place for &gt;7 days)</td>
</tr>
</tbody>
</table>


2007
The presence of highly similar notes within the MIMIC-III dataset
Rodney A Gabriel, MD, Sanjeev Shenoy, MS, Tsung-Ting Kuo, PhD, Julian McAuley, PhD, Chun-Nan Hsu, PhD
University of California, San Diego, La Jolla, CA, USA

Introduction
With the advent of the electronic medical record system, there became a widespread presence of copy-and-pasting when composing clinical notes. Highly similar or even identical notes are problematic when one is compiling statistics or training predictive algorithms that model the language or attributes in notes. For example, duplicates may cause an outlier detection algorithm to erroneously identify a rare condition as being common or a predictive model may erroneously identify correlations between symptoms.

Methods
We sought to explore the presence of highly similar notes within the Multiparameter Intelligent Monitoring in Intensive Care (MIMIC-III) dataset.1 We used an approximation algorithm that minimizes pairwise comparisons and consists of three phases: 1) Minhashing2 with Locality Sensitive Hashing3. Minhashing was first used in the Alta Vista search engine to detect duplicate web pages scalable to the entire World Wide Web1; 2) a clustering method using tree-structured disjoint sets; and 3) classification of these highly similar notes via pairwise comparisons of notes in each cluster. We use Jaccard similarity (JS) to calculate the similarity between notes — in which all notes within a cluster has a JS above a certain threshold with at least one other note in that cluster. Classification of notes were grouped into three defined categories: 1) “exact copies” – defined as notes that have a JS of 1.0 and are for the same patient with an identical date stamp; 2) “common output notes” – defined as notes that have a JS of 1.0 but are for different patients and/or different date stamps; and 3) “template” – defined as notes with JS greater than or equal to 0.7. We explored all clinical notes with greater than 10 words from MIMIC-III. To validate results, the JS of 10 million random pairs of notes from MIMIC III were calculated. All pairs generated with JS >0.3 were then checked to see if they were either correctly or incorrectly placed in the same cluster.

Results
First, we tested the algorithm on a sample set consisting of 500 notes, in which 54 were considered near-duplicates or duplicates with at least one other note. Our algorithm correctly clustered 94% of the notes and provided no false positives. We then explored 2,065,096 notes from MIMIC III and identified 110,029 clusters of notes with high similarity (JS greater than or equal to 0.7). Among these clusters, there were a total of 310,227 notes. Validation results demonstrated that there was only 1 note pair out of 291 random pairs that was incorrectly clustered while 99% of random pairs (110 out of 111) with JS >=0.7 were appropriately placed in a cluster. Of all the notes, 61,672 were exact copies, 26,403 were common output notes, and 222,152 were templates. Fifty random clusters were selected and upon review of those notes, there was 100% agreement between manual review and assignment from our algorithm.

Conclusion
We found that there were multiple instances of exact copies, common outputs, and template notes form the public domain MIMIC-III dataset. It is unclear for the reasons of highly similar notes, but this could be related to pervasive practice of copy-and-pasting, note template utilization, common outputted notes from automated machines, and technical errors in note processing. The identified clusters will be made available for public knowledge and research efforts.

References:
Feasibility of State Public Health Agencies Linking National Program of Cancer Registries and Birth Certificate Data: Outcomes and Lessons Learned

Roland E. Gamache, PhD MBA1, Timothy Jay Carney, PhD MPH MBA2, Mary Ann Cooney, MPH MSN2, Emily Moore, MPH2

1 Gamache Consulting, Rockville, MD, USA; 2Association of State and Territorial Health Officials, Arlington, VA, USA

Introduction

This pilot project will assess the feasibility of linking data from cancer registries and vital records systems to answer questions about birth outcomes for cancer survivors, as there have been limited studies on this subject.1,2 State public health agencies support many information systems to collect, store, and manage critical information on the health of the communities and populations they serve. These information systems are often siloed in different registries, dedicated to meeting the specific needs of particular programs and diseases. This project is a collaboration between three states, the Association of State and Territorial Health Officials (ASTHO), and the Centers for Disease Control and Prevention (CDC).

Methodology

Three states participating in this pilot project will link previously collected National Program of Cancer Registry (NPCR) data of female residents diagnosed with a reportable cancer between the ages of 0 – 20 years with the birth certificate files of those women’s subsequent children born in the same state using a probabilistic approach. In each state, a comparison group of female non-cancer survivors with a live birth will also be selected. The linkage protocol will utilize the Link Plus, a probabilistic linkage program developed by CDC, to perform the linkages.3 Manual review and Lexus-Nexus Accurint software will help determine true matches.

Results

Results from this project will yield information on the utility of the protocol, threshold values for the linking algorithm, and a longitudinal assessment of the quality of the matching variables over a fifteen-year period. Additionally, the number of links and the frequency of missing values will be evaluated. Furthermore, we will assess the quality and representativeness of maternal and infant outcome data of the cancer survivors. Table 1 represents data for each of the three states and the number of potential comparisons.

Table 1. Number of Potential Comparisons

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>State 1</td>
<td>2,762</td>
<td>197</td>
<td>3.354</td>
<td>141,377</td>
<td>133,947</td>
<td>132,409</td>
<td>130,280</td>
<td>128,748</td>
<td>130,946</td>
<td>132,951</td>
<td>2,260,170</td>
</tr>
<tr>
<td>State 2</td>
<td>2,591</td>
<td>185</td>
<td>3.446</td>
<td>126,845</td>
<td>122,350</td>
<td>120,389</td>
<td>119,831</td>
<td>119,002</td>
<td>120,875</td>
<td>121,585</td>
<td>2,066,611</td>
</tr>
<tr>
<td>State 3</td>
<td>3,326</td>
<td>238</td>
<td>4.039</td>
<td>117,294</td>
<td>114,531</td>
<td>114,008</td>
<td>113,091</td>
<td>113,489</td>
<td>114,375</td>
<td>114,465</td>
<td>1,945,899</td>
</tr>
</tbody>
</table>

Conclusion

Lessons learned about the feasibility of this type of linkages will inform future public health surveillance programs. Further, this pilot project enhances the field’s understanding of maternal cancer on birth outcomes. Moving toward interoperable data that can be exchanged and linked across systems has the potential to provide more complete views of population health while leveraging the strengths and infrastructure of existing systems.

References

An Exploratory Study of Older Adults’ Strengths, Needs, and Outcomes Using Electronic Health Record Data in a Senior Living Community

Grace Gao, DNP, RN-BC, Madeleine Kerr, PhD, RN, Ruth Lindquist, PhD, RN, FAAN, Chih-Lin Chi, PhD, MBA, Karen A. Monsen, PhD, RN, FAAN
School of Nursing, University of Minnesota, Minneapolis, MN, USA

Problem
Living well with chronic conditions is a goal for the growing aging population. The Institute of Medicine describes the concept of living well as an achievable healthy state characterized by optimum physical, mental, and social wellbeing. The predominant problem-oriented health care infrastructure, however, creates a data gap related to the documentation of individuals’ strengths and wellbeing in electronic health records (EHRs). Without adequate data to describe a comprehensive assessment of an individual’s health, there is a lack of necessary information to generate integrated knowledge and incorporate person-centered wisdom into practice. Studies are needed to describe whole-person knowledge representation that communicates not only older adults’ problems/needs, but also their strengths; and to explore associations with health care interventions and outcomes, toward bridging the data gap that persists in problem-oriented clinical practice and documentation.

Purpose/Aims
The purpose of this study is to describe and compare strengths and needs of older adults, interventions they receive, and related Knowledge, Behavior, and Status (KBS) outcomes documented by the standardized terminology, the Omaha System, using the EHR data in a senior living community to inform care practice and documentation. The aims of this study are to (1) describe the strengths and needs of residents, nursing interventions they receive, and their initial and final KBS outcome scores in the EHR dataset; (2) examine the relationships among residents’ strengths, needs, and the interventions they receive; and (3) explore the relationships among residents’ strengths and needs, and their KBS outcomes.

Methodology
This study employs a descriptive, comparative design using existing EHR data generated from multiple facilities of a senior living community in a Midwest metropolitan area. A convenience sample comprises de-identified data from documentation of strengths, needs, interventions, and outcomes for over 900 assisted living residents. Descriptive statistics will be used to summarize strength and need attributes, interventions, and KBS outcomes. The sample will be split between those with a higher vs. lower proportion of strengths to needs. Inferential statistics will be used to compare interventions and KBS outcomes between the higher strength group and the lower strength group.

Results
Preliminary results demonstrate that it is feasible to collect comprehensive holistic data using the Omaha System as a framework for strengths, needs, interventions, and outcomes. Final results about residents’ strengths and needs attributes, relationships among strengths, needs, and interventions, and relationships among strengths, needs, and KBS outcomes associated with the three research aims will be reported in the poster.

Conclusion
The study will discover new knowledge regarding relationships among strengths, needs, interventions and outcomes of older adults and create a platform for further research use of a whole-person, strength-based ontology to expand beyond a problem-oriented information structure in practice and documentation.

References
Leveraging Crowdsourcing to Help Classify Social Media Data for Medical and Patient Safety Insights

Alex Gartland¹, Arooj Akhtar, PharmD, MBS², Richard Le¹,2, Greg E. Powell, PharmD, MBA³
¹University of North Carolina at Chapel Hill, Chapel Hill, NC; ²ZeroChaos, Orlando, FL; ³GlaxoSmithKline, Research Triangle Park, NC

Abstract
Classifying social media data for insights is time consuming and costly. This research evaluates whether crowdsourcing, via Amazon Mechanical Turk (MTurk) can be used as a platform to accurately and efficiently classify basic content in social media posts. We concluded that crowdsourcing is a cost effective and efficient method for classifying basic social media content which allows trained drug safety experts to focus on assessing social media data for medical and patient safety insights.

Introduction
The classification/curation of social media data for use in drug safety is both time consuming and costly. This research aims to demonstrate how crowdsourcing can be used to accurately and efficiently classify social media data, potentially reducing associated cost, allowing trained drug safety experts to focus their time and expertise on assessing social media data.

Materials and Methods
A dataset consisting of 15,000 uniquely authored social media posts was previously curated by GlaxoSmithKline drug safety professionals and served as the gold standard for comparing crowdsourced classifications. Amazon Mechanical Turk (MTurk) was chosen to crowdsource basic classification of social media posts using the gold standard dataset. The intent was to measure cost, accuracy and time (Phase one) and quality and time (Phase two). Phase one, consisting of three identical batches of 500 randomly selected posts, was tested at three variable price points -- $0.03, $0.03 with a $1.00 bonus per 50 posts, and $0.05. Each post was worked upon by three unique MTurk workers (Turkers) and compared to the gold standard using a majority voting system. Phase two consisted of 5,000 randomly selected posts where each post was reviewed by one Turker and then compared to the gold standard. To help ensure quality we implemented measures to detect (e.g. use of bots) and then applied an accept/reject system to posts completed by Turkers.

Results
The results from Phase one yielded an overall percent accuracy with the gold standard as follows: 92.7% for the $0.03 batch, 92.7% for the $0.03 batch with a bonus, and 92.4% for the $0.05 batch. All three batches were completed, reviewed, and accepted/rejected within approximately 72 hours of posting. Phase two took less than 48 hours and had an overall percent accuracy of 91.6%.

Discussion
These results show that the average Turker can review and classify social media posts for poster type and other basic content of interest with over 90% accuracy as compared to a trained drug safety expert at a fraction of the cost. Additionally, it was found that there was no substantial difference in accuracy when considering the compensation rate for an assignment. Additional research is required to further assess the value of using crowdsourcing to evaluate various social media types (images, video, etc.) and to perform more advanced classification/curation for the purpose of obtaining drug development and safety insights.

Conclusion
Crowdsourcing is a cost effective and efficient method for classifying basic information contained in social media posts which allows trained drug safety experts to focus their time and expertise on assessing social media data for medical and patient safety insights.
Clinicians’ Perceptions of Usefulness of the PubMed4Hh App for Clinical Decision-Making at the Point of Care

Kyungsook Gartrell, RN, PhD1, 2, Caitlin W. Brennan, PhD, APRN2, Fang Liu1 MS, Gwyneth R. Wallen, RN, PhD2, Paul Fontelo, MD, MPH1
1National Library of Medicine, Bethesda, Maryland; 2National Institutes of Health/Clinical Center Nursing Department, Bethesda, Maryland

Introduction

Evidence-based practice in healthcare has been facilitated by ubiquitous wireless mobile devices. Access to online health information affects clinicians’ decision-making, through improvement in knowledge. Point of care decision support tools provide clinicians with useful information in a timely manner for informing patient care and improving patient-physician communication. However, research evidence is still lacking on the effectiveness of clinical decision support for clinicians at the point of care.

PubMed4Hh is a Web app as well as an app for iOS and Android devices from the National Library Medicine (NLM) for finding relevant health information from biomedical literature. The purpose of this study was to explore how evidence provided as abstracts and the bottom-line (TBL) summaries, accessed with mobile devices in wireless network, affects clinical decision-making at the point of care.

Methods

This pilot study was conducted at the National Institutes of Health Clinical Center (NIH CC) from February 2016 to January 2017 as a collaboration between the Lister Hill National Center for Biomedical Communications, NLM and the NIH CC Nursing Department. The original PubMed4Hh website was modified to include features which highlighted key terms, as well as an online survey questionnaire for clinicians to rate the usefulness of abstracts and TBLs. NIH clinicians’ group e-mail distribution lists were used to send invitations and reminders to participate in the study. Participants were instructed on how to search and were asked 1) to rate the usefulness of retrieved TBLs and abstracts, 2) the reason of scoring on a 7-point Likert scale for the usefulness rating, and 3) the location of use of the PubMed4Hh (e.g., point of care or office). The three questions were collected anonymously and archived on a server at the NLM. The study was approved by the Office of Human Subjects Research Protections of the NIH and exempted from full Institutional Review Board review.

Results

Study participants reviewed 34 abstracts and 40 TBLs. Ninety-nine percent (73/74) of scores were based on the review of only one citation. The overall usefulness Likert score (1=the least useful, 7=the most useful) was 6.00±1.15, and there was a statistically significant difference in mean scores for abstracts (6.38±0.74) and TBLs (5.68±1.33): t (62.76) = -2.88, p=0.005. The Likert scale mode for abstracts and TBLs were both ‘7’ (18/34=52.9% for abstracts and 13/40=32.5% for TBLs). Eighty-nine percent (66/74) of usefulness reviews provided rationales for the rating Likert score. The most frequent reason (27.3%) was that it confirmed current or tentative diagnostic or treatment plan, followed by ‘led to new diagnostic skill, additional test, new management’ (21.2%), then ‘modified clinical skill, test, or treatment plan’ (13.6%). There was no difference between abstract and TBLs for providing the rationale of usefulness. Also, there was no statistically significant difference in usefulness scores among three reasons (confirmed, led, and modified). More than half of abstracts or TBLs were used at the office (n=36, 58.9%) compared at the point of care (n=20, 35.7%). The TBLs were used more in the office (n=23, 79.3%), and abstracts were used more at point of care (n=14, 51.9%): p=0.031. There was a statistically significant difference in usefulness mean scores for point of care (6.55±0.76) versus the office (5.75±1.30): t (54) = 2.56, p=0.015.

Conclusion

Our results show that retrieving relevant health information from biomedical literature using PubMed4Hh may be useful for decision-making. In this study with clinicians’, information access at the office was higher (79.3%) than at the point of care. Abstracts were used more often but not significantly and no significant difference was found in the rationale for scoring the evidence (confirmed, led, and modified) although participants seemed to favor “confirm” more.
Clinical Information Displays in the Emergency Department:
Evaluation of Cognitive Support to Improve Patient Safety

Carolina Gatica, RN1, Carolina Diaz, RN1, Juan Pablo Salazar, MD2, Po-Yin Yen, RN, PhD3, Mario Barbé, MD4, Marcelo Lopetegui, MD, MS3,4
1Departamento de Enfermería, Clínica Alemana de Santiago, Chile; 2Departamento de Urgencia, Facultad de Medicina Clínica Alemana Universidad del Desarrollo, Chile; 3Department of Biomedical Informatics, The Ohio State University, OH, USA; 4Departamento de Informática Biomédica / Centro de Innovación e Investigación en Informática Biomédica, Facultad de Medicina Clínica Alemana Universidad del Desarrollo, Chile

Introduction. Electronic health records (EHR) have the potential to improve quality and patient safety by serving as a ubiquitous source of reliable information. Clinical dashboards are being incorporated within EHRs aiming at improving cognitive support (providing the right information, in the right format, when and where required)1. However, previous studies have reported that, despite the use of EHRs, information gaps still exist, mainly due to difficulty accessing the information1. Several efforts have been made to improve usability and workflow integration of these systems to achieve effective cognitive support. However, unlike the ability of immediately reviewing data from paper charts, clinicians must securely access the electronic health records, for a single patient at the time, and often share limited workstations.

We hypothesize that the continuous availability of timely relevant information to nurses in the emergency department could fill information gaps, resulting in a safer healthcare. This poster will present initial assessments of the pilot phase for a study that aims to test this hypothesis on a large academic hospital in Chile. We present the development of clinical information displays that integrate to our existing EHR, which provide relevant information, alerts and reminders.

Methods. Following a user-centered design process, we iteratively worked with clinical staff to identify opportunities to improve patient safety given the appropriate cognitive support were provided. Three themes emerged: frequency of vital signs measurements, assessment of patient’s risk of falling, and awareness of allergies. We defined the context of use for each intervention, designed the components for the application and evaluated accordingly. We developed a web-application optimized for tablet computers that integrated with our home-grown EHR to retrieve real-time clinical information. Only non-sensitive information was made available in the information display modules.

In this pilot, we evaluated the impact of the intervention on adherence to evidence-based nursing protocols in the emergency department (which regulate the minimum frequency of vital signs assessments) and awareness of allergies, thus aiming at improving patient safety. Due to the nature of the pilot phase, only four tablets were installed: each one outside a given room, right by the door; other rooms were not intervened. Clinical staff including nurses and nurse assistants were trained in group regarding relevant functionalities. A visual and audible alarm triggered on the app every time vital signs expired for a patient, which is variable depending on their pain levels: 60 min for 0-2/10, 30 min for 3-6/10, 15 min for 7-10/10. In order to consider repeated observations, we used mixed models with random effects, adjusting for triage.

Results. Three weeks of preliminary data were analyzed. Frequency of vital signs measurements were compared for patients in a room with tablet against patients in rooms without tablet (excluding patients who switched from a room with tablet to a room without tablet and vice-versa). Considering only measurements that did not meet the required minimum interval, those in rooms with a tablet had vital signs taken 7.2 minutes earlier than those without, adjusting for triage.

Discussion. We present an early prototype of the application with positive preliminary results. Ongoing assessment of new modules will be presented in a future report, including a perception of usefulness and acceptability, and analysis of six months of data to address potential Hawthorne effect and to evaluate alert fatigue.

References

2013
A Flexible Computational Neuroinformatics Workflow for Computing Functional Networks in Epilepsy Neurological Disorder

Arthur L. Gershon, Ph.D.¹; Bilal Zonjy, M.D.²; Curtis Tatsuoka, Ph.D.²; Kaushik Ghosh, Ph.D.³; Samden D. Lhatoo, M.D.²; Satya S. Sahoo, Ph.D.¹

¹Department of Population & Quantitative Health Sciences, School of Medicine, Case Western Reserve University, Cleveland, OH; ²Department of Neurology, School of Medicine, Case Western Reserve University, Cleveland, OH; ³Department of Mathematical Sciences, College of Sciences, University of Nevada Las Vegas, Las Vegas, NV

Introduction & Motivation. The derivation of brain connectivity measures is essential to understand cranial function & neural circuit formation, and how these change via natural aging or traumatic disorders. Epilepsy is a neurological disorder affecting more than 50 million people with seizures. Epileptic seizure events form intracranial functional networks, and accurate spatial/temporal classification of these is critical for pre-surgical evaluation of patients not responsive to anti-epileptic medication. There are, however, challenges with studying functional connectivity in epilepsy, including handling large data volumes, & modeling/storing/analyzing complex data.

We describe the creation of a novel signal data format called Cloudwave Signal Format (CSF) featuring: (1) support for ontological annotation, (2) adaptability for use with distributed file systems such as the Hadoop Distributed File System (HDFS), and (3) ability to retrieve channel-specific signal data segments for time-series analysis. The effectiveness of this new format is shown by its use within a flexible & scalable neuroinformatics computational workflow for quantitatively measuring functional connections in epileptic events. The workflow and associated computational libraries are being developed as part of the National Science Foundation (NSF)-funded Advanced Computational Neuroscience Network (ACNN).

Computational Workflow & Cloudwave Signal Format. At present, electroencephalograph (EEG) recordings of epileptic activity are stored in the European Data Format (EDF) [1]. Although EDF supports the storage of large volume of signal data in a single file, it has drawbacks, such as limited support for parallel processing and methods of storing data in such a way as to obstruct time series analyses necessary for determining connectivity.

To address these and other limitations of the EDF format, we developed CSF, and showed its compatibility with the Hadoop parallel computing technology stack over a 31-node computing cluster [2]. The first step of our workflow converts given EDF files to a collection of CSF files per user options, and reorients data values as time series.

Using the resulting CSF files, the pipeline applies user-specified correlation measures to pairs of signals input by the user as active in an epileptic event. At present, we are implementing three correlation measures to derive functional connectivity: (1) Pearson’s linear regression coefficient [3], (2) a non-linear regression coefficient by Pijn [4], and (3) a (non-linear) frequency-based measure known as mean phase coherence [5]. We compute matrices from these measures that are, after applying normalization to identify statistically significant connections, used to create network graphs that can be analyzed for patterns.

Conclusion. We describe a scalable computational neuroinformatics workflow based on a novel CSF data format for electrophysiological signal data, which we are using to derive brain functional connectivity measures in epilepsy.

Acknowledgement. This work is supported in part by the NIH-NIBIB Big Data to Knowledge (BD2K) 1U01EB020955 grant and NSF grant# 1636850.

References

2014
A Data-driven Method for the Early Identification of Diabetes and Prediabetes

Travis R. Goodwin, MS\textsuperscript{1}, Michael E. Bowen, MD, MPH, MSCS\textsuperscript{2}, Sanda M. Harabagiu, PhD\textsuperscript{1}

\textsuperscript{1}The University of Texas at Dallas, Richardson, TX; \textsuperscript{2}The University of Texas Southwestern Medical Center, Dallas, TX

Introduction

Type-II diabetes is projected to affect as many as one in three adults by 2050 in the United States\textsuperscript{1}. This is exacerbated by the fact that an estimated 8 million people with diabetes and 80 million people with prediabetes remain undiagnosed or unaware of their conditions\textsuperscript{1}, despite the prevalence of well-established diabetes screening criteria. To improve the early detection of individuals at risk for diabetes and prediabetes, we present a data-driven method that automatically infers a patient’s risk for diabetes/prediabetes at any arbitrary time in the patient’s future without relying on pre-specified criteria. Our method jointly learns how to (1) cluster patients into an optimal set of latent sub-populations based on patient’s diabetes trajectories and (2) infer a personalized predictive model of diabetes/prediabetes risk for each latent sub-population.

Methods

The data-driven model was trained using a cohort of 974 patients from the MIMIC-III clinical care database who were eventually coded with an ICD-9 code indicating type-II diabetes (i.e., 250.00). Every patient \(n \in [1, N]\) was associated with a contiguous sequence of \(T_n\) dates in which the patient had recorded laboratory tests. Each date \(t \in [1, T^n]\) was associated with (1) the set of all laboratory tests and values recorded on that date, \(L_n^t\), (2) the diabetes status of the patient, \(D_n^t\), and (3) the elapsed time \(e_n^t\) between date \(t\) and the previous date, \(t−1\). The patient’s diabetes and prediabetes status was determined using HbA\textsubscript{1C} as the criterion; as in Bowen et al. (2015)\textsuperscript{1}, diabetes was determined as an HbA\textsubscript{1C} $\geq 6.5\%$ and prediabetes as $5.7\% \leq$ HbA\textsubscript{1C} $< 6.5\%$. We implemented the predictive models associated with each latent sub-population using gradient boosting machines. The model, illustrated in Figure 1 and extended from Goodwin and Harabagiu (2015)\textsuperscript{2}, considers \(V\) as the total number of laboratory tests recorded in the database, and learns (1) the optimal assignment \(z^{*}\) of all \(N\) patients into a given number of latent sub-populations and (2) a (different) set of optimal weights for every laboratory test for the predictive models associated with each latent sub-population.

Results

We evaluated the performance of our model using 5-fold cross validation in two experimental settings: predicting (1) the patient’s HbA\textsubscript{1C} or (2) whether the patient is at risk for developing diabetes/prediabetes. In both cases, for every date \(t\) associated with every patient \(n\), we compared the predictions produced by the model given \(L_n^t\) and \(e_n^{t+1}\) against the laboratory values or diabetes status at date \((t+1)\). In all experiments, we found considering $> 8$ latent sub-populations provided no significant improvements. To assess the model’s ability to predict a patient’s HbA\textsubscript{1C}, we computed the Mean Absolute Error (MAE) between the percentage predicted by the model and the HbA\textsubscript{1C} recorded in the patient’s labs and found that, on average, the percentage predicted by the model differed from the HbA\textsubscript{1C} test by 0.73 of a percent. The model’s ability to predict diabetes/prediabetes risk was assessed as a multi-class classification problem, and measured using Precision, Sensitivity/Recall, and $F_1$-measure. Table 1 illustrates these results.

Conclusions

The ability to discover latent patient populations positively impacts the quality of the prediction results, both in terms of diabetes/prediabetes and HbA\textsubscript{1C}, leading to potentially improved screening and identification of patients with undiagnosed diabetes/prediabetes. Continued study is needed to identify the characteristics of the latent sub-populations inferred by the model to provide new perspectives on early identification of diabetes and prediabetes.

References

Vitals Risk Index: A Simplified Pediatric Early Warning Index that Relies Solely on Objective Measures

Tyler Gorham, PhD1, Steve Rust, PhD1, Jing Yang, MS1, James Lin1,2, Richard Hoyt1, Swan Bee Liu, MS1, Jeffrey Hoffman, MD1, Tensing Maa, MD1, Yungui Huang, PhD1, Simon Lin, MD, MBA1

1The Research Institute at Nationwide Children’s Hospital, Columbus, OH; 2Ohio State University College of Medicine, Columbus, OH

Introduction

Delayed intervention for deteriorating pediatric ward patients increases morbidity and mortality. Several pediatric early warning systems (PEWS) that reflect patient condition have been in existence for decades. At Nationwide Children’s Hospital (NCH), Monaghan’s PEWS is utilized to evaluate floor patients and determine the need for assessment and consultation by the ICU to reduce the likelihood of cardiopulmonary emergencies (Code Blue). The PEWS scoring system requires subjective assessment by caregivers at least once every four hours, a frequency that can be burdensome to hospital staff. Moreover, many likely preventable Code Blue events still occur despite use of the PEWS to initiate preventive measures. In this study, we sought to develop a completely objective, automated real-time risk stratification model based only on vital sign measures to preemptively identify pediatric inpatients with a high risk for Code Blue events.

Methods

EHR records for the time period between 1/1/2011 and 11/30/2016 were retrospectively reviewed. During this period, NCH had 260 pediatric encounters involving Code Blue events on the inpatient wards. The rest of 84,748 pediatric inpatient encounters were classified as controls. This data was split into two datasets, one for model development and one for model testing. All the encounters with admission dates before 2015 were employed to develop a weighted logistic regression model. The independent model variables (heart rate, O2 flow rate, O2 saturation, respiratory rate, systolic blood pressure, and temperature) were selected and the associated clinical values were converted into scores of 0, 1 or 2 based upon the deviation from the normal range suggested by the clinical experts. During model training, model stability was assessed by repeated 10-fold cross-validation. The risk score generated from the model was translated into a Vitals Risk Index (VRI) with the range from 0 to 100. Model performance was characterized on the held-out testing dataset.

Results

This testing set consisted of 63 code events, which happened outside of ICU unit, and 22,826 controls. For the test data, the VRI achieved 81.0% area under the receiver-operator characteristic (ROC) curve (AUC), and 54.0% true positive rate and 7.8% false positive rate for a VRI threshold of 42.3. Evaluated on the same test dataset, PEWS has a 73.3% AUC, and true positive rate of 33.3% and false positive rate of 7.8% for a PEWS threshold of 4. Beyond very small false positive rates, the VRI exhibits superior performance when compared to PEWS.

Conclusions

Our findings suggest that we have developed an automated ward risk stratification model, which performs better than the current warning system in identifying deteriorating pediatric inpatients in NCH. Such a tool may help perceive potential cardiopulmonary arrests while not generating too many alerts that would fatigue the caregivers. Relying solely on objective measurements would reduce nursing workload, whereas it may miss signs of patient deterioration that would be captured by subjective evaluation. Longer-term nursing workload reduction can be achieved by identifying which elements of subjective patient evaluation can be replaced by automated assessment.

Reference

State-Level Adoption of National Guidelines for Norovirus Outbreaks in Healthcare Settings: Implications for Decision Support

Carl J. Grafe, MS¹, Catherine J. Staes, BSN, MPH, PhD¹, Kensaku Kawamoto, MD, PhD¹, Matthew Samore, MD¹, R. Scott Evans, MS, PhD¹

¹Department of Biomedical Informatics, University of Utah, Salt Lake City, Utah

Introduction

Computerized decision support systems can help investigating officials to follow best practices when responding to infectious disease outbreaks,¹ but variation in local guidelines may make such systems more difficult to develop and implement. In 2011, the U.S. Centers for Disease Control and Prevention (CDC) published guidelines for the prevention and control of norovirus outbreaks in healthcare settings.² Many U.S. state and local jurisdictions have since published or updated their own norovirus or general acute gastroenteritis outbreak guidelines with recommendations tailored to the perceived needs of their individual jurisdictions. Failure to adopt national guidelines could result in the use of ineffective strategies for outbreak response and lead to disparities in outbreak data collection and reporting between jurisdictions. This study aimed to identify the extent to which state-level guidelines adhere to national guidelines for norovirus outbreak response in healthcare settings.

Methods

State guidelines were obtained from internet searches and direct email or webform contact with state public health officials in early 2016. Guidelines specific to healthcare settings were then reviewed and categorized based on alignment with elements of the CDC’s guidelines. General recommendations, such as increased hand hygiene and patient cohorting, were grouped broadly under one of the CDC’s guidelines’ twelve evidence-based recommendations for prevention and control. Specific recommendations, such as number of stool specimens or time ranges for patient isolation, were tabulated based on the specific guidance indicated. Recommendations on topics that were not covered in the CDC’s guidelines were included if they occurred in at least 10 states’ guidelines.

Results

Guidelines were obtained from 41 of 45 (91%) state health departments (the remaining five did not respond). The state guidelines generally covered the CDC’s twelve primary recommendations, but specific recommendations within those categories varied considerably. For example, there were five different recommended time ranges for isolation of ill patients, six different recommended time ranges for exclusion of ill employees, and 21 different recommended ranges for numbers of stool specimens to collect. High levels of variation were also found in recommendations that were not found in the CDC’s guidelines, including 11 unique recommendations for outbreak detection, 14 for outbreak resolution, six for determining mode of transmission, and six for conducting environmental health assessments. These varying recommendations were likely to impact key aspects of outbreak response, such as determining which outbreaks to investigate, which cases to include, and which data fields to collect. Ultimately, 31 states (76%) contradicted at least one recommendation from the CDC’s guidelines, but 22 of these 31 states (71%) had the CDC’s guidelines cited in their documents or hyperlinked on their websites.

Conclusion

This study identified substantial variation in state healthcare-associated norovirus outbreak response guidelines. Such variation could necessitate taking a state-by-state or jurisdiction-by-jurisdiction approach to decision support, with associated increases in complexity and cost. These findings are useful in determining the scope of decision support development projects and in the interpretation of national outbreak data and analyses. More research is needed to understand why this variation exists, how it might impact outbreak outcomes, and where further evidence-based recommendations or alternate methods for communicating national guidance might be beneficial.

References

Critical Appraisal of Models for Prediction of Readmission (CAMPR): A Quality Tool to Assess Models that Predict Hospital Readmissions

Lisa V. Grossman¹, Rollin R. Reeder²,3, Colin G. Walsh²,3,4, Devan Kansagara⁵, David K. Vawdrey⁶, Hojjat Salmasian¹,6

¹Department of Biomedical Informatics, Columbia University, New York, NY, ²Departments of Biomedical Informatics, ³Medicine, ⁴Psychiatry, Vanderbilt University, Nashville, TN, ⁵Oregon Health & Science University / Portland VA Medical Center, Portland, OR, ⁶The Value Institute at NewYork-Presbyterian Hospital, New York, NY

Objective

Despite the rapidly growing number of studies devoted to the development and validation of models predicting hospital readmission risk,¹,² few data exist on the quality of these models, their clinical relevance, and institutions' success in implementing them. We developed "Critical Appraisal of Models for Prediction of Readmission" (CAMPR), a quality tool that assesses the design and implementation potential of readmission risk prediction models, as well as the necessary environment to effectively implement one.

Methods

We conducted a national online survey of domain experts in the development, validation, and implementation of models for predicting readmission. We publicly distributed the survey to the American Medical Informatics Association (AMIA) mailing lists and also sent it directly to the primary authors of studies on readmission prediction from two recent systematic reviews.¹,² The survey assessed perceived barriers to model quality, use, and implementation. Two independent researchers (LVG, RRR) conducted thematic analysis of qualitative content in NVivo Version 11 (QSR International). The survey results have been previously reported.³ Based on the thematic analysis, the authors identified and classified measures to operationalize two quality dimensions of readmission risk models: (1) design, and (2) implementation.

Results

The preliminary CAMPR tool contains 51 questions, categorized into 15 key domains and further classified into two quality dimensions. Domains under design include overall quality, predictors, timeframe, heterogeneity and subgroups, data availability and quality, and clinical relevance. Domains under implementation include data availability, validation and generalizability, resources, vision and buy-in, workflow integration, clinical relevance, and continuous improvement.

Conclusion

CAMPR is a potential new tool to assess quality of current readmission risk prediction models and guide the development of future models. Our next steps include refinement and reduction of items through expert review, as well as establishing the reliability and validity of the tool by applying it to the previously published literature.

References

A Machine Learning-Based Approach for Identifying Atopic Dermatitis in Adults from Electronic Health Records

Erin N. Gustafson, PhD, Al’ona Furmanchuk, PhD, Jennifer A. Pacheco, Firas H. Wehbe, MD, PhD, Kathryn L. Jackson, Abel N. Kho, MD, MS, William K. Thompson, PhD, Jonathan I. Silverberg, MD, PhD, MPH
Feinberg School of Medicine, Northwestern University, Chicago, IL

Introduction
Atopic dermatitis (AD) is a chronic inflammatory disease associated with intense itch and skin hyper-reactivity to environmental triggers that are innocuous to non-atopic individuals. The complex etiology of AD drives a growing need to acquire large-scale cohorts with detailed granular data for clinical research. Here we present an electronic health record phenotype algorithm for this purpose, developed as part of the NIH-funded electronic Medical Records and Genomics network. Diagnostic criteria for AD rely on clinical observations of disease manifestations and laboratory measurements. AD is difficult to distinguish from other dermatoses due a complex, multi-dimensional clinical presentation (i.e., atypical signs and symptoms). As a result, estimation of AD prevalence, clear understanding of sub-phenotypes, and nationwide guidelines for AD prevention may be inconsistent or inaccurate. Our algorithm is designed to address these challenges.

Methods
This work was based on the analysis of an existing registry of 562 patients from Northwestern Medicine with at least one AD diagnosis code. Following chart review, a diagnostic criteria was applied to assign AD diagnosis (278 cases). Our approach expanded on previous phenotype algorithms using diagnosis codes by including information extracted from encounter notes using Natural Language Processing (NLP). We utilized AD diagnostic criteria to create a phenotype-specific dictionary of more than 400 AD concepts for our NLP pipeline. We opted to use lasso logistic regression (rather than a decision tree) due to its high interpretability. Counts of NLP-derived concepts and other numerical input from coded sources were log-transformed and served as features in the regression, which output a predicted class label (i.e., presence or absence of AD diagnosis). The regression was trained using 10-fold cross validation to choose the best value for the parameter lambda. This phenotype algorithm will be applied at Northwestern to identify additional AD cases in order to characterize sub-phenotypes of AD using hierarchical clustering. Within each cluster (i.e., sub-phenotype), we will examine the most influential features that determine cluster membership. For example, we predict that patients who have the same comorbid conditions (or risk factors) will fall within the same sub-phenotype, and that features representing these conditions will be highly influential for that cluster.

Results and Conclusion
The results of the lasso logistic regression improved upon previous algorithms, with a positive predictive value of 85.2%, a recall of 82.1%, and an F-measure of 83.6%. Improvements to model performance could be achieved by refining the NLP approach (e.g., capturing mentions disease change over time) or the machine learning approach (e.g., ensemble methods such as random forest classification). These initial results demonstrate that current diagnostics and understanding of AD could be significantly improved with machine learning and NLP. As patients within a sub-phenotype are likely to have similar disease severity and response to medications, identification of sub-phenotypes of AD and their unique clinical presentations has implications for developing guidelines for proactive measures in delivering adequate health services to the right risk groups.

References
Inpatient Perspectives and Information Needs for Error Prevention

Shefali Haldar¹, Sonali R. Mishra¹, Ari H. Pollack, MD¹², Wanda Pratt, PhD¹
¹University of Washington, Seattle, WA; ²Seattle Children’s Hospital, Seattle, WA

Abstract
Hospital safety has long been acknowledged as a problem in the United States. Although patients and caregivers have great potential to help prevent these errors, less is known about their perspectives and information needs regarding these quality and safety problems. To fill this knowledge gap, we surveyed 246 individuals at two hospital sites. Our analysis informs the design of inpatient tools to further involve patients and caregivers in error prevention.

Introduction: For almost twenty years, much attention has been given to the problem of preventable medical errors in United States hospitals (1). Despite this attention, a recent estimate suggests medical errors are the third leading cause of death (2). Because inpatients and caregivers are witnesses to all care they receive, they have great potential to play a key role in preventing a broad spectrum of quality and safety problems, and in avoiding harm. However, inpatients and caregivers currently do not have sufficient tools to help them identify these errors and manage their effects. To design powerful technologies for patients and caregivers to prevent error, we must first understand their experiences and know their information needs.

Methods: We created a web-based, anonymized survey asking patients and caregivers to describe their experiences with undesirable events, which we define in our survey as something that was (1) a small or big concern, (2) was unpleasant or caused harm, and (3) could have been avoided. Recruitment took place at two sites: a pediatric hospital and adult hospital. Patients and caregivers were approached in-person and were eligible if they were: well enough to provide informed consent, comfortable communicating in English, had experienced an undesirable event during a previous or the current hospital stay, and had spent at least one night in the hospital during their current visit. Participants were asked to identify as either a patient or caregiver, describe their undesirable event experience in their own words, and specify what information would have been important to know at the time of the event. Two research team members conducted an iterative and inductive qualitative analysis to examine the types of events that participants described. Descriptive statistics were used to analyze the information needs identified.

Results: Our research team approached a total of 606 patients and caregivers, of which 312 (51.5%) were eligible for participation. Out of 312, we received 246 survey responses (response rate 77.56%). Our qualitative analysis revealed 19 types of undesirable events, many of which fall outside of traditional clinical definitions of error. Additionally, a wide variety of information needs were identified, spanning all aspects of the hospital stay (e.g., introduction to care team, institution-specific resources, discharge details). The relationship between the type of undesirable event and types of information needs proved to be complicated, as each person’s experience and needs manifested in a unique way. Although every experience was different, we found patterns of information needs among patients and caregivers. For example, patients tended to favor information about their care team and daily schedules, while caregivers wanted information regarding treatment plans and who to contact for help or questions. The use of a qualitative, open-ended method revealed findings that suggest hospitalized individuals want information at a deeper level: not just the what, but also the why behind care decisions. Knowing these details can help facilitate more substantial and informed conversations with their care team as an undesirable event unfolds. However, this information is not easily accessible to patients and caregivers, and their needs go beyond what is offered by existing patient portals and Electronic Health Records. Therefore, more work is necessary for patient-facing systems to incorporate inpatient perspectives and information needs regarding safety issues in the hospital.

Conclusion: Understanding what types of problems patients and caregivers experience, and knowing the information these individuals need to prevent these problems, will help guide the design and implementation of inpatient systems that will successfully involve inpatients and caregivers in error identification, prevention, and reporting.

References
2. Makary MA, Daniel M. Medical error—the third leading cause of death in the US. BMJ. 2016 May 3;353:i2139. Available from: http://www.bmj.com/content/353/bmj.i2139
Development of a VA Clinical Trial Matching Application (POP CTMatch) to Enhance Clinical Ability to Bring Trials to Veterans

Robert B. Hall MS\textsuperscript{1}, Elly J. Cohen PhD\textsuperscript{2}, Thomas P. Bechtold MBA\textsuperscript{3}, Susan L. Colen\textsuperscript{2}, Karen E. Pierce-Murray RN BA\textsuperscript{1}, Nhan V Do MD\textsuperscript{1}, Louis D. Fiore MD\textsuperscript{1}

\textsuperscript{1}MAVERIC, VA Boston Healthcare System, Boston, MA; \textsuperscript{2}University of California San Francisco (UCSF) Department of Surgery, San Francisco, CA; \textsuperscript{3}QuantumLeap Healthcare Collaborative (QLHC), San Francisco, CA

Introduction

In 2015, the Veteran Affairs (VA) launched the Precision Oncology Program (POP) to offer genomic analysis as part of routine clinical care for patients with solid tumors. The POP program will include several applications as part of a knowledge-based suite of services including a molecular tumor board and web-based clinical trial matching application. There are multiple barriers to consider for adoption of matching programs such as difficulty of finding relevant trials and usability of available tools. We partnered with the CTMatch team at UCSF and QuantumLeap Healthcare Collaborative to develop a matching application for non-small cell lung cancer (NSCLC) on the CTMatch platform (POP CTMatch). Our goals were to identify and accurately match to trials at both VA and non-VA sites, and to create a usable interface for providers that would enhance adoption.

Overview of Application and Technical Features

POP CTMatch uses data from ClinicalTrials.gov (CT.gov) that have been tagged for NSCLC as well as basket trials in the National Cancer Institute (NCI) Physician Desk Query (PDQ) Cancer Trials Registry that have been tagged as “unspecified adult solid tumor, protocol specific”. Cancer subtype, stage, and biomarker status (EGFR, ALK, KRAS) were manually curated with agreed upon coding rules while eligibility criteria of gender, age, location, and recruiting status were automatically obtained from CT.gov data feed. The list of matching trials highlighted studies within the VA system that were actively recruiting patients with easy access to non-VA sites nationwide that could be filtered by patient’s location. Patient data, entered on a web-based form, consisted of date of birth, gender, cancer subtype, stage, and biomarker status. To provide greater currency, we focused on open and recruiting trials that were updated and verified within the past two years. Further design objectives included: displaying an easy-to-read format of relevant trial information, ordering trials by proximity to the patient location, easy access to eligibility criteria through pop-ups, links to CT.gov for additional trial information, and display of last updated and last verified dates. This prototype seeks to maximize sensitivity of the search, while future versions will include more patient relevant data and deeper trial curation to increase specificity.

Evaluation

We evaluated the pilot by assessing quality of matching and provider usability. Because of the POP CTMatch curation process, trial matches were more refined than with CT.gov based on the markers of cancer status (stage, biomarker, subtype). Comparisons of patient use cases between the applications found a larger number of relevant trial matches with POP CTMatch than search criteria entered through CT.gov. Usability was measured with the System Usability Scale (SUS) and compared to experience with CT.gov. System usability of POP CTMatch was rated higher than CT.gov. An evaluation from five VA Boston clinical providers found the average SUS score for POP CT Match was 94 ± 6.3 while the average SUS score for CT.gov was 34.5 ± 14.2 (p-value < 0.0001).

Conclusion

The POP CTMatch trial matching apparatus provides succinct results to assist clinicians in identifying novel therapeutics for cancer patients including trials targeted to genomic alterations. The approach focuses on matching via patient criteria, highlights trials available in the VA Healthcare System, and provides functionality to assist trial investigation activities. Application usability proved better than experience with ClinicalTrials.gov. Our team continues to identify means to improve the matching algorithm and system interface in order to enhance usability and greater adoption as part of the suite of Precision Oncology informatics applications.

2021
Understanding Use of Health IT to support Treatment of Youth with Substance Abuse Disorders using Evidence-Based Practices

Saira N. Haque, PhD, MHSA, Shellery Ebron, MS, Erin Mallonee, MS, Laura Dunlap, PhD
RTI International, Research Triangle Park, NC

Abstract

Analysis of experiences with technology to support implementation of evidence based practices to care for youth with substance abuse disorders for SAMHSA Youth Cooperative Agreement grantees. Uses of technology include Electronic Health Records, web-based training and technical assistance, health information exchange, mHealth and social media. Grantee-reported experiences with technology include time spent manually entering data in several systems and indicated that existing technology was not designed for their needs. Other findings included use of social media to connect to youth and associated privacy and security concerns. As the grant continues, we will continue to identify how grantees use technology and associated barriers and facilitators.

Introduction

Addressing substance abuse disorders in youth is a significant priority for the United States because youth make up a significant portion of the substance abuse treatment population. The Substance Abuse and Mental Health Services Administration (SAMHSA) has funded grant programs to support use of evidence based practices (EBPs) in caring for youth with substance abuse disorders. Caring for youth with substance abuse disorders requires that information be shared by multiple providers across organizations and systems. Health Information Technology (IT) is an important component of infrastructure that can support use of clinically appropriate EBPs. We sought to understand how grantees using EBPs to care for youth with substance abuse disorders use health information technology.

Methods

We analyzed information gathered during the course of a larger program evaluation including grantee-prepared progress reports, applications and notes from meetings, which were then coded and thematically analyzed in Atlas-ti, a qualitative analysis program. We then extracted themes related to HIT use, barriers and facilitators.

Results

Grantees noted that services were delivered across providers and settings. Thus HIT use varied within grantee organizations. Grantees used HIT as a communication device, document repository, and to promote client and family engagement. These uses were supported by Electronic Health Records (EHRs) and related systems, telehealth, mobile health, social media and web-based platforms for training. EHRs are used for documentation, reporting and decision support. Key challenges were cost, confidentiality, interoperability and functionality concerns. Meaningful use incentives do not apply to non-physician providers, so providers reported financial barriers to obtaining HIT. Confidentiality was a concern noted for care coordination in the absence of a robust health information exchange and when using social media to engage with youth and their families. Interoperability made coordination across disparate providers difficult due to a lack of uniform data standards. Finally, grantees noted that system requirements for caring for youth were not necessarily met by available systems and cited missing functionality they would prefer.

Conclusion

Behavioral health treatment programs for youth are not typically integrated, but occur in silos with substance abuse treatment often provided separately from mental health services and other systems-related services. Each of these silos has its own information systems and requirements, which were not necessarily connected through HIE. Future work could include requirements gathering for different types of HIT for those who care for this population and identification of best practices in system use for this population.

References

The Effect of Relational Coordination on Handoff Routines: A Case Study

Saira N. Haque, PhD, MHSA1, Carsten S Østerlund, PhD2, Lawrence M. Fagan, MD, PhD3
1RTI International, Research Triangle Park, NC; 2Syracuse University, Syracuse, NY; 3Stanford University, Palo Alto, CA

Abstract

We sought to understand the impact of relational coordination on the handoff routine. We identified handoff routines in the interventional cardiology unit of a community hospital and administered the relational coordination survey to their participants. We found significant differences in how those in different roles perceived relational coordination. When relational coordination between roles is low, systems and processes such as information systems must bridge the gap across handoff routines. This can be used to inform system design.

Introduction

Participants in a handoff routine must effectively communicate information about the patient and his or her expected trajectory for a successful handoff1. Handoff routines often differ from their ideal-types, but these differences do not always lead to exceptions. We posited that relational coordination is one reason why some handoff routines lead to exceptions and others do not. Relational coordination is a measure of the strength and quality of relationships between roles, including communication and shared goals2. Previous work has studied relational coordination with delivery of acute, emergency, trauma and older patients. However, the effect of relational coordination on handoffs is not as well studied. Thus, we sought to understand the impact of relational coordination on handoff routines.

Methods

We identified handoff routines and their participants in the interventional cardiology unit of a community hospital in the Midwest as part of a larger study examining coordination1. We observed handoff routines in practice and identified where handoff routines differed from their ideal-types and where there were exceptions.

We administered the relational coordination survey2 via paper to handoff routine participants. Participants in the handoff routines were in a number of occupational roles (cardiologists, nurses, technologists, administrative staff. There were 22 respondents to the survey which we analyzed using the Kruskal-Wallis method for non-parametric statistics.

Results

There was a significant difference in relational coordination across role groups (Kruskal-Wallis test, p<.05). That means the perceptions of each role about the frequency, accuracy and timeliness of communication, shared goals, mutual respect and knowledge of each other’s work vary. We found that physicians experienced the highest degree of relational coordination, followed by scrub techs and cath lab nurses. Nurses in other units experienced lower relational coordination. In all analyses, relational coordination between handoff participants was related to handoff exceptions. As relational coordination increased, the number of handoff exceptions decreased. Exceptions were closely aligned with differences for handoff routines external to the unit but not for handoff routines internal to the unit.

Conclusion

This study illuminates the importance of relational coordination between roles. Where relational coordination is low, handoff routine exceptions are more likely and systems and processes must bridge the communication gap. Where relational coordination between roles is high, then formalized mechanisms to facilitate communication are of less importance. In addition, the study highlights the importance of promoting relational coordination within and across hospital units.

References

From Scoping Review to Metadata: An Evidence-Based Approach to Identifying Informed Consent Metadata for Biorepositories

Marcelline Harris PhD RN1, Frank J. Manion PhD(c)1,2, Hsing-Yi Song MPH2, Muhammad Amith PhD(c)2, Yongqun He PhD DVM1, Cui Tao PhD2
University of Michigan, Ann Arbor MI; University of Texas Health Science Center, Houston TX

Introduction and Background The rapid growth in the number of biorepositories, attributed to technical and informatics capabilities, is characterized by increasing complexity in the management and use of the specimens, data, and information managed by repositories and the networks in which they may participate. Concurrently, there are increasing demands to introduce standard practices to assure the quality of specimens, associated data, and overall operations; nontechnical operations such as governance and oversight of informed consent processes are emphasized. On a routine basis, biorepositories deal with managing consent, collection access, specimen and data use and sharing limitations, etc. To date, research on informed consent had emphasized the perspective of those donating specimens, however, the management of informed consent information from the perspective of biorepository data and information management has received far less attention. The complex set of regulatory, legal, privacy, security requirements, and information flows involved in biorepository specimens and data, as well as paper-based workflows in legacy repositories, challenges our ability to effectively build information systems that support discovery and sharing of research data, specimens and other research artifacts at scale, and fidelity to the intent of the specimen donor.

Methodology Our overall goal is the development of an ontology of informed consent metadata in the context of biorepositories. In this project, we report on a scoping review to identify relevant metadata from the perspective of biorepository operations, and the application of NLP processes to enrich metadata content. We are conducting a scoping review, a type of systematic review that is particularly appropriate when the need is to systematically map the literature available on an evolving topic in order to identify key concepts, theories, sources of evidence, and gaps in the research. The research question driving the review is: What is the evidence base for metadata needed to manage informed consent information across biorepositories, and support best practices in the management of that informed consent information? Eligibility criteria include: a) best practice guidelines (e.g., ISBER, OECD, NCI) b) the biorepository classification literature; and c) literature that specifically addresses informed consent and biorepository operational perspectives. Literature: We conducted a broad search of MEDLINE for the timeframe Jan 2010 through current on Feb 25 2017, using combinations of MeSH terms and keywords including: (biobank OR biorepository OR biospecimen resource) AND informed consent; (biobank OR biorepository OR biospecimen resource) AND informed consent information? Eligibility criteria include: a) best practice guidelines (e.g., ISBER, OECD, NCI) b) the biorepository classification literature; and c) literature that specifically addresses informed consent and biorepository operational perspectives. Literature: We conducted a broad search of MEDLINE for the timeframe Jan 2010 through current on Feb 25 2017, using combinations of MeSH terms and keywords including: (biobank OR biorepository OR biospecimen resource) AND informed consent; (biobank OR biorepository OR biospecimen resource) AND informed consent AND classification; (biobank OR biorepository OR biospecimen resource) AND informed consent AND classification. Selection: Abstracts and articles were reviewed for suggestions of relevant metadata and data elements. Analysis: After identifying relevant content within each document, we first apply open information extraction, a natural language processing method that extracts knowledge triples irrespective of domain specificity, and then apply our custom interface for ClausIE to attain a high accurate extraction of subject, predicate, and object information. We did not pre-select the documents or modify the existing verbiage as previously done. The output produced subject-predicate-object triples. A spreadsheet that allows for online collaboration is being used to iteratively refine columns labeled ‘category’, ‘relationship’, and ‘value set’.

Results 283 abstracts were retrieved, 10 duplicates were removed, 106 abstracts were excluded for non-relevance, leaving 106 articles and documents retrieved for full text review. Our qualitative summary of the scoping review indicates that while the literature calls out the specific need for policies for managing information about informed consents relevant to biorepository specimens and data, we found few suggestions/indications of higher level metadata needs for consent management. Best practice guidelines clearly address governance of processes related to informed consent information management, but the focus is on general principles. While metadata needed to address principles are suggested/indicated, metadata needs at the detailed level of implementation are not evident. The literature on classification of biorepositories was highly focused on implementation and operational perspectives, including indications of metadata for managing informed consent information. Jointly, these three literature sources provide evidence of information and metadata needs that should be enabled by a metadata ontology. The NLP-enrichment processes are augmenting the identification of triples need for the ontology, with 2,131 entities and 614 predicates currently identified.

Conclusion Next steps involve normalization, developing the ontology and engaging domain expert stakeholders in validation.

Acknowledgements This project is supported by NIH grant NHGRI (U01HG009454).
Development & Implementation of a Clinical Decision Support Tool for Familial Hypercholesterolemia

Ali Hasnie M.D.1, Pedro Caraballo M.D.2, Robert R. Freimuth Ph.D3, Iftikhar Kullo M.D.1

1Department of Cardiovascular Diseases, Mayo Clinic 2Division of General Internal Medicine, Mayo Clinic, Rochester, MN, 3Department of Health Sciences Research, Mayo Clinic, Rochester, MN

Introduction:
The 2009 Health Information Technology for Economic and Clinical Health (HITECH) Act mandated hospitals and medical clinics to demonstrate meaningful use of electronic health records (EHR). Subsequently, there is increasing interest in leveraging the EHR to assist providers at the point of care, including reminders to screen for certain conditions [1], implement preventive measures [1], and even to assist providers in the diagnosis and management of certain conditions [1]. We sought to develop and implement a clinical decision support (CDS) tool to assist providers in caring for patients with Familial Hypercholesterolemia (FH), a relatively common genetic disorder associated with increased risk of coronary heart disease.

Methods:
We utilized a mixed methods approach to obtain physician input to inform development of a clinical decision support (CDS) tool for FH. A total of 1166 physicians were queried on their opinions on key elements of a CDS tool using an enterprise-wide survey. Specific feedback was elicited relating to: 1) the primary focus of the CDS Alert (diagnosis vs. management), 2) the utility of an order set, and 3) the ideal location within the EHR for an alert to appear. In addition, focus group discussions were held to guide the development of the tool as well as gather specific recommendations for the development and implementation of the CDS tool.

Results:
The response rate to the survey was 17.4% (n=203). The vast majority (97%) of respondents favored development of a CDS tool for FH and believed it would be helpful in the management of an FH patient. Subsequent focus group discussions revealed several key themes. Participants prioritized guidance on the initial management, diagnostic workup, and appropriate referrals of patients suspected with FH. Opinions regarding a passive alert and an active alert were divided with majority preferring a passive alert. Based on this feedback we developed a prototype CDS tool incorporating key elements listed in Table 1.

Table 1. Key elements of a CDS tool for FH

| Essential facts related to the diagnosis and treatment of FH |
| Reminder to rule out secondary causes of hypercholesterolemia |
| Reminder to screen family members |
| Guidance to initiate/optimize lipid lowering therapy |
| Alert presented upon reviewing the laboratory data (e.g. lipid profile) |
| An ORDER-SET that includes cardiovascular genomics consultation, lipid profile, and reminder to test family members |

Figure 1: Steps in developing a CDS tool for FH. The final phases will involve testing the tool and deploying it in the EHR.

Conclusion:
Based on feedback from a survey and focus discussions, we developed a prototype CDS tool for FH. We will use this tool in a sample of genetically confirmed FH patients. Usability and outcomes of the tool will be assessed and refinements made. Following this, the tool will be deployed in the Mayo Clinic EHR and implementation metrics and outcomes will be assessed. The impact of this tool could be significant given the low awareness, control and detection of FH [2].

References

2025
Screening the Contribution of Medical Reports to Electronic Diagnostic Systems

Peter J. Haug, MD1,2, Jeffrey P. Ferraro, PhD1,2, Nathan Dean, MD1,2
1Intermountain Healthcare, Salt Lake City, Utah; 2University of Utah, Salt Lake City, Utah

Introduction
In recent years, we have sought more efficient ways to build diagnostic tools from accumulated EHR data. We previously described a tool designed to largely automate the initial development of diagnostic, clinical decision support systems1. This application used disease models embedded in an ontology to drive extraction of data stored in an Enterprise Data Warehouse (EDW), to select features relevant to diseases of interest, to develop diagnostic systems for these diseases, and to evaluate the accuracy of these automatically derived systems. These rapid-turnaround prototypes help determine whether to develop a fully realized diagnostic decision support system.

Experience with several diseases2,3 has convinced us that this approach leads to promising diagnostic applications. Relevant diagnostic data are often available in health information systems. However, much of the most informative clinical data for diagnostic applications are embedded in free-text documents. In order to use these data effectively, these medical reports must undergo Natural Language Processing (NLP) to produce the needed clinical information.

In a previous evaluation of pneumonia diagnosis1, the system automatically extracted all of the structured data necessary but relied on an existing NLP sub-system to parse radiology reports. Development of this text-processing tool added substantially to the overall effort necessary to produce the diagnostic system.

This observation led us to experiment with reduced-effort approaches to NLP that can be incorporated into a system designed to automatically screen existing, computerized data for diagnostic power. The approaches described here use off-the-shelf tools to process reports and generate features. They rely on either minimal human annotation or pre-existing annotation derived from diagnostic conclusions routinely captured in the EDW (for instance, ICD codes).

Methods
We tested these two approaches to extracting relevant report data. In the first, physicians perform minimal annotation: they mark each report as indicating either “positive”, “negative”, or “possible” pneumonia. For this analysis, “positive” and “possible” were collapsed into a single category. In the second approach, no annotation is required. The report is assigned “positive” or “negative” based on the discharge diagnosis recorded in the EDW. In either case, we then apply an algorithm that 1) breaks each report into N-grams, 2) assigns TF-IDF values to each n-gram for each report, 3) applies a feature selection algorithm to reduce the number of available features, and 4) constructs and evaluates several predictive, machine-learning models to determine the operating characteristics of the report classifier.

Results
We assessed AUROC and F-Score to determine the potential usefulness of these report classifiers in a diagnostic CDS system. Two separate collections of chest imaging reports were used in these two evaluation.

Minimal Annotation: We used unigrams, bigrams, and trigrams extracted from 2198 chest imaging reports and tested in 1132 independent reports; this approach produced an AUROC of 0.932, and an F-Score of 0.925.

No Annotation: Using unigrams, bigrams, and trigrams extracted from 4938 chest imaging reports, we tested 2544 independent reports; this approach produced an AUROC of 0.933, and an F-Score of 0.917.

Conclusion
The goal of this research has been to determine whether sufficient information can be extracted from medical reports to automate the production of prototypic diagnostic systems. Familiar tools from the realm of information retrieval are used to screen the free-text information in the EDW and estimate its potential contribution. The objective is an economical, scalable way to predict the suitability of information in the EDW to contribute to a meaningful diagnostic application. The promising results seen here warrant additional effort towards refining the system for clinical use.

Using Modular Reasoning to Improve Classification Performance of the SNOMED CT Ontology With Added Description Logic Features

Robert R. Hausam, MD
Hausam Consulting LLC, Midvale, Utah

Introduction

SNOMED CT is a large, comprehensive clinical ontology based on the OWL 2 EL profile (EL++) description logic. This allows for efficient and tractable reasoning, but lacks more advanced description logic features, including universal restriction and negation. The addition of logical negation to the ontology allows creating new fully defined concepts such as “Non-bacterial infectious pneumonia” \(^1\), as well as fully defining and properly classifying existing concepts (e.g., “127302008 |Traumatic brain injury with no loss of consciousness (disorder)|”).

Currently available description logic reasoners such as FaCT++ and HermiT support negation and other advanced capabilities, but this additional capability typically comes with a significant performance cost, often making it impossible or impractical to classify a large ontology such as SNOMED CT. On the other hand, faster OWL 2 EL reasoners, such as ELK and Snorocket, perform well in classifying SNOMED CT, but lack the advanced logic capabilities.

A novel approach to providing enhanced reasoning capability while maintaining an adequate level of performance utilizes modular combinations of OWL reasoners. \(^2\) This approach uses a module extraction technique to divide the reasoning workload between a more expressive OWL 2 reasoner (e.g., FaCT++ or HermiT) and an efficient reasoner (e.g., ELK) and assigns the bulk of the reasoning workload to the latter. The MORe reasoner from the University of Oxford utilizes this technique.

Methods

An updated version of the MORe reasoner was developed as a plugin for the Protégé ontology editor tool. An OWL 2 ontology version of the SNOMED CT Jan. 2017 International Edition was generated using the distributed Perl script. This ontology was enhanced to add the fully defined concept of “Non-bacterial infectious pneumonia” (the definition includes “not (‘Causative agent’ some ‘Bacteria’)”). In addition, to achieve the expected classification results with the logical negation, it is necessary to add disjoint axioms between the “Bacteria” and “Virus” (and other) organism concept subhierarchies. Only with the disjoint axiom does the reasoner infer that “Virus” and its subtypes are not also subtypes of “Bacteria”. It is also necessary to add a universal restriction closure axiom (e.g., ‘Causative agent' only 'Virus') to the definitions of the expected subtypes of “Non-bacterial infectious pneumonia”, including “Viral pneumonia”, as only with the closure axiom does the reasoner infer that “Viral pneumonia” does not have additional unknown causative agents which may be a type of bacteria or other “non-viral” organism.

Results

Classification of the entire SNOMED CT ontology with the additional negated concept and supporting logic was performed using the expressive OWL 2 reasoner Fact++ in approx. 11.9 minutes. Classifying the same ontology using the MORe reasoner combining the ELK and Fact++ reasoners was performed in 5.2 minutes, a 56% reduction in classification time.

Conclusion

The initial results of using the modular combination reasoner for reducing classification time are promising. This likely can be enhanced using different strategies for module segmentation and different combinations of reasoners, and this will be explored in continuing work.

References

Facilitators and Barriers to PHR Adoption Experienced by Pediatric Primary Care Providers

Kisha H. Hawthorne, MHA, MBA, Ph.D. Candidate1,2, Anthony Luberti, MD2, Patricia Lee, BS2, Erin Sladek2
1Drexel University, College of Informatics and Computing, Philadelphia, PA; 2Children’s Hospital of Philadelphia, Philadelphia, PA

Introduction

The overall adoption of Personal Health Records by both physicians and patients is growing; however the pace of PHR adoption in the pediatric community is not as high. Limited studies regarding PHR adoption have been completed specifically focusing on the pediatric provider community. Facilitators and barriers impacting tethered PHR adoption are explored in this study conducted at one of the nation’s leading pediatric institutions. The current patient PHR adoption ranges from 5% to 10% for all but one of the practice (68%) in the study. The organizational average adoption rate is 30%.

Two hundred providers practicing at eleven academic primary care practices were invited to participate in the research study. Provider roles included in the study were pediatricians, nurse practitioners and medical directors. The Ambulatory Care Network leaders assisted in recruiting participants. All primary care practices were based in a large urban city center, the encompassing county and contiguous counties.

Data collection was conducted using REDCap (Research Electronic Data Capture). The survey questions were constructed based on an extensive literature review of PHRs, patient portal and technology acceptance and adoption models. The 21 question survey tool focused on provider perceived patient PHR adoption, practice setting and location, factors impacting adoption, PHR utility and functionality, health literacy, enrollment and clinical workflow, marketing and health promotion, and individual participant technology and PHR use.

Results

Fifty five survey participants were asked to select their practice location(s), setting choices urban or suburban and rate the perceived adoption rate. All participants used a sliding scale 0 to 100, to indicate what they believed the overall perceived PHR patient adoption rate, % of patients using the PHR, in their practice. The adoption rating scale was determined to be low = <34, medium = 34-67, and high = >67 (Figure 1). In 100% of provider survey participant’s opinion, their practice cares for disadvantaged patient populations. Twenty-two participants (40%) indicated that their practice was suburban and the remaining 33 participants (60%) classifying as the practice as urban.

Factors and Barriers Impacting Adoption

Study participants were asked, “Do any of the following factors(8) impact PHR utilization?” Socioeconomic levels (78%), clinical issues, i.e. non-standard PHR activation or enrollment workflow or clinical staffing inconsistencies; (57%), family structure, i.e. parents, grandparents, and guardians are surrogates for patients, (44%); ethnicity (35%); and patient age (28%) represents the top five answered identified. While patients’ privacy and access differ in the pediatric community, only 22% or 12 participants identified this as a factor impacting adoption. Regarding barriers, providers were asked to choose from 6 barriers, rating patient PHR adoption as high, medium or low. Limited access to technology (60%), digital health literacy (58%) and lack of network or wireless service (53%) were the top three barriers.

Conclusion

Pediatric providers face unique opportunities with PHR adoption including significantly improved mobile device access to the Internet, growth of consumerism and heightened desire for patient engagement. However, challenges universally expressed still included limited access to technology devices, digital health literacy, PHR promotion and enrollment. The next phase of the research will consist of pediatric primary care focus groups and comparison of actual enrollment data to validate provider perceptions.
A Visual Analysis Tool of Qualitative Eligibility Criteria of Clinical Studies

Zhe He, PhD1, Zhiwei Chen, BS1, Sarah Mixon1, Jiang Bian, PhD2

1Florida State University, Tallahassee, FL; 2University of Florida, Gainesville, FL

Introduction

Clinical studies are conducted for testing the efficacy and safety of treatments or interventions such as a new drug. Clinical trials have been widely accepted as a gold standard in modern medical research. Nevertheless, they are often criticized of lacking generalizability. For example, qualitative eligibility criteria with moderate or strict restrictions were frequently observed in cancer clinical trials, which hampers their generalizability. In previous work, we developed a rule-based parsing tool called qUalitative eLigibility cRiteriA parser (ULTRA) that leverages lexical patterns to structure qualitative eligibility criteria automatically [1]. To allow trial designers to visualize the temporal constraints of eligibility criteria, we built a Web-based visual analysis tool called VATEC (Visual Analysis Tool of Qualitative Eligibility Criteria of Clinical Studies). Using this tool, a user can select trials with certain characteristics (e.g., medical condition, study phase, intervention type), view the frequently used qualitative eligibility criteria in these trials, and visualize the temporal constraints that are often applied on a certain criterion.

Method

Our recent publication [1] describes the detailed process of using Stanford CoreNLP and lexical patterns to structure free text eligibility criteria with temporal constraints. VATEC was built with the Python Flask framework and deployed through the Amazon Web Services Elastic Beanstalk. The backend of VATEC is a MySQL relational database with eligibility criteria structured by ULTRA. VATEC has four modules: a condition selection module, a query builder module, a criterion frequency analysis module, and a visualization module. VATEC allows its users to visualize the usage pattern of a qualitative eligibility criterion in clinical studies such as temporal constraints, anchor events of the temporal constraint (e.g., “History of myocardial infarction within 6 months prior to randomization”), and other criteria that co-occur with the selected one. We used the Google Charts API to visualize the distribution of the temporal constraints of a certain eligibility criterion in a set of similar cancer studies. A typical workflow for a user is as follows: (1) select a medical condition; (2) select one or more study descriptors (e.g., study phase, intervention type) and the number of frequent qualitative features to be retrieved; (3) choose a qualitative eligibility criterion; and (4) retrieve the visual analysis results. The data about these trials are collected from ClinicalTrials.gov.

Results

The VATEC tool is publically available at https://is.gd/VATEC. To demonstrate the functionality of VATEC, let us use the following example. A user wants to examine a criterion, “malignancy”, frequently used in Phase II interventional “lung cancer” studies, and to visualize the temporal constraints applied on this criterion in these trials. As shown in Figure 1, among the 791 trials that use the criterion, 170 (21.5%) trials applied temporal constraints. Among these 170 trials, 115 (67.6%) trials used 60 months as the temporal constraint (e.g., “no malignancy within the last five years”), and 36 (21.2%) trials used 36 months as the temporal constraint. To evaluate the accuracy of the extraction of criteria and temporal constraints, we randomly selected 200 eligibility criteria sentences from cancer studies on ClinicalTrials.gov. A biochemistry undergraduate student (SM) manually extracted the criteria and the temporal constraints to form the gold-standard testing data. We then compared the performance of ULTRA with MetaMap. The results show that our tool achieved a better recall than MetaMap (95.6% vs. 74.3%). The accuracy of extracting and unifying the temporal constraints is 100%.

Conclusion

In this work, we developed a web-based tool to visualize the temporal constraints of the eligibility criteria in cancer studies. Further usability assessment is needed to test the usability of the tool in the trial design.

Reference


Figure 1. Distribution of temporal constraints applied on “malignancy” in lung cancer trials
Quality Improvement Dashboard to Monitor Oral Manifestations of Dysglycemic Patients

Harshad Hegde, M.S.1, Alok Sagar Panny, M.S.1, Rajesh Koralkar, MSHI2,
Amit Acharya, M.S., Ph.D1,2

1Institute for Oral and Systemic Health, Marshfield Clinic Research Institute, Marshfield, WI; 2Biomedical Informatics Research Center, Marshfield Clinic Research Institute, Marshfield, WI;

Abstract: Diabetes mellitus (DM), the most common chronic disease encountered in dental practice has a large impact both on systemic and oral health. The abstract provides a brief description of a dashboard which summarizes various oral manifestations in patients with dysglycemia being developed by a large health system providing medical and dental.

Introduction: The relationship between periodontal disease (PD) and DM is bi-directional. Diabetic patients are more susceptible to periodontal destruction and PD which in turn can lead to poor glycemic control. Understanding this relationship is important for dental and medical providers to aid in informing and counseling patients to seek additional medical or dental assessments as needed. To improve the oral health related quality of life among these patients there is a need to develop strong data visualization platforms to summarize various oral manifestations in patients with dysglycemia (pre-diabetes, gestational diabetes, Type I DM and Type II DM). The Dental Quality Analytics Dashboard (DQAD) graphically presents information of dysglycemic patients served by the dental and medical centers with in Marshfield Clinic Health System.

Methods: An extensive literature review was done to better understand various oral manifestations of DM, gestational diabetes, prediabetes and relationship between periodontitis and DM. A list of oral health determinants which would have an influence on the blood-glucose level of a diabetic patient and systemic factors which could impact the oral health of diabetic patients were listed out and measures and summaries were developed. A high fidelity prototype was developed to define the information hierarchy and facilitate better understanding of the interface design from a usability standpoint. Although no formal usability testing was conducted at this point, the study team did have a usability expert in formulating initial user interface design as part of this prototype development. The data for all the summaries and measures will be obtained from Marshfield Clinic Enterprise Data Warehouse and will be imported into a SQL Server database by employing the Online Analytics Processing (OLAP) cube design. ASP.NET MVC framework along with C3-JS charting library will be used to develop the front-end. The data could also be filtered based on billing type, individual dental centers, providers, and reporting periods.

The ‘Summary’ tab will display the following: a. periodontal status: provides a summary of periodontal health status of each cohort of dysglycemic patients - healthy, gingivitis and mild/moderate/severe and aggressive periodontitis, b. oral health status: provides a summary of the oral health status of each cohort of dysglycemic patients - gingivitis, periodontitis, xerostomia, dental caries, candidiasis and oral cancer, c. periodontal parameters: summarizes various periodontal parameters among different cohorts of dysglycemic patients. The count of dysglycemic patients with probing depth >=4mm in at least two probing sites, clinical attachment loss of >=4mm in 1 or more teeth, number of teeth absent >=4, bleeding on probing index and calculus index >20%, d. co-morbidities of diabetic patients: indicates the counts of dysglycemic patients with hypertension, family history of diabetes, smoking status, obesity, periodontitis and dyslipidemia, and e. dental procedures vs diabetic patients: summarizes distribution of various dental treatments received by dysglycemic patients like non-surgical periodontal treatments, complete and partial dentures, surgical tooth extractions, root canal treatments.

The ‘Measures’ tab will display the following: a. undiagnosed dysglycemic referral statistics: is a process measure which indicates the actual number of patients who followed up with a medical provider after they were screened for their glycemic index at the dental center based on the presence of certain risk factors and referred to a medical provider based on high risk. This measure will indicate the role of providers to improve the care of undiagnosed patients with risk factors of diabetes and b. change in blood glucose vs periodontal treatment: is an outcome measure which will indicate the impact of a periodontal treatment like scaling and root planning on the glycemic level of a diabetic patient.

The ‘Trends’ tab will display the monthly progression of measures and ‘Comparison’ tab would facilitate the user to compare the centers and providers with each other for a given measures.

Conclusion: This DQAD could be instrumental in providing data relevant to oral health of dysglycemic patients both at center level and provider level. This will allow dental centers to design better intervention and quality care targeting dysglycemic patients for a given oral condition and also empower dental providers with information pertaining to various oral manifestations that exist among their patient population. The next steps would be conducting usability sessions and analyzing “think aloud” thoughts and post-session usability ratings from evaluators and eliminating any design vulnerabilities and ensuring the system adaptability to address the consumer needs and preferences accordingly. Also implementing the DQAD into clinical practice by integrating it into the existing Electronic health record (EHR) or as a stand-alone system are some of the future directions.
Extracting Condition-action Statements in Medical Guidelines

Hossein Hematialam, M.S.1, Wlodek Zadrozny, Ph.D.2
1,2Univeristy of North Carolina at Charlotte, Charlotte, NC

Medical guidelines establish criteria regarding diagnosis, management and treatment in specific areas of healthcare. The most important parts of the guidelines describe actions of medical personnel to be performed under certain conditions. We propose and evaluate a supervised machine learning model extracting condition-action sentences. We establish baselines for hypertension1 and rhinosinusitis2 guidelines, as well as compare our method against previously done work on asthma guidelines.

We used part of speech (POS) tags as features in our classification models. Using POS tags instead of semantic pattern rules makes our model more domain-independent, and therefore more suitable for establishing baselines, not only for text mining of medical guidelines but also in other domains, such as text mining of business rules. We annotated part of asthma guidelines used previously by other researchers3, hypertension and rhinosinusitis guidelines to create gold standards to measure the performance of our condition-action extracting models. We classified them into three classes: condition-action, condition-consequence (effect, intention, and event) and action. ZeroR(majority classifier), Naive Bayes, J48, and random forest classifiers were applied. The results are based on 10-fold cross-validation on respective datasets. Precision for finding Condition-Action statement in asthma guidelines is 0.5 which is lower than 0.88 which was reported as Wenzina and Kaiser’s3 precision for doing the same task. The difference is due to our use of completely automated feature selection when training on an annotated corpus, and not relying on manually created extraction rules. In addition, their results demonstrate recalls on activities with specific patterns.

The results show that generally random forest classifier seems to work best in extracting Condition-Action statements. (0.84, 0.39) and (0.93, 0.43) are the (precisions, recall) pairs for extracting condition-action statements in rhinosinusitis and hypertension guidelines. Our work established baselines for automated extraction of condition-action rules from medical guidelines, but its performance is still inferior to a collection of manually created extraction rules. We are currently augmenting our model with semantic information and evaluating improvements based on “events” vs. “entities”.

References

Semantic Association for Literature Based Discovery

Sam Henry, M.S. and Bridget McInnes, Ph.D.
Virginia Commonwealth University, Richmond, VA

Introduction

This research explores the effectiveness of association measures at ranking the output terms of a literature based discovery (LBD) system. The amount of biomedical text published is growing exponentially and researchers are finding it increasingly difficult to keep up with new findings, even inside their area of expertise. LBD attempts to address this situation by automatically uncovering new, potentially meaningful relations between terms that could lead to new discoveries. A core challenge of LBD is that systems generate more potential discoveries than can be analyzed by a human, therefore effective ranking measures are essential.

Method

Using the traditional ABC co-occurrence model of LBD we begin with a start term, A, from which A implies B relationships are found via co-occurrences in text. Using the generated B terms, B implies C relationships are found. From these, therefore A implies C relationships are inferred to produce a list of output terms. This method generates hundreds or even thousands of C terms, many of which are uninformative and uninteresting, therefore ranking is critical. We evaluate the ranking procedures of: Average Minimum Weight, where the minimum value between each A to B and B to C is averaged over all B terms for each C term; Maximum B to C, the maximum B to C value over all B terms for each C term; Linking Term Count (LTC), where the count of unique B terms for each C term.

Generally, the co-occurrence frequency between terms is used as the “value” in these procedures; we modify them by replacing the co-occurrence frequency values with association measure values. Association measures quantify the likelihood of two terms occurring together in text versus by chance. Both the terms’ individual occurrence frequencies, and their mutual co-occurrence frequencies are taken into account. This study presents a comprehensive comparison between association measures, including Log Likelihood Ratio, Left Tailed Fisher Test, Pearsons Chi Squared, Dice Coefficient, Odds Ratio, Mutual Information, Jaccard Measure, and Phi Coefficient. We use data prior to year 2000 as the training set, and data published after year 2000 as the test set. We use the 2015 MetaMapped MEDLINE baseline as our dataset, and apply concept filtering to restrict B and C terms to specific UMLS semantic types.

Evaluation and Results

We evaluate the ranking methods with discovery replication and time slicing evaluation. Discovery replication reproduces previous discoveries, and the rank of the C term of interest is reported. The higher the rank, the better the system. We replicate three benchmark discoveries, Raynaud’s Disease - Fish Oil, Migraine - Magnesium, and Somatomedin C - Arginine. For time slicing evaluation, we divide the dataset into testing and training portions. The training portion is used to generate C terms, and the test portion is used to simulate to be discovered knowledge. 100 A terms are randomly chosen from the training set, and C terms are generated. Using the C terms generated, and the test set as a gold standard, precision and recall curves, Mean Average Precision, and precision at k (precision using only the top k ranked terms) are calculated. Table 1 shows the results of discovery replication. The results show the efficacy of using association measures, but more analysis is required. Further results and analysis are shown on the poster.

<table>
<thead>
<tr>
<th></th>
<th>Total</th>
<th>LTC</th>
<th>Frequency</th>
<th>Log Likelihood</th>
<th>Left Fisher</th>
<th>Chi Squared</th>
<th>Dice</th>
<th>Odds</th>
<th>Mutual Info</th>
<th>Jaccard</th>
<th>Phi</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fish Oil</td>
<td>51931</td>
<td>2344</td>
<td>2345</td>
<td>15025</td>
<td>28396</td>
<td>45759</td>
<td>4966</td>
<td>3220</td>
<td>18791</td>
<td>1487</td>
<td>3220</td>
</tr>
<tr>
<td>Magnesium</td>
<td>56243</td>
<td>192</td>
<td>192</td>
<td>31970</td>
<td>51915</td>
<td>22973</td>
<td>31970</td>
<td>2408</td>
<td>3930</td>
<td>2408</td>
<td>22973</td>
</tr>
<tr>
<td>Arginine</td>
<td>87671</td>
<td>17</td>
<td>38</td>
<td>2020</td>
<td>35133</td>
<td>29422</td>
<td>30766</td>
<td>1352</td>
<td>86523</td>
<td>70</td>
<td>393</td>
</tr>
</tbody>
</table>

Average Minimum Weight

<table>
<thead>
<tr>
<th></th>
<th>Total</th>
<th>LTC</th>
<th>Maximum B to C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fish Oil</td>
<td>51931</td>
<td>2344</td>
<td>45759</td>
</tr>
<tr>
<td>Magnesium</td>
<td>56243</td>
<td>192</td>
<td>31970</td>
</tr>
<tr>
<td>Arginine</td>
<td>87671</td>
<td>17</td>
<td>35133</td>
</tr>
</tbody>
</table>

Table 1: Discovery Replication Results for three discoveries and ranking methods.
An assessment of the contribution of technology on documentation practices in an electronic health record (EHR): evaluation of the introduction of two computer-based tools in a clinical research environment

David Herion, M.D1, Matt Fred, M.D.2
1NIH Clinical Center, Bethesda, MD; 2WorkingDiagnosis, LLC, Newark, NJ

Abstract

Metrics associated with the use of computer-based clinical documentation tools can be used to assess the usability of an electronic health record (EHR)

Introduction

Usability of clinical documentation tools in an EHR is a timely and important issue. Electronic documentation of clinical practice introduced both significant benefits, as well of substantial burdens. Although it has produced improvements in the portability of patient records, as well as billing and, to some extent, improved communications among care teams, it has been shown in some cases to have a significant detrimental impact on clinicians’ work, including time management, productivity and frustration as measured by satisfaction. In response to these impacts, efforts have been made to improve the experience of clinical documentation.

Human assets, medical “scribes”, have been introduced in various clinical care environments in an effort to reduce the burden of EHR documentation. Data regarding the efficacy or negative consequences of this practice, its “usefulness”, is wanting, leaving a limited understanding of the impact of this approach on the clinical practice. Another approach has been the development of computer-based tools to facilitate documentation in an EHR.

One difficulty in assessing the usability of these tools is lack of comparator systems in a matched cohort of users. We therefore used an “experiment of nature” to assess the usability of the EHR for clinical documentation.

Methods

Two different third-party custom applications designed to facilitate clinical documentation were sequentially deployed in a single EHR over a five-year time period. Rates of user adoption and intensity of use were compared. The breadth of functionality used (the percent of modules or components accessed) was also measured as an indicator of interest of and demand for use of the application, another aspect of usability.

Results

There was a log-phase growth in the adoption of the use of the first tool and a stable distribution of intensity of use as measured by the distribution of sustained, partial and non-adopters of the technology. The second tool had a similar rate of adoption, but a lower percentage of overall intensity of use. The analysis of the breadth of use of functions was inconclusive.

Conclusion

The adoption and intensity of use of two similar technologies designed to facilitate clinical documentation provide a potential metric of the usability of an EHR.
The Experience of Students Building an Effective Clinical Data Warehouse for a Free, Student-Run Clinic

Matthew C. Hess, BS⁴, Steve Layfield, BS⁴, Timothy Kennell, Jr., BS⁴, and Seung Park, MD⁵

University of Alabama School of Medicine; Equal Access Birmingham⁴
Community Health Systems, Inc.; Equal Access Birmingham⁵

Objective:
Many free, student-run clinics depend on the availability of free or open-source electronic health record (EHR) software; however, many of these systems do not provide easy access to data for later use in research. For our student-run clinic, we developed an in-house data warehouse that would simultaneously streamline checking patients in the clinic and store this information for later use in research. This system runs alongside the clinic’s current EHR solution. Here, we discuss the design process of the system, its benefits to the clinic, and the survey responses regarding system usability after one year of use.

Materials and Methods:
After a two-week long, intensive informatics course, we analyzed the current method of check-in and data collection at our student-run clinic. We then designed and constructed a data warehouse implemented in a Raspberry Pi using a LEMP (Linux, nginx 1.11, MariaDB 10.0, and PHP 5.5.9) stack architecture. The user-interface of the data collection form was designed with web technologies (HTML5, CSS3, and JavaScript) and the Bootstrap 3.0 framework. The system allows basic retrieval of data in a CSV format; however, all data is accessible to those with credentials to pull from the database. The system is limited to the clinic’s secure, local area network and is physically locked up when the clinic is not in operation.

After one year of use, we designed a custom, 10-question survey for the current students that ran our student-run clinic. The questions covered user demographics, user satisfaction, and length and form design.

Results:
We have created a secure, customized data warehouse solution for our student-run clinic that efficiently collects data for purposes of later research and streamlines the check-in process for patients at the clinic. This system allows patients themselves to provide their own information, students to add any additional information or view the data for clinical use, and researchers to explore patterns in data collected. The use of a normalized data structure allows easier addition of new data elements to the structure. Additionally, use of the bootstrap framework allows patients to use tablets to easily fill out the check-in form.

Our survey received an 89% response rate with 88% of respondents indicating that they were very satisfied after 1 year of use. The system was designed to enhance research data collection without impairing clinic flow, and the majority of respondents (70.6%) found the system most useful to the check-in aspect of clinic flow. Despite only one year of collection so far, the data warehouse facilitated many research projects including a longitudinal study on demographic and social factors affecting patient retention in care as well as providing easily accessible data for grant applications for clinic funding. Survey respondents’ subjective comments also identified minor areas for improvement such as modifying specific intake form questions, enhancing patient interaction with the tablets, and further optimizing the use of the system in clinic workflow.

Discussion:
Our experience indicates that it is possible for students to construct a customized, usable system for assisting with workflow in a free, student-run clinic. Given the overall positive feedback from the survey, we anticipate the system to continue to be used for data collection, making it more valuable for research data as time passes. We will continue to incorporate user feedback to improve the system for the future.

Conclusion:
We have constructed an in-house data warehouse for our free, student-run clinic that has received strong indications of usability and functionality from a survey. While there are still improvements that we can make to the system, it has proven useful in its current state. This entire project was student-driven and student-engineered, allowing medical students a valuable experience in clinic and informatics administration.
Upstream Detection of Clinical Decision Support Malfunctions

Thu-Trang T. Hickman, MPH¹, Lewis Silletto², Adam Wright, PhD¹,³,⁴
¹Brigham and Women’s Hospital, Boston, MA; ²Brown University, Providence, RI; ³Partners HealthCare, Somerville, MA; ⁴Harvard Medical School, Boston, MA

Introduction
An alert at our hospital suggesting a thyroid-stimulating hormone (TSH) test for patients who had been on amiodarone, an antiarrhythmic medication, for more than a year, stopped firing. The alert logic checks whether a patient is on amiodarone but had not received TSH testing within the last 12 months, and then suggests ordering the test. The malfunction occurred when the internal code for amiodarone changed from 40 to 7099 and the alert logic was not updated to reflect this change; thus, the alert only fired for patients on amiodarone coded as 40 and not for patients on new prescriptions of amiodarone, coded as 7099. Automated techniques for identifying CDS malfunctions exist, and most rely on monitoring alert firing rates. While these can detect some malfunctions, there are delays in detecting more subtle anomalies resulting from code changes. In this example, patients who were on amiodarone prior to the code change still had 40 listed in their records. The alert still worked correctly for these patients, so the malfunction only created a gradual drop off in the rate of firing. Furthermore, it took a year for the anomaly to become visible at all because the alert only fired for patients who had been on the medication for a full year. Many CDS alerts rely on “upstream” data elements such as medication codes and lab results to be generated. As such, we hypothesized that CDS malfunctions which do not manifest in alert firing data could be detected more quickly and easily by searching for changes in these upstream data elements. We developed and tested anomaly detection algorithms which analyzed upstream prescription data.

Methods
We developed two automated techniques to detect changes in medication codes: the existing code technique and the new code technique. While our algorithms focused on medication code changes, they can be applied to any upstream input code changes (e.g. lab tests). The existing code technique searches for medication codes being replaced by other codes while the new code technique searches for medications being prescribed for the first time, which may be replacing existing codes. We developed these techniques using data from the Longitudinal Medical Record, the legacy outpatient electronic health record (EHR) system at Partners Healthcare System, as well as from Epic, the current EHR system. We collected data on all new prescriptions of all drugs between January 1, 2008 – May 31, 2015. Using the two techniques, we identified possible CDS malfunctions within the EHRs at PHS. To assess the efficacy of these methods, we performed case studies and reviewed a random sample of anomalies detected using the methods.

Results
We identified five actual code changes using the existing code technique. For example, a new extended-release form of carbamazepine was added to the EHR’s drug dictionary (other extended-release forms were already in the dictionary). A rule is in place which suggests monitoring carbamazepine levels. It identifies patients taking carbamazepine by looking to see if one of several specified carbamazepine codes is on the patient’s medication list. The list of codes in the rule was not updated, so the rule did not fire for patients taking the new form of the medication. This was a new malfunction discovered using upstream detection that was not discovered when analyzing alert firing data. Additionally, we tested the new code method against previously known malfunctions, such as the amiodarone example, and succeeded in identifying the malfunctions earlier than methods relying on alert firing data.

Conclusion
CDS malfunctions caused by internal code changes are difficult to detect by monitoring alert firing data. By analyzing “upstream” data, these types of malfunctions can be detected more quickly and easily. Thus, upstream detection can allow CDS malfunctions to be resolved earlier, ultimately improving patients’ safety.
Behavioral Health Provider Perspectives on Health Data Sharing
Megan Hiestand1, Adela Grando, PhD1,2, Anita Murcko, MD1, Hiral Soni1, Julia Varkey1, Caroline Horrow2, Michael Saks, PhD1, Richard Sharp, PhD2, Mary Jo Whitfield3, Mark Callesen, MD3, Christy Dye4, Darwyn Chern, MD4
1 Arizona State University, Arizona; 2 Mayo Clinic, Minnesota; 3 Jewish Family and Children’s Services, Arizona; 4 Partners in Recovery, Arizona;

Abstract
Integrating behavioral and physical health is the key to value-based care. Little is known about data sharing preferences and consent practices for individuals with behavioral health conditions. This study focuses on identifying behavioral health provider perceptions about patient data sharing practices, preferences and perceived impact on care resulting from enhanced patient control of record types during consent for data sharing.

Introduction
Most protected health information is currently shared under a broad consent model. The Office of the National Coordinator identified the need for patients to have more granular control of their information when sharing with providers and others. Regulations such as 42 CFR Part 2 that extend additional protections to certain types of health records containing sensitive information, oblige the behavioral health community to develop tools that allow granular control of data by patients. We are developing My Data Choices (MDC), funded by the National Institute of Mental Health, as an electronic consent tool to offer behavioral health patients more control over their medical records. To inform the construct of MDC, we are collaborating with two community behavioral health clinics: one focused on general mental illness and the other treating patients with serious mental illness (SMI.) This study summarizes our findings from interviews with 12 representative behavioral health providers about current consent processes at their facility and their perceptions about patient-driven granular data control.

Methods
Building on the lessons learned from a semi-structured interview script consisting of 15 questions was designed by our multidisciplinary research team. Data reflecting interviewee characteristics and detailed perceptions on the impact of mental health stigma, quality of life, information sensitivity and behavioral health patient-motivation for sharing data were covered by this instrument. Subjects were based on role, longevity with the facility and experience with the consent process. All interviews were conducted at the respective facilities. Interviews were recorded and transcribed. Two independent coders performed qualitative analysis using the free Coding Analysis Toolkit (CAT) and conflicts were resolved by a third coder.

Results
Six providers from each facility, completed the semi-structured interviews in 30 to 60 minutes with a mean time of 44 minutes. From the preliminary analysis of the interviews, we identified the following:
Differences in perceptions on patient-driven granular consent between medical and social work providers. Compared with social work providers (clinical coordinators, therapists, and case managers), medical providers (psychiatrists, nurse practitioners, etc.) expressed stronger concern about the potential impact on patient safety and care coordination that may result from allowing patient more control over access to medical records.
Differences in data types that providers wish to access. Medical providers were interested in full access to patients’ treatment plans and medication lists, while social work providers wanted access to patient’s history of abuse.
Stigma, sensitive data and information sharing in the context of culture, religion, and diagnosis. All providers observed that mental illness is not recognized as a legitimate medical condition by many Hispanic, Asian subpopulations and certain religions. HIV status, substance abuse history and legal history (past and present) were identified as sensitive information. Many participants stated that SMI patients feel “labeled” and more fearful of discrimination from outside providers, potential employers and others.
Need for improved consent processes and supporting tools. All providers reflected on the complexity of the process and agreed that patients probably do not comprehend each form they sign and provided support for the development of a better process and new tools.

Future Work
We are currently fielding a companion survey of 90 behavioral health patients and legal guardians on their data sharing opinions and preferences. The final analysis of the provider interviews described above will include a comparison to the patient responses. These analyses will inform the development of the MDC consent tool.

References

2036
A Unified Chart Review Tool Integrating OMOP CDM and Unstructured Data
Charity A. Hilton, BA¹, Jon D. Duke, MD, MS²
¹Georgia Tech Research Institute, Atlanta, GA; ²Georgia Institute of Technology, Atlanta, GA

Introduction
Chart review of medical records is a necessary part of clinical research. Validation of clinical outcomes, for example, often requires examination of both structured data (e.g., diagnosis codes) as well as unstructured data such as clinical notes and radiology reports. Creating a unified interface to facilitate rapid review of structured and unstructured data is thus an important consideration in chart review and abstraction.¹ Furthermore, to maximize interoperability, chart review tools should support standard data models where they exist.

Methods
We developed an integrated chart reviewing tool to rapidly explore patients with data in the OMOP common data model (CDM) as well as in free text documents. Our review tool uses OHDSI (Observational Health Data Sciences and Informatics) cohorts as the point of entry for chart abstraction. OHDSI cohorts provide a common structure which can be correlated to clinical phenotypes. We used a minimalistic schema to incorporate clinical text narratives, using Apache Solr, an open source text search platform. The application organizes every data point to a cohort index date, supporting the reviewers’ ability to review data temporally. We support navigation filters for rapid traversing the chart based on index date (e.g., 365 days after index) as well as limiting content based on text found in structured and unstructured sources (the latter using Solr). The tool’s design provides visual clarity for rapidly disambiguating data domains (e.g., drugs, procedures). This tool has been used effectively in our institution for expert annotation and scanning of cohorts to assess phenotypic accuracy.

Discussion
Use of the OMOP common data model provides our chart review tool with portability across multiple institutions. Similarly, by allowing a lightweight schema for ingesting documents into Solr, we are able to link unstructured data from multiple sources flexibly. Future additions include features to better support validation of clinical phenotypes, such as structured annotations and comments.

References
Modeling the Cognitive Diagnostic Process, Mapping Factors Associated with Diagnostic Errors, and Analysis of Measures to Prevent Diagnostic Errors

Teruyoshi Hishiki, MD, PhD1, Rie Ichikawa, MD2,3, Jimpei Misawa, PhD2, Yukihiro Maeda, MA2, Akiko Shibuya, RN, PHN, PhD2, Yoshiaki Kondo, MD, PhD2
1Department of Information Sciences, Toho University, Chiba, Japan; Department of Health Care Services Management/2Pediatrics, Nihon University, Tokyo, Japan

Problem Addressed
Graber et al.1 reported a taxonomy of cognitive factors associated with diagnostic errors that can be classified into 25 types in the following categories: knowledge, information retrieval, information processing, and verification. They used their taxonomy to analyze individual cases of diagnostic error. We believe that, instead of making one choice from among all of the error types, it could be useful to summarize or serialize the error types, with the aim of analyzing and preventing diagnostic errors. Here, we sought to define the temporal order of diagnostic processes.

Purpose
We sought to define a cognitive diagnostic process in order to understand the cognitive factors that may lead to diagnostic errors. We also demonstrated its applicability to the analysis or prevention of decision-making errors in the diagnostic process by identifying which cognitive processes are frequently reported in the literature and identify opportunities to prevent decision-making errors in the diagnostic process.

Method
We defined cognitive functions according to the elements of the 25 cognitive factors linked to diagnostic errors in the report by Graber et al.1 We then mapped these cognitive functions according to the presumed order over time in the diagnostic process. We next searched PubMed using the index terms “Diagnostic Errors” and “Decision Making”, separated the contents of the retrieved abstracts into individual words, and counted the frequency of each of these words. Based on the results, we created a search filter to retrieve PubMed entries on decision-making errors in the diagnostic process. We used this filter to collect entries relevant to diagnostic decision-making errors for a representative disease, porphyria. We chose porphyria because it is frequently overlooked in the diagnostic process and its clinical presentation varies between patients. In addition, we expected the data management would be simpler owing to the rarity of the disease and the relatively low number of PubMed entries. We reviewed the abstracts of the PubMed entries to retrieve descriptions of factors that may hinder its diagnosis or advice on its diagnosis. Finally, we examined which cognitive processes are represented by these descriptions.

Interim Results
We were able to summarize 18 of the 25 cognitive factors into 4 cognitive processes in the information processing stage and 5 cognitive processes in the verification stage. To develop the search filter, we selected 62 words from the abstracts of 475 entries in the initial search (“Diagnostic Errors” AND “Decision Making”) and we classified the words into 4 groups: “diagnosis”, “decision making”, “error”, and “clinical”. We joined the words within the same group using “OR” and concatenated the groups with “AND” to develop the search query. When we combined this filter with the index term porphyria, a total of 250 out of 4735 entries were extracted that included the keyword porphyria and had an abstract. The abstracts of about half of these entries included descriptions to indicate that the article described the analysis or prevention of decision-making errors in the diagnostic process. We are currently analyzing which cognitive processes are related to these descriptions.

Conclusion
Plotting the cognitive diagnostic processes over time and mapping the cognitive factors associated with diagnostic errors should facilitate the analysis of decision-making errors in the diagnostic process that have been reported in the literature. We believe that the cognitive processes defined here can be used to analyze other cases of diagnostic errors, and to develop interventions aimed at preventing these cognitive errors.

References
Data-Driven Mobile Medical Immunization Clinic for Veterans

Christopher Hollweg, MD, MPH1, Andrew Muth, MD1, Valentin Rivish, PhD2, Pamela Garcia-Filion, PhD1, Hamed Abbaszadegan, MD, MBA2
1University of Arizona College of Medicine, Phoenix, Arizona; 2Phoenix VA Healthcare System, Phoenix, Arizona

Purpose
Despite best efforts the immunization rates among veterans remains low. Methods of routine clinical reminders for providers, annual mailings, and standing nursing orders have had variable levels of success. The use of mobile medical vehicles (MMVs) has shown to have an impact on clinical care including immunization rates, but limited work has been done with regards to optimizing the location of these mobile clinics. The purpose of this informatics project is to evaluate the use of data analytics and geomapping to mobilize a medical vehicle for immunization services to locations with the greatest underutilization of services.

Methods
The project utilized a randomized prospective design to identify residential areas with the lowest immunization rates among veterans for a community immunization event. Using the health record system at a metropolitan healthcare system, veteran location was geocoded to the centroid of the residential address on record. The intervention evaluated was a data-driven geolocation approach to optimize immunization delivery (data-driven optimized) versus a convenience location (non-optimized). The target vaccinations were influenza, pneumococcal and Tdap. Residential zip code areas were rank ordered based on lowest immunization rates. From the top 10 areas, 3 zip codes were randomly selected. Two were matched with an adjacent (top 10) zip code based on similar socio-economic statistics (i.e. median age, median household income, and median home sales price), and the third was matched with itself (as it had no adjacent top 10 zip code to pair with). Within the 3 zip code pairs, one zip code received a MMV event at a data-driven optimized location versus a conveniently located MMV event at the paired zip code. The project occurred during February 2017 over a 3 week period, and the order of scheduled events was randomized for the 6 sites. Veterans were notified twice of the event, once by mail with a letter designating what specific vaccine(s) that veteran was due for, and the other was either secure messaging via patient portal or telephone call. At events, veterans were asked to fill out a form indicating how they were contacted for the event, and which method they preferred.

Results
During the project’s course, the MMV delivered 767 immunizations to a total of 461 veterans out of a target population of 4,917. Use of a data-driven approach increased the immunization rate up to 2.4-fold [1.19 (pair-1) 2.4 (pair-2)], compared to the non-optimized approach, among areas with an adjacent zip code. For the self-paired zip code, a 4.8% increase in vaccination rate was seen from the data-driven optimized event, despite a previous non-optimized event for the same zip code. The data-driven optimized locations had 4.0% more veterans attending. All p-values were < 0.05. The average distance from the veteran to the location optimized vaccination events was 0.7 miles and 7.8 miles for the non-optimized events.

Conclusion
The results of this project indicate that a data-driven approach to MMV placement has a statistically significant impact on both veteran attendance and the amount of clinical care able to be given in the form of immunizations administered. This suggests that this model for MMV deployment and mobile clinical care is more effective than a non-data-driven approach. These methods can likely be applied to other mobile care modalities that use MMVs.
Standardization of Prescribing Data in PCORnet: RxNorm Concept Unique Identifiers in Multi-Site Research

Casie E. Horgan¹, MPH, Jessica L. Sturtevant¹, MS, L. Charles Bailey, MD, PhD², Lindsey E. Petro¹, Julia A. Fearrington, MPH¹, Juliane Reynolds, MPH¹, Jeffrey S. Brown, PhD¹, Jason P. Block, MD, MPH¹

¹Harvard Medical School and Harvard Pilgrim Health Care Institute, Boston, MA; ²Perelman School of Medicine and University of Pennsylvania, Philadelphia, PA

The National Patient-Centered Clinical Research Network (PCORnet) is a large, distributed research network (DRN) facilitating multi-site observational research and clinical studies. PCORnet uses a common data model (CDM) to standardize data across sites to allow for distributed querying. PCORnet’s CDM is based largely on the well-established Food and Drug Administration’s Sentinel Initiative CDM, with modifications to include Electronic Health Record (EHR) primary data sources and new data domains including prescribing and vital measurements. Given the PCORnet CDM is still early in its development, ongoing characterization and knowledge management will be integral to the long-term success of PCORnet.

The PCORnet Antibiotics Study (ABX)—which investigates the effects of type, timing and amount of antibiotics prescribed during the first two years of life on weight outcomes later in childhood—was the first study to use the PCORnet Prescribing table. As such, the ABX study team developed, tested, and implemented a Study-Specific Data Characterization (SSDC) distributed program to assess network data heterogeneity and usability via detailed domain and variable analyses specific to the ABX study cohort and protocol. Using SSDC, the ABX team was able to identify several key Prescribing table issues requiring remediation:

1. **Concept Unique Identifier (CUI) Missingness**: RxNorm is a clinical drug nomenclature system developed by the National Library of Medicine. The system uses CUIs to provide a standard representation for medications. Missingness was identified via checks that assess population of the RxNorm_CUI CDM variable. Among the 31 responding sites, approximately 4% of CUI codes were missing. However, missingness varied widely, from 0-95% of CUIs at individual sites. Because prescriptions (Rxs) for antibiotics with missing CUIs cannot be identified by analytic programs and would subsequently lead to misclassification of exposure, CUI missingness should, ideally, be addressed through complete mapping of Rx records from the EHR to the CDM.

2. **CUI Specificity**: CUI codes have varying levels of specificity ranging from the most granular specific generic or branded drug formulations—which include ingredient, dose, and drug form—to least granular ingredient-only codes. SSDC identified certain sites mapping exclusively to ingredient-only codes. For the ABX study, sites which used ingredient codes were typically unable to differentiate between amoxicillin and amoxicillin/clavulanate—an important distinction when comparing effects of broad versus narrow spectrum antibiotics. Researchers should ensure drugs are mapped to codes with the most granular level of specificity possible.

3. **Date Completeness**: The PCORnet CDM has three date fields in the Prescribing table—Rx_Order_Date, Rx_Start_Date, and Rx_End_Date. These variables had differential completeness across sites. To account for differential availability of these critical study variables, we had to implement a hierarchy to indicate which dates should be used to identify records. The ABX study instituted a hierarchy that relies first on Rx_Order_Date and, if Order_Date is missing, on Rx_Start_Date.

4. **Unique Prescriptions**: Examination of RxNorm_CUI frequencies suggested that some sites were using multiple codes at different levels of specificity to represent the same Rx, thereby creating duplicate records. This can cause critical issues if steps aren’t taken to de-duplicate records either in CDM formation or in analyses. To address this issue the ABX study team instructed sites to select the single code that provides the greatest degree of specificity for each Rx.

Overall, ABX investigations identified several important issues that required additional time and effort for the sites and the analytic team to resolve before the PCORnet Prescribing table could be used for the study. Knowledge management and application of lessons-learned will be essential for the long-term usability of prescribing data for PCORnet research. In the future, further characterization and exploration into different types of cohorts (adult vs. child) or medications (chronic vs. acute) will be necessary to expand Prescribing table knowledge.

**References**

Data Heterogeneity Among CMS Data Assets: Challenges and Opportunities for PCOR

Lauren Hovey, MA\(^1\), Rina Dhopeshwarkar, MPH\(^1\), Prashila Dullabh, MD\(^1\)

\(^1\)NORC at the University of Chicago

Introduction
The Learning Health System requires an infrastructure that leverages heterogeneous data for secondary research in order to generate clinically relevant insights that can be translated into practice. Given their utility for research in general and patient-centered outcomes research (PCOR) in particular, we conducted a literature review of the Centers for Medicare and Medicaid Services’ (CMS) data to examine data heterogeneity, and strategies and considerations for researchers interested in leveraging them for PCOR.

Research Objective
We define data heterogeneity as the variability among data (e.g., structure, format, terminology), given different sources, modes of collection, and intended uses. CMS data are a rich resource for PCOR, given their national reach and the diversity of data contributors. This gives rise to heterogeneity, which CMS has made significant efforts to address to increase its usability and availability for research (e.g., as part of the Open Data Initiative).\(^1\,^2\) Certain challenges remain, which we explore in more detail and couple with proposed solutions.

Study Design
We conducted a review of peer-reviewed and grey literature, targeting Medicare and Medicaid claims, marketplace data, quality reporting and assessment data, and beneficiary data under CMS purview. The search returned 355 articles, which we culled to 38 via title/abstract screening and full text review. A supplementary search and review of reference sections returned an additional 13 articles.

Principal Findings
Several specific types of issues consistently challenge the conduct of secondary research with heterogeneous data: delayed or missing data, non-standard data formatting, and state-level variation, all of which present methodological/validity challenges. Delayed or missing data results from incomplete reporting or collection, delays in availability, changes in patient eligibility for Medicare/Medicaid that truncate records, or policy changes that affect how and which data is collected. Data quality issues can be mitigated through linkages between data sets that fill gaps within a single source. Non-standard formatting arises from both provider- and source-level variation, including lack or inconsistent application of standards, and underlying differences in the fields within a data source. Electronic health records are a rich source of clinical data, therefore data element and content standards achieved via more rigorous certification criteria could address a considerable amount of heterogeneity and incompleteness. State-level variation typically arises from differences in Medicaid reporting requirements across states, and the incentives for managed care organizations involved in quality control. Statistically, imputations, attribution algorithms and risk adjustment can be applied to overcome data bias and enhance research reliability and generalizability. Finally, best practices and guidance related to building sustainable PCOR data infrastructure (e.g., distributed research networks) help orient the data collection process around secondary research, through which data heterogeneity and quality issues can be collaboratively addressed.

Conclusions
Heterogeneity among CMS data is a natural consequence of the breadth and diversity of its sources. Although heterogeneity creates challenges for conducting secondary research, CMS invests considerable time and effort in cleaning and harmonizing the data made available to researchers, and the literature reflects successful use of this data for PCOR. A close study of the underlying sources of data heterogeneity as well as practical solutions through standardization and data linkages may aid PCOR researchers to more efficiently address data quality issues.

References
A Dashboard to Help Clinicians Identify Hospitalized Pediatric Patients at High Risk of Medication Errors

Jonathan D. Hron, MD

Boston Children’s Hospital, Boston, Massachusetts

Introduction

Hospitalized pediatric patients are particularly vulnerable to medication errors. Factors associated with increased risk of medication errors include patient age, health literacy, number of medications, number of high risk medications and chronic conditions. While there is no consensus on what constitutes high risk medications in pediatrics, one area of concern is the need for compounding medications that are not commercially available in liquid form. Using these concepts, an interactive electronic dashboard was created to help identify hospitalized pediatric patients at high risk for medication errors.

Methods

Using the Boston Children’s Hospital enterprise data warehouse a query was created to identify all hospitalized patients. The query included all documented medications and prescriptions from each patient’s home medication list, excluding durable medical equipment and placeholders such as “no home medications.” Also included were the chronic problem list, hospital service and hospital unit caring for the patient. This report was used to create an electronic dashboard displaying a combination of interactive tables and graphs. The EDW is updated at midnight each day such that any report generated will display data for all bedded patients on the day prior. Filters for medications, chronic problems, hospital service and unit allow the user to select specific populations of patients. A sample dashboard was reviewed on March 8, 2017. The median and range were calculated for number of home medication and chronic problems. By filtering to only patients with a medication including the word “compound,” a subset of patients receiving liquid medications not commercially available were identified.

Results

The dashboard identified 286 inpatients on 43 services across 19 hospital units. In total there were 2,560 items on the patients’ home medication lists (median 6; range 1-61). Forty seven patients (16%) had more than 15 items on their medication list. There were 1,751 chronic problems recorded (median 4; range 0-44). The hospital services with the most number of items on the home medication list per patient were the Young Adult Unit (serving young adults with cystic fibrosis and other chronic pulmonary diseases) (27.8 items per patient), Pulmonary (19.3 items per patient) and the Intermediate Care Program (an ICU step-down unit) (17.1 items per patient). There were 25 patients receiving at least one compounded medication that is not commercially available (range 1-4 compounded medications per patient).

Conclusions

An interactive electronic dashboard was used to identify hospitalized pediatric patients, with filters that allow the user to quickly identify patients who may be at high risk of medication errors. While only a baseline, the data suggest a significant portion of hospitalized patients are at risk. Utilizing dashboards such as this may help clinicians and hospital administrators direct additional resources to patients in need. Future studies will attempt to correlate medication errors and adverse drug events with specific patient populations, investigate whether use of the dashboard may reduce such events as well as what types of interventions are most effective.

References

A System for Monitoring Stability, Usage and Data Completeness in OpenMRS Electronic Health Records in Rural Rwanda

Fraser HSF, MBChB MSc1, Mugisha M PT, MPH2, Richards J, MS, MPH, PhD3, Chelliah W MSc4, Ngenzi J, BSc PhD2, Gaju E, BSc, MSc5, Umubyeyi A MD, PhD2

1University of Leeds, Leeds, UK; 2University of Rwanda, Rwanda; 3U.S. Centers of Disease Control and Prevention, Atlanta, Georgia, USA; 4Jembi Health Systems, Cape Town, South Africa, 5Rwanda Biomedical Centre, Kigali, Rwanda

Abstract

EHR systems in low income settings can be unreliable with poor usage rates and data quality. We implemented a tool in 160 clinics in Rwanda to transmit data to central system (DHIS2) to monitor and improve EHR function.

Introduction:

In Low and Middle Income Countries (LMICs) there is increasing use of EHRs for clinical care, quality improvement, program management, disease surveillance and research. OpenMRS is an open source, modular EHR first implemented in Kenya and Rwanda in 2006[1], used in over 50 LMICs. Several countries have rolled out OpenMRS to hundreds of sites and are integrating the system into a national eHealth architecture. There are important challenges in achieving and maintaining stability, usage levels and data quality in EHR implementations in LMICs, and a lack of early warning of problems. As part of a CDC funded impact and cost study of OpenMRS use for HIV care in Rwanda, we developed tools to monitor the performance of systems in near real time.

Methods:

An OpenMRS software module was developed to track (1) server up and down time, (2) daily usage based on data entered, (3) completeness of key indicator variables. The latter include total numbers and numbers in last working day of: all patients; new patients; patient encounters; CD4 counts; and HIV Viral Load results. The first version of the Server Monitoring Tool (SMT) was field tested at 20 sites in 2014. A second version was developed in 2015 with capability to send data to the Rwanda national health monitoring system based on DHIS2[2], speeding recognition of malfunctioning sites. Uploaded data was validated against local reports in EHRs.

Results:

The SMT was successfully installed in 160 sites in September 2016. Fifty six sites automatically send updates to DHIS2 if internet is functional. 49 sent at least once per month but OpenMRS typically lacks a direct internet connection and staff copy reports from the SMT into DHIS2. Data can be viewed in DHIS2 reports. Of 25 sites reporting on 28th February 2017 mean uptime was 50.1%, range 7%-100% and 6 sites had 85% or better uptime.

Discussion:

Monitoring data is being used to rapidly detect power, network and IT problems and ensure system are functional during clinic hours. Early warning of missing data (e.g. viral loads) is important to ensure good HIV care, allow monitoring of general data quality, and assist study site selection, and will be complemented by a user survey.

Conclusions:

Despite all challenges faced in implementation sites, real time monitoring (SMT) of EHRs in LMIC is an innovative technology to support usability, user satisfaction, and data quality, and enable early interventions to address problems and improve use confidence. It is a key component in our process evaluation of OpenMRS.

Acknowledgements: Funding: CDC Global HIV&TB; EU Horizon 2020 Program MSC fellowship# 661289

References


2043
Real-world Evidence for Phenotyping Chronic Diabetic Kidney Disease

Lynn S. Huang¹, Xufei Zheng², Fang-Chi Hsu¹, Jianzhao Xu³, Maggie CY Ng³, Donald W. Bowden³, Barry I Freedman⁴, Jing Su¹

¹Department of Biostatistical Sciences; ³Center for Genomics and Personalized Medicine Research; ⁴Department of Internal Medicine, Wake Forest School of Medicine, Winston-Salem, NC, USA; ²School of Computer and Information Science, Southwest University, China;

Abstract
We developed a novel algorithm to accurately and reliably identify chronic diabetic kidney disease (CKD) stages and onset time using electronic medical records. The phenotyping results demonstrated significant improvement over current diagnosis-based phenotypes and make evidence-based clinical research of the progression CKD possible.

Introduction
CKD is a major, life-threatening complication of diabetes. In 2010-2011, diabetes accounted for 44% of new kidney failure cases [1] and a third of adults with diabetes have CKD [2]. CKD is classified in severity-based stages according to the reduction in kidney function using estimated glomerular filtration rate (eGFR). Clinical research and practice require accurate and reliable identification of CKD stages and onset time, but such information is largely missing or inaccurate. Electronic medical records (EMRs) allow evidence-based CKD stage phenotyping. Previous attempts at staging CKD using EMRs have focused on CKD staging, without identifying onset time, and have relied on physicians’ written notes rather than laboratory results.

Results and Discussion
According to current clinical guidelines and practice, we developed a novel, EMR-based algorithm (Fig. 1) to accurately and reliably detect CKD stages and onset time. Renal function was estimated by eGFR according to serum creatinine level using the CKD-EPI 2009 standard [4]. CKD stages were determined by persistent evidence of eGFR reduction, in which a single below-threshold eGFR is confirmed by additional evidence in the next 1-12 months. The algorithm’s novel use of hierarchical stage definitions, where evidence of more severe CKD stages indicates the previous passage through less severe stages, allows both identification of CKD stage and onset time. This algorithm is robust against transient eGFR changes due to acute kidney diseases, trauma, medication, or clinical procedures. Nephrologists reviewed and confirmed the accuracy of the phenotyping results. We implemented the algorithm and phenotyped CKD stages for 79,434 adult type 2 diabetes patients at Wake Forest Baptist Health. Dramatic improvement was shown over the ICD-based diagnosis records. For example, among the 16,885 identified Stage 2 CKD cases, only 10.1% had the corresponding ICD documentation. On average, CKD Stage 2 diagnosis was documented 10.84±5.51 years after the detected onset time.

Conclusion
The novel EMR-based phenotyping algorithm can accurately and reliably detect CKD stages and onset times. This approach improves over current CKD staging approaches by identifying CKD stage onset time, removing the influence of temporary eGFR changes, and relying directly on laboratory results. The evidence-based phenotyping of CKD stages provides critical stage information for clinical research and patient care.

References
Semantic Analysis of a Nursing Home EHR for Sensor Data Annotation

Zhongkai Zeng, Mihail Popescu, PhD
University of Missouri, Columbia, MO

Abstract

Eldercare monitoring using non-wearable sensors is an emerging solution for improving care and reducing costs. However, attaching clinical significance to the large amount of data generated by the sensor network is challenging for the clinical personnel. Using electronic health records (EHR) in conjunction with a monitoring system enables sensor data annotation with clinical terms that could ease clinicians’ interpretation. The question is if a nursing home EHR contains enough clinical information to enable such a task? We performed a semantic analysis of a dataset that consisted of clinical records of 54 residents from Tiger Place, an aging in place facility, collected over five years. We found an average of 1.5 clinical terms/resident/day available for sensor data annotation.

Introduction

Sensor networks have emerged in the last decade as a possible solution to reducing cost and improving quality of eldercare. Tiger Place is an aging in place facility for seniors located in Columbia, Missouri. Each resident included in the study (54 as of September 2016) has a computer in his/her apartment that collects data from a wireless sensor (bed, motion and gait) network. The health data for each resident is stored in a home-grown nursing EHR (see more details at http://eldertech.missouri.edu/papers). We intend to use the extracted clinical terms to annotate sensor data using a “guilt by association” approach (widely used in annotation of biological sequences) in which similar sensor patterns are assumed to be produced by similar resident clinical conditions.

Methods

We extracted text from three “description” fields found in three tables of our HER: Visit, Incident and Medication Error (see Table 1). We then used MetaMapLite (https://metamap.nlm.nih.gov/ MetaMapLite.shtml) to extract 13 semantic type categories (see Table 1) from each description field that we assumed to be linked to clinical conditions.

Results

The number of terms obtained from each table and for each semantic type is shown in Table 1. We obtained a total of about 139,000 terms which translates in about 1.5 terms/resident/day. Most terms were found for “finding” (does turn, does shave, able, etc), “pharmacologic substance” (Levaquine, Calmoseptine, Xanax, etc) and “body part or organ” (foot, hip, intestines) categories. However, most useful for our purpose seem to be the “sign and symptoms” category: pain, back pain, weakness, etc. We note few problems: some terms does not seem to make sense (phsu: pharmaceutical preparations; fndg: in care) and the surprising lack of the “disease and syndrome” term category.

Acknowledgment

This work was funded by the National Library of Medicine grant #R01LM012221.

References

Annotation of Research Common Data Elements Using Clinical Terminologies

Vojtech Huser, Cynthia Burke, Minh-Diep Nguyen, Liz Amos
National Library of Medicine, NIH, Bethesda, MD

Introduction

In recent years, Common Data Elements (CDEs) have been developed to unify data collection in clinical research studies. Since 2015, a CDE repository maintained by the National Library of Medicine allows retrieval of CDEs across several initiatives. With wider availability of de-identified patient-level clinical trials’ data across research studies, organizing data elements using a hierarchical system can improve ability of researchers to discover relevant data to re-use. Free text and synonym-extended search is often used to discover data elements, however using hierarchical relationships available in medical terminologies (such as Systematized Nomenclature of Medicine Clinical Terms; SNOMEDCT) may provide better search features. We present our pilot work on annotation of CDEs using a framework that is inspired by the SNOMEDCT compositional grammar (see snomed.org/scg). In addition to cross-trial data retrieval, such annotation can also help during trial design stage or in discovering overlaps across various CDE initiatives.

Methods

From an informatics and statistical perceptive, we structured all CDEs by their data type, such as boolean, date, text or number. In selecting the CDEs to analyze, we only included initiatives that in addition to individual CDEs also distribute 100 or more pre-configured case report forms (e.g., “A questionnaire to assess migraines and headaches”). CDE initiatives that fulfilled these criteria were PROMIS, PhenX and CDEs of the National Institute of Neurological Diseases and Stroke (NINDS). We later ruled out PROMIS because their CRFs contained exclusively CDEs of value-list data type. NINDS forms were also ruled out due to ongoing development. For finding concepts, we used the official SNOMED International browser. To evaluate mapping quality, we classified CDE-SNOMEDCT_Expression mappings into classes of exact match, partial match and no match.

Preliminary Results and Conclusion

Using PhenX REDCap download option, we obtained input data on 564 forms (or possible form sections) that reflected various PhenX protocols (e.g., Sleep Apnea child protocol). These forms contained a total of 22,705 individual CDEs. In a pilot research mode, we selected the following data types: date (example: “Date chemotherapy completed?”), boolean (example: “Are you currently in treatment for substance abuse?”) and number (example: “How old was your child when he/she completely stopped breast-feeding or being fed breast milk?”). A team of annotators annotated a total of 228 data elements. An example of an annotation would be (simplified view of the full expression): “The date the sleep disorder started?” : 298059007 Date of onset (observable entity)| + 39898005 |Sleep disorder (disorder). We were able to employ existing SNOMEDCT attributes within the draft concept model for observable entities (e.g., 704318007 |property type (attribute)|, or 370132008 |Scale type (attribute)|. The poster will present detailed data on exact matches identified, SNOMEDCT compositional grammar observations, and draft CDE annotation guidelines formulated by our team. The project website at https://github.com/lhncbc/CDE contains interim results for those outputs. Our pilot study suggest that CDE annotation can facilitate data discovery and CDE quality assurance. For some CDEs, we further considered ways of using the annotations to relate case report form data to observational data formatted in a common data model (e.g., the Sentinel model). For example, the annotation for the data element ‘Year inserted permanent pacemaker?’ follows a pattern that can be automatically converted into an observational data query. During the study, we have made several new concept submissions to SNOMEDCT (US content request system) with many of them accepted. We also hope to contribute to existing IHTSDO initiatives around observable entity concepts and compositional grammar. Through our participation on the NIH CDE Task Force, we hope to provide initial input for a discussion on annotation guidelines that can optionally be adopted by the CDE initiatives.

References

Transformation of Disparate Terminology Standards to a Common Model

Eric Y. Hwang, MD, AMA1, S. Sarah Choi, BS, PMP1
1Deloitte, Arlington, VA

Introduction

The healthcare environment today contains countless medical terminology standards developed by disparate Standards Development Organizations (SDO), each unique and serving a specific function. To reflect the rapid evolution of medicine and technology, the content that represents these various domains is continuously updated and is released at varying frequencies and schedules. In addition, Knowledge Workers will often times create localized derivative content that is dependent on SDO content. Therefore, it is critical that this medical content is updated as soon as it is available to accurately perform quality assurance on systems consuming the content in order to support patient care.

Purpose

Storing disparate data points in a common model is vital to a terminology ecosystem1, 2. By using a common model, unrelated SDO content (e.g., SNOMED, LOINC, RxNorm) or locally created content (e.g., maps or values to support institutional requirements) represented by different data models, can share a similar construct and be stored together. Because content is continuously updated and released, it is critical for this common model to properly segregate content such that it is shielded from another content's update. This, in turn, allows systems and users to have access to the latest available content.

Approach

To simulate maintenance of SDO and local content, our solution first transformed version (V1) of disparate SDO content (SNOMED and LOINC, each released in a different format) into a common model. We then transformed the Regenstrief Institute and SNOMED International Collaboration Files (SNOMED/LOINC mapping) to simulate the injection of a map set into the same model. This native content was transformed into this common model by using code written in Java to consume individual native content source files, then parse pertinent data points to fit within the common model objects. Along with version information, each unit of content (SNOMED, LOINC, SNOMED/LOINC map) is stored together in a single database, but is also conceptually segregated from each other. We then updated V1 to the next version (V2) of SNOMED without impacting LOINC or SNOMED/LOINC Collaboration data. Subsequent queries were then executed to identify maps that were impacted due to the simulated SNOMED update to illustrate a content update. Export of each unit of content was then created to illustrate that a common query could be used despite disparate native SNOMED and LOINC formats.

Discussion

Using a common model to store and maintain SDO and local content has several technical and clinical advantages, specifically it:

- Provides a single model and single extraction mechanism to store and query against different clinical content
- Quickly generates a report of change types and newly available content for knowledge workers to assess impact1
- Allows reusable Java code in subsequent releases to transform source data into the common model as well as isolating these codes to address any source information model changes, format changes, or newly introduced information
- Uniformly exports content in a format that is defined by institutional requirements for easier downstream data consumption, regardless of the original native content format

In summary, rather than treating these released units of content as individual silos, a single common model would allow disparate content to be handled uniformly.

References

Unsupervised Analysis of Activity Patterns in Eldercare Monitoring

Omar Ibrahim, Mihail Popescu, PhD, James Keller, PhD
University of Missouri, Columbia, MO

Abstract

Eldercare monitoring using non-wearable sensors is an emerging solution for improving care and reducing costs. Abnormal sensor patterns produced by certain resident behaviors could be linked to early signs of illness. We propose an unsupervised method for detecting abnormal sensor patterns based on a sensor sequence clustering approach. A preliminary analysis of the method was conducted on data collected in an eldercare facility.

Introduction

Sensor networks have emerged in the last decade as a possible solution to reducing cost and improving quality of eldercare. Tiger Place is an aging-in-place facility for seniors located in Columbia, Missouri\(^1\). Each resident included in this study (37 as of March 2017) has a data logger in his/her apartment that collects data from a wireless sensor network. Each sensor network consists of several types of sensors mounted throughout the resident’s apartment, including motion and bed sensors. The health data for each resident is stored in a home-grown nursing EHR (see more details at http://eldertech.missouri.edu/papers).

Methods

For each day, the sensors firings were converted into sequences of discrete symbols where each symbol represents one sensor type. The activity (behavior) of a resident is represented by a sequence of sensor firings as it would be represented by his/her genome. In this study, we used 11 different bed and motion sensors. We split the daily sensor sequence in subsequences using a separation threshold of 30 seconds which would provide enough granularity to capture daily activities. Our pilot dataset consisted of 14 days of resident sensor data. To ground truth our approach, we employed a normal/abnormal labeling of each day based on clinically validated health alerts from our EHR. Days on which a health alert was generated by a fall or other health event were labeled “abnormal”, while days on which no alert was generated were labeled “normal”. To explore behavior patterns captured by our sensors, we clustered both the normal and the abnormal datasets using hierarchical clustering and used Calinski-Harabasz index to find the most probable number of clusters.

Results

In figure 1, we show the clusters obtained for normal (1.a) and abnormal (1.b) days projected in 3D using t-SNE\(^2\). First, there are more clusters (activities) on normal days (21 clusters) than on abnormal days (17). Second, the cardinality of some clusters, such as bed restlessness (figure 1.c and 1.d) and bathroom visits is greater on abnormal days, while the cardinality of living room activity is greater on normal days. Finally, after generating the time histogram for each cluster, the timing of activity on normal days seems to be more consistent than that of abnormal days. We use above observations to differentiate normal/abnormal days and produce more meaningful health alerts.

Acknowledgment

This work was funded by the National Library of Medicine grant #R01LM012221.

References

Clinical Informatics 101 Training in Family Medicine

Tochi Iroku-Malize MD MPH MBA1,2, Barbara Keber MD1,2
1Zucker School of Medicine at Hofstra/Northwell, Long Island, NY;
2New York State Academy of Family Physicians

Abstract

The creation of board certification in the field of clinical informatics by the American Board of Medical Specialties brought to light the need for medical education to offer earlier exposure to the subject matter so that learners would understand the options of pursuing further training, utilization in primary care as well as the potential for a subsequent clinical informatics career. We created a workshop to introduce the basics of clinical informatics which was presented at three family medicine residency programs, the departmental academic day and the state family medicine specialty society scientific assembly. We will describe the workshop developed.

Introduction

The family medicine department of Zucker School of Medicine at Hofstra/Northwell (formerly Hofstra Northwell School of Medicine) and Northwell Health, faced the challenge of introducing clinical informatics into the continuum of medical education for students, residents and physicians. This need was enhanced by the expansion of our health system, the acquisition of various electronic medical record systems, meaningful use attestation in the ambulatory settings, as well as the then impending ICD-10 activation.

Methodology

We first looked to current curriculum guidelines for clinical informatics within our specialty. The American Academy of Family Physicians had developed guidelines in 1996 with the most recent revision in 2014. They justified the need for clinical informatics training by stating that “The ultimate concern of the physician is patient welfare, yet the medical knowledge required of physicians is beyond the brain's physical capability… physicians must leverage information technology to help ensure safe, high quality care... acquisition, input, retrieval, analysis, and sharing of clinical and administrative data are crucial components of physician proficiency in this expanding field.”1 We then reviewed what was available from the clinical informatics specialty organization, namely American Medical Informatics Association (AMIA). The 10x10 course was taken by key faculty to get the latest information and to update baseline knowledge. Curriculum from graduate clinical informatics programs were reviewed and information from the Office of the National Coordinator for Health Information Technology (ONC) was put into lay terms for the introductory course.

The workshop was created utilizing a variety of learning methods that involved individual answers, team based learning and large group report outs. The timing for the workshop was 90 minutes with time left for questions and answers. The main topics covered were:

- Key definitions (audience response system)
- Resources and support tools for clinical decision making (clinical case based scenario)
- Basic computer systems and networks (hands on teambuilding activity)
- Data modeling and database systems (office based scenario regarding immunizations)
- Policies and procedures to ensure security and integrity (hospital based scenario)
- Application of clinical informatics (small group activity)
- New technologies (audience response system)

A total of 8 workshops were held (3 residency programs, department academic day and state scientific assembly) where we reached 80 medical students, 66 residents, and over 50 clinicians. A survey was given to those at the state session which assessed learner knowledge of clinical informatics, ability, formal and informal training, relevance in current or future career. There were 30 respondents of which 72% were male. 3% were medical students, 16% were residents, 7% junior attendings (<5yrs in practice) and the majority senior attendings. What follows is preliminary analytics of the data. With regards to self-rated ability in EMR utilization 38% felt their skills were excellent while 24% felt their abilities were poor/fair. In terms of utilizing decision support tools, 27.5% felt their skills were excellent compared to 20% who rated their abilities as poor/fair. These results were mirrored with regard to consulting online reference materials. The majority of respondents (regardless of year of practice) felt that formal and informal training in medical school and residency was poor, and that clinical informatics was very important for physician related tasks or responsibilities.

Conclusion

Training medical students, residents and physicians in basic clinical informatics is vital to the ability of physicians to navigate the current health care system. The goal of our program was not to create advanced learners of informatics, but rather to serve as a method of introducing the concepts. Future studies need to be done to determine the utility of the workshop in creating a difference with regards to practice and in the long term knowledge, skills and attitudes gained during the sessions.

References

   (retrieved March 10th, 2016)
linguistic Summarization of Sensor Data Leading to Health Events

Akshay Jain, Mihail Popescu, PhD, James Keller, PhD
University of Missouri, Columbia, MO

Abstract

Sensor networks installed in homes of elderly hold promise in helping them live independently. However, understanding the large amount of data generated by these sensors presents a unique set of challenges. One solution to this problem is to display the relevant sensor data in a linguistic format when a possible health event is detected. We present an approach for the linguistic summarization of relevant information based on Linguistic Protoform Summaries (LPS). The application of the method is illustrated with data obtained from a real eldercare setting.

Introduction

In order to help elderly population in living independently for longer periods of time, apartments at TigerPlace, an aging in place facility located in Columbia, Missouri, are equipped with motion, bed and gait sensors (see more details at http://eldertech.missouri.edu). The variety and volume of the sensor data and derived measurements produce unique visualization and interpretation challenges for clinicians. In this work, we design an eldercare monitoring system that summarizes in natural language only the part of the sensor data related to a given health event.

Methods

In a survey of 22 clinicians, we found correlations between changes in sensor measurements such as bathroom activity, bed restlessness and apartment motion and 13 clinical conditions often found in elderly. Here, we illustrate our linguistic summarization method using two Urinary Tract infection (UTI) cases. As indicated by the survey, a change in one of the activities alone may not indicate a change in overall activity pattern of a person, but if the change is present in multiple sensor feature streams, then it can be of real concern. To this end, we summarized the data comprising of multiple sensor features in natural language using Linguistic Protoform Summaries (LPS).

Results & Conclusions

We show our summarization method for two weeks of data leading to a UTI diagnosis (1,2 in Figure 1) and two previous weeks (3,4), which were considered normal based on our EHR data. The summaries that distinguished the UTI and normal days are shown in Table 1, along with their Truth Values. The Truth Values of the summaries show that, as compared to normal days, the UTI diagnosis event has more days with bathroom motion and bed restlessness being high concurrently. This can be used treated as a data summary, as well as to trigger alerts. For example, a Truth Value higher than 0.5 can be used to trigger a UTI alert for days 1 and 2 but not for days 3 and 4. Based on our clinician input, we plan to design linguistic templates for 13 chronic conditions found in elderly, as outlined in our survey.

Acknowledgment

This work was funded by the National Library of Medicine grant #R01LM012221.

References


Table 1: Truth values of the summaries for the sensor data in the four days presented in Figure 1

<table>
<thead>
<tr>
<th>Linguistic Protoform Summaries</th>
<th>Case 1</th>
<th>Case 2</th>
<th>Case 3</th>
<th>Case 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>In the past two weeks, there were few days with high nighttime bathroom motion and bed restlessness</td>
<td>0.7</td>
<td>0.7</td>
<td>0.0</td>
<td>0.2</td>
</tr>
</tbody>
</table>

Figure 1: Day 1, 2, 3, 4 (left to right). 1, 2 are days leading to a UTI diagnosis, while 3,4 are normal days.
Business Process Modeling in the Arden Syntax

Robert A. Jenders, MD, MS, FACP, FACMI
Charles Drew University & University of California, Los Angeles, CA

Abstract

Context: Arden Syntax encodes knowledge as Medical Logic Modules (MLMs) to deliver clinical decision support (CDS). Business processes are part of CDS, but Arden lacks formal business process constructs. Objective: Assess the need for such constructs by evaluating the presence of business processes in MLMs. Methods: A corpus of 325 MLMs was examined. Result: At least one business process was referenced in every MLM. Conclusion: Arden Syntax could be improved with formal business process constructs.

Introduction

Arden Syntax is an American National Standards Institute formalism supervised by Health Level Seven International for representation of procedural medical knowledge with the goal of sharing units of knowledge known as MLMs. A knowledge base consists of MLMs, and an MLM, which may contain business processes, consists of 3 required categories: Maintenance (software attributes of the MLM such as author); Library (with Purpose and Explanation slots for unstructured descriptions of what the MLM does); and Knowledge (with Data and Logic slots for explicit, structured encoding of executable knowledge). While Arden Syntax has a rich set of features for representation of clinical knowledge, it lacks explicit constructs for business process modeling. Vendors and implementation sites using Arden have constructed workarounds in order to represent business processes. Business Process Model and Notation (BPMN) and the related Business Process Execution Language (BPEL) have been suggested as models for representation of workflows. These formalisms represent entities such as events (things that happen, such as escalation and conditional), activities (things that are done, such as indivisible tasks) and workflows among them. The present work assesses the prevalence of business processes in knowledge bases encoded as MLMs in order to gauge the utility of introducing business process constructs in the Arden Syntax.

Methods

A convenience sample of MLMs pooled from 5 sources—3 university hospitals and 2 vendors—was examined. Business processes, either structured in the Knowledge category or unstructured in the Maintenance category of each MLM, that align with constructs in BPMN 2.0 (such as events, activities and workflows) were identified.

Results

A total of 325 MLMs were examined. MLMs concerned mainly with lab tests were the most common (137/325 = 42%), followed by clinical assessment or classification (67/325 = 21%) and medication (41/325 = 13%). The remainder addressed administrative and miscellaneous topics. These included 430 trigger mappings that could be modeled as events such as escalation or conditional events, but often with incompletely defined workflow (e.g., responsible clinician, escalation pathways, etc). The most common events were hospital admission and a laboratory test result, while ordering of a medication or a laboratory test was the most common task. In addition, 1027 output definition statements were identified, and these would align with message flows between activities and processes containing them. Further, the Purpose and Explanation slots in the Library section of nearly all MLMs referenced business processes in narrative format that typically were not represented explicitly in the Logic slot of the Knowledge category.

Conclusions

Business processes are a common feature of knowledge represented in a convenience sample of Arden Syntax MLMs. These processes are represented in part through explicit data mappings of events, activities and message output destinations. However, more nuanced processes are present only in narrative format because of the absence of specific formal constructs that would support their representation. Arden Syntax could benefit from inclusion of additional formal business process constructs, particularly conditional events, escalation, tasks and message flows.

References

MALTASE: a Mobile Application To improve patients’ Access to their data Sharing preference

Chao Jiang, MS, Xiaoqian Jiang, PhD, Shuang Wang, PhD, Diana Guijarro, Elizabeth Bell, MPH, Imho Jang, Gwangnho Yun, Masud Rahman, Lucila Ohno-Machado MD, PhD, Hyeoneui Kim, PhD
University of California San Diego, La Jolla, CA, USA;

Abstract
iCONCUR (informed CONsent FOR Clinical data and biosample Use for Research) is a novel web-based tiered informed consent management system (https://ctri-iconcur.ucsd.edu/). To improve accessibility to this system, we developed mobile iCONCUR application. This application allows patients to control their decisions on sharing over twenty five categories of medical and genomic data for research. Using our application, patients can create, update, and store their data sharing preferences to iCONCUR at any time.

Introduction
The goal of iCONCUR is to facilitate the process of specifying data sharing preferences in an academic medical data center for research\(^1\). To further improve accessibility and user experience with iCONCUR we developed an iOS application in addition to the existing web portal.

Methods
Our application was built on Xcode 7.3.1 with Swift 2.3. By using the latest SWIFT 3.0 framework, all the dependency packages or libraries can be easily managed by the CocoaPods, which improves the security of the application for future development and maintenance. For the convenience of integration with the web version, we are using the same database as backend for the APP. As patients might be familiar with the existing web version of the iCONCUR portal, our implementation of mobile version ensures to reproduce the user experience of the web version and makes it as user-friendly as possible on the mobile side. For example, we redesigned the layout of the questionnaire for mobile users.

There are two major efforts that we addressed in the development. One is how to dynamically interact with database and display corresponding changes; another is to deal with the limited widget functionality for the check button. We took UITableView as our major display widget, which supports dynamic content displayed and makes the application more flexible to update. Combining a third package of check button with UITableViewCell allows us customize the row of each table, which addressed main problems in our development. See (Figure 1) for a visualization of the interface. In addition to the core functionalities, we added some useful functions such as email validation, password retrieval etc. We also covered the SQL injection attack in the development of server side.

Conclusion
Using this mobile APP, patients have more options to access their data sharing preferences. Furthermore, our application is very flexible and can be easily adapted for other similar tasks such as medical questionnaire. Such extension is relatively straightforward due to the adoption of CocoaPods, which allows hassle-free maintenance and upgrade. Also our preliminary demo video of the APP is available on https://www.youtube.com/watch?v=KLRFoHk9IHY.

Acknowledgement
This work has been supported by grants NIH/NHGRI R01HG008802 and NIH/NHGRI R00HG008175.

References

2052
Teacher evaluation and its correlated factors: a case study of introductory health informatics education

Xia Jing, MD, PhD¹, Timothy Vickers, MA², Brian O. Phillips, M.Ed¹, Kathy Jefford, BS³

¹Department of Social & Public Health, College of Health Sciences and Professions, ²Center for Teaching and Learning, ³College of Osteopathic Medicine, Ohio University, Athens, OH 45701

Introduction

Teaching is one of the three pillars, along with research and service, of academic success for a faculty member. Teacher evaluation is a critical means for new faculty members to receive feedback about their teaching. In some cases, it may be the only formal evaluation of a faculty member’s teaching performance. There are a variety of factors that can contribute to a teacher’s evaluation. We selected an introductory health informatics course at a public university to analyze the correlation between the teacher evaluation results and a series of factors over the course of six semesters between 2014 and 2016.

Methods and Results

A total of 291 students were enrolled in the course over the three-year period (six semesters). During this time, there were six evaluations, which included self-rated student scores. We calculated the average and median final grades, average assignment feedback per student per assignment (word counts), survey return rates, and student attendance rates per semester. We then compared the teacher evaluation results with these factors and calculated the correlation coefficients for these factors. Our case study shows that student self-rated scores (0.82), student attendance rates (0.62), and survey return rates (0.60) have high correlation coefficients with teacher evaluation results. Figure 1 provides a summary of the main results. Students’ average grades (0.28) and median grades (0.42) have a relatively low correlation with teacher evaluation results.

![Figure 1. Results of teacher evaluation and three other aspects of teaching over six semesters](image)

Discussion

Schlenker and McKinnon¹ have demonstrated statistically significant correlations between fewer absences and more favorable teacher evaluation results. Benton and Cashin’s review shows relatively low correlations between students’ final grades and teacher evaluation results. Our case study shows the important correlation between self-rated student scores, student attendance rates, and survey return rates with teacher evaluation results in health informatics education. Attendance rate can be utilized as an early indicator of teaching effectiveness for faculty members to adjust teaching strategies early in the semester. Our results, however, should be validated across institutions and across courses, since the correlation may not be unique to this course. Our next step will be cross-course validation.

References

Identifying Factors Associated with Problem List Omissions in Electronic Medical Records (EMR), to Assess Underdiagnosis in Obesity

Akshat Kapoor, PhD, MS\textsuperscript{1}, Juhee Kim, ScD, MS\textsuperscript{1}, Xiaoming Zeng, MD, PhD\textsuperscript{1}, Andrew Anderson, ME\textsuperscript{1}

\textsuperscript{1}East Carolina University, Greenville, NC, USA

Introduction

Obesity is a continuing national epidemic and the condition can have physical, psychological as well as a social impact on one’s well-being\textsuperscript{1}. Consequently, it is critical to accurately diagnose and document obesity in an electronic medical record (EMR), so that the information can be used and shared to improve clinical decision making and health communication, and in turn, the patient’s prognosis. It is therefore worthwhile to identify the various factors that play a role in documenting obesity diagnosis and methods to improve current documentation practices.

Method

IRB approval was obtained prior to commencing any study activities. We used a retrospective cross-sectional design to analyze outpatient EMRs of patients at East Carolina University (ECU) Physicians. ECU Physicians is the medical practice of the Brody School of Medicine at ECU, spread across 26 locations in Eastern North Carolina. Obese adult patients who completed an outpatient visit at any of the ECU Physicians locations, at least once between 1/1/2016 and 12/31/2016 with a corresponding BMI measure, were identified. Obese patients were identified using the measured BMI (\(\geq 30\)) entry in the EMR, recorded at each visit. Patients with any pregnancy related diagnosis were excluded, in order to avoid confounding the results. Documentation of obesity in the problem list was identified using the corresponding ICD-10 codes (E66 and below). Patients were categorized into two groups (diagnosed or undiagnosed), on the basis of a documented diagnosis (or omission) of obesity in the EMR problem list. We compared the two groups, to determine differences in gender, age group, race and patient portal use by performing statistical analysis.

Results

A total of 10,208 unique patient records of obese patients were included for analysis, of which 4,119 (40\%) did not have any documentation of obesity in their problem list. Chi-square analysis between the diagnosed and undiagnosed groups revealed significant associations between documentation of obesity in EMR and age group (\(p<0.0001\)), gender (\(p<0.0001\)), race (\(p<0.0001\)) and patient portal use (\(p=0.0133\)). Results of binary logistic regression indicate that Female (62\%, 1.3 OR), Black patients (65\%, 1.6 OR) aged 65 and below (66\%, 3.9 OR) who used the patient portal (62\%, 1.1 OR) were more likely to have documentation of obesity in their EMR, as compared to patients who were Male, White, aged 66 and above and non-users of patient portal; correspondingly. Additionally, morbidly obese patients (BMI\(\geq 40\)) were far more likely (80\%, 3.6 OR) to be diagnosed as compared to other obese patients.

Conclusion

Meaningful Use Stage 1 mandated the recording of height, weight, blood pressure and BMI as core measures of vital signs, in the EMR at each patient visit. EMR designers and developers must consider investigating automated decision support techniques to populate and update problem lists, based on the existing, recorded BMI in the EMR, in order to prevent omissions occurring from manual entry. Further studies are recommended to investigate the causes that lead to a lower rate of obesity documentation for certain sociodemographic groups. Providers must also investigate patient engagement and patient portal use as potential means of improving the quality of existing problem lists, by allowing patients to review their own problem lists.

References


Evolution of Research Topics in MEDLINE

Andrzej Kastrin, PhD¹, Thomas C. Rindflesch, PhD², Dimitar Hristovski, PhD¹
¹University of Ljubljana, Faculty of Medicine, Institute of Biostatistics and Medical Informatics, Ljubljana, Slovenia; ²National Library of Medicine, Lister Hill National Center for Biomedical Communications, Bethesda, Maryland, United States of America

Introduction

We investigated the dynamics and evolution of MEDLINE using network analysis and co-word methodology. Here we present findings expanding on our previous work in which we initiated research on the evolution of research topics in MEDLINE¹. Our hypothesis is that the history of the biomedical domain can be summarized in a series of co-occurrences of keywords (i.e., MeSH terms) that are associated with each MEDLINE citation and that identify its important topics. We model scientific topics as evolving communities of MeSH terms over time and explore how the temporal characteristics of the MeSH network can be used to provide insight into the historical evolution of scientific thought in biomedicine. As proof of concept, we implemented a computational experiment based on over 20 million documents in MEDLINE, from 1966 through the end of 2014. To the best of our knowledge this is the first such analysis conducted on the entire MEDLINE database. In particular, this analysis is of special interest to researchers who seek to acquaint themselves with scientific topics, trends, and collaboration opportunities.

Methods

We processed all of MEDLINE up to the end of 2014, only including citations tagged with major MeSH descriptors. In the constructed network, nodes represent major MeSH descriptors, and edges between two descriptors represent co-occurrence of those descriptors in the same MEDLINE citation. To capture critical events, we converted the entire network into static subnetworks at yearly snapshots, from 1966 through 2014. These subnetworks embed the communities (groups of nodes) of MeSH terms through which their development in time can be observed. We used the Louvain community detection algorithm to identify communities of nodes in each subnetwork. After discovering communities, we computed relationships between them, in order to track their evolution over time. The Jaccard coefficient was first used to determine whether two communities match. Next, we characterized the content of each community by computing its density and centrality. Density measures both the strength of the edges that tie the cluster of MeSH terms together as well as the cluster’s capacity to develop over time. Centrality measures the degree of interaction of a cluster with other parts of the network. Finally, we created a strategic diagram for each time slot, which is a graphical representation of the structure of the particular scientific field. We can identify four types of clusters according to the quadrant in which they appear in the diagram: (i) motor themes, which are both well developed and important for a research field; (ii) specialized and peripheral themes; (iii) themes which are either emerging or disappearing; and (iv) themes which are important for a research field but are not developed.

Results and Conclusion

This study identifies the major research focuses and the current status and trends in the life sciences. It provides a description of the intellectual structure and dynamics of the entire field of biomedicine from the perspective of frequently appearing MeSH descriptors. The results of this study show that (i) using MeSH terms is plausible for tracking historical events in the biomedical domain; (ii) the evolution of MEDLINE occurs in an incremental fashion; (iii) over the years more and more diverse research disciplines are involved in the complex process of scientific evolution, and links among them become stronger; and (iv) different research areas have different dynamic evolution patterns. When compared to existing research, this work is innovative in three respects: (i) the experimental design incorporates the longitudinal framework based on dynamic communities, (ii) we provide visualization of the evolution of research topics, and (iii) we propose a simple community labelling approach based on MeSH terms.

References

Development and Implementation of Clinical Data Management Technology for Clinical Research

Shogo Kato, Ph.D.1, Yoshihiko Morikawa, Bachelor2, Seiji Mitsui, Bachelor1, Takeshi Kuriyama, Bachelor1, Takahisa Ogasawara, Bachelor1, Kazuyuki Saito, Bachelor1, and Naohisa Yahagi, Ph.D.3

1National Center for Clinical Research and Development, Tokyo, Japan; 2Tokyo Metropolitan Children’s Medical Center, Tokyo, Japan; 3The University of Tokyo, Tokyo, Japan

Introduction
Advances in information management technology create opportunities for researchers to more easily share information in clinical research. Such technology allows clinical research and patient care to become more integrated and interactive. Data sharing is necessary to improve the quality of healthcare and to accelerate the progress in clinical research from bedside to community.

As the Electric Medical Record (EMR) in Japan has been evolved from ordering system, items in EMR are not standardized, because it has not been designed to share data among multiple sources. Therefore, collecting, integrating and reconstructing the high quality data into a centralized system from many institutes is challenging.

This study aims to evaluate the importance and effectiveness of code mapping to assemble data for clinical research from different formats and multiple data sources into a centralized system.

Method
The Division of Data Science for Clinical Research at National Center for Child Health and Development, Tokyo, Japan, implemented clinical data management system (CDMS), which is a user-friendly, versatile, and secure network system for clinical data. The data was actually collected from 35 clinics and 11 hospitals geographically dispersed whole Japan. There was no such report before, at least in Japan.

The CDMS used a metadata-driven model that contained built-in knowledgebase to control medical terminology. It was well-constructed knowledgebase consisted of multi-indexed data such as physical examination, laboratory, diagnosis and treatment. The data dealt in HL-7 standard, however, the code for items from EMR such as drug, ICD and laboratory was various among each institute or vendor.

The items in the central database were mapped into the standard code from local codes, by one engineer in systematic method using a supporting software. In addition, items were mapped in manual processing by the engineer, through communicating with each institute and vendor, to improve and ensure the quality of clinical data. As a verification, the percentage of mapped items among actual registered data in each classification by the end of 2016 was evaluated.

Results
In the initial state, the average percentage of items with standard code of prescription, injection, laboratory and disease, among nine institutes, was 13.4 ± 14.1%, 2.9 ± 2.9%, 2.3 ± 1.3% and 95.8 ± 3.0%, respectively.

By the operation, the proportion improved to 94.8 ± 3.3%, 96.9 ± 2.6%, 73.9 ± 12.7% and 97.2 ± 2.8%, respectively.

In the actual running data, the percentage was 96.0 ± 3.1%, 98.6 ± 1.0%, 78.8 ± 5.5 and 99.6 ± 0.9%, respectively.

Conclusion
The mapping rate before operation was extremely low and it was dramatically improved by the operation both in systematic and manual processing. Development and implementation of clinical science technologies demand a close interaction between clinical research and clinical practice. The global standardized master code, which is independent from vendors or institute, should be developed to share information more easily.
Assessment of Clinical Decision Support Resolution: A Comparison Between a Generic and Targeted CDS Notification

Jessica Ken, MPH¹, Alina N. Bogdanov, MA¹, Joseph J. Vasey, PhD¹, Shruti Gangadhar, MS¹, Richard M. Loomis, MD¹
¹Practice Fusion, San Francisco, CA

Introduction

When implemented thoughtfully, clinical decision support (CDS) can improve provider experience and patient outcomes¹. Although providers find CDS useful², notification resolution remains low³. We define “notification resolution” as providers completing a clinical action, such as a screening or exam, recommended by the CDS notification. Our objective is to use large scale electronic health record (EHR) data to compare a CDS notification that displays based on patient demographics (generic) with one based on demographics plus diagnoses (targeted).

Methods

De-identified data from Practice Fusion’s EHR database between September 5, 2015 and May 31, 2017 was used to identify patients whose providers received a diabetic foot exam CDS notification (N = 1,103,252) and patients whose providers received a clinical depression screening CDS notification (N = 13,588,699). Diabetic foot exam CDS displays for adult patients with a type I or II diabetes diagnosis (targeted); the clinical depression screening CDS displays for patients ages 12 and up (generic). Descriptive statistics and logistic regression were conducted.

Results

Descriptive statistics show that resolution rate (number of patients with a resolution divided by total number of patients with a CDS notification) of diabetic foot exam CDS was significantly higher than that of clinical depression screening CDS (12.34%, 6.57%, p < 0.0001). In both cohorts, older patients and patients with a higher Charlson comorbidity index (CCI) had higher resolution rates. Patients in medium-sized practices (500-1,499 monthly patient visits) and in practices with higher technology use (e.g. electronic prescribing and lab ordering) also had higher resolution rates. Compared to other specialties, cardiologists (12.42%) and hospitalists (11.64%) resolved clinical depression screening CDS at a higher rate. Podiatrists (23.42%) resolved diabetic foot exam CDS at a higher rate. Controlling for practice, provider, and patient characteristics, a diabetic foot exam CDS is much more likely to be resolved (OR=1.367, CL=1.358-1.376, P<0.0001) than an alert for clinical depression screening.

Discussion

The large number of patients with a resolved CDS highlights the potential impact of CDS programs. Understanding how to best target a specific CDS is essential to patient care and reducing provider alert fatigue.

Conclusion

Our study suggests that providers are more likely to resolve alerts attributed to a narrower patient population or specific condition. Further, specialists related to the alert type are also more likely to resolve the alert.

References

Challenges Impacting Data Collection from Electronic Health Record (EHR)
Systems in Small and Medium Sized Practices in the Midwest

Pauline Kenly1, Luke V. Rasmussen1, Andy Schriever2, Faraz Shafique Ahmad1, Kathryn L. Jackson1, Richard Chagnon1, Theresa L. Walunas1, Abel N. Kho1
1Northwestern University Feinberg School of Medicine, Chicago, IL; 2Park Street Solutions, Naperville, IL

Introduction: Since its enactment, the Health Information Technology for Economic and Clinical Health (HITECH) Act has led to a widespread adoption of Electronic Health Records (EHR) systems in clinical practice. Accessibility and utilization of the wealth of information contained in EHR systems are becoming important in guiding patient care and future health research. The AHRQ-funded Healthy Hearts in the Heartland (H3) research program's goal is to work with small and medium sized practices (practices with 20 or fewer primary care providers) in the Midwest with the intention of implementing quality improvement (QI) strategies for cardiovascular care utilizing EHR data. The study focuses on evaluating four clinical measures: A spirin when Appropriate (NQF#0068), Blood Pressure Control (NQF#0018), C holesterol Management and Smoking Cessation (NQF#0028). The open-source popHealth software (https://github.com/OSERHA/popHealth) was implemented to calculate the measures and provide a central dashboard across practices for feedback and audit.

Methods: We assessed each EHR system capability to output clinical data in standard formats: Consolidated Clinical Document Architecture (C-CDA), Clinical Document Architecture (CDA) or Continuity of Care Documents (CCD) with an option to export records in batch. We also assessed if the EHR allowed custom report development or direct database export. We relied on vendor documentation to determine how patient data was structured.

Results: The study enrolled 226 small and medium practices with 21 EHR systems represented. Of the 21 EHRs, 7 were assessed for data export capability. Table 1 shows the distribution of data export capabilities for assessed EHRs. Four EHR systems allowed for data export formats in CCD, C-CDA and custom reports. The three output formats had limitations in providing complete data for measure calculations. For example, data elements such as smoking history and RxNorm codes for medications were missing in some of C-CDA data exports. Three EHR systems enabled direct database connection, allowing us to write data extraction queries against patient level data. This method provided more complete data. Of the 14 EHR systems that were not assessed, we attempted to contact vendors to identify data export options available, with insufficient responses.

Table 1: Export Capabilities from 7 Assessed EHRs

<table>
<thead>
<tr>
<th>EHR Data export capability</th>
<th>EHRs with capability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard Report (CCD, C-CDA)</td>
<td>3</td>
</tr>
<tr>
<td>Custom Report</td>
<td>1</td>
</tr>
<tr>
<td>Direct Database Queries</td>
<td>3</td>
</tr>
</tbody>
</table>

Discussion: Quality EHR data not only promotes health research, it also provides reliable data for submission for quality improvement programs such as Meaningful Use. EHR data also guides patient care through coordination of care and referral documents between providers. The challenges of exporting EHR data evident in the H3 research study underscores the necessity for continued efforts towards advocating for better quality EHR data exports, formats and standards. It also highlights the need for easier access to EHR systems and vendor documentation. One such promising initiative by the Office of the National Coordinator for Health Information technology, would be to have EHR vendors enable open Application Programing Interfaces (API).

Acknowledgements: This research was supported by AHRQ grant number 5R18HS023921-02.

References
Leveraging Value Sets from the Value Set Authority Center (VSAC) in a Standards-Based Clinical Data Repository

Richard Kiefer¹, Luke V. Rasmussen², Jennifer A. Pacheco², Peter Speltz³, Joshua C. Denny³, William K. Thompson², Jyotishman Pathak⁴, Guoqian Jiang¹
¹ Mayo Clinic College of Medicine, Rochester, MN; ² Northwestern University, Chicago, IL; ³ Vanderbilt University, Nashville, TN; ⁴ Cornell University, New York City, NY

Introduction
The National Library of Medicine (NLM) Value Set Authority Center (VSAC) stores value sets used in Clinical Quality Measures. The Observational Health Data Sciences and Informatics (OHDSI) program offers a wide range of tools for large-scale integration and analytics of observational health data. The objective of this study is to develop an application to import VSAC-hosted value set definitions into an OHDSI repository. This will facilitate the reuse of well-defined value sets from existing value set repositories.

Value sets are groups of concept codes from one or more vocabularies, which are used to define clinical concepts. These value sets are created by subject matter experts and stored in public repositories, such as VSAC. In our ongoing Phenotype Execution and Modeling Architecture (PhEMA)¹ project, we utilize Common Terminology Services 2 (CTS2)² services, an HL7 standards-based approach that allows an application to retrieve value sets from any repository which is CTS2 compliant. OHDSI³ is a collaborative effort to create open source methods and tools to analyze observational health data. To accomplish this goal, OHDSI requires the use of a Common Data Model (CDM)-based repository. A collection of standard vocabulary concepts is stored as a concept set within the CDM Concept table, but reuse of existing value sets requires manual transcription into an OHDSI instance. By leveraging application programming interfaces (APIs) across CTS2, VSAC and OHDSI, we are able to bridge these repositories to facilitate value set reuse.

Methods and Results
A command-line utility was developed using Java to make CTS2 representational state transfer (REST) calls to access value set definitions, and parse the JavaScript Object Notation into the individual concepts contained in the value set. We used a CTS2 wrapper implemented by the PhEMA project to retrieve the value set definitions from VSAC.

We used OHDSI Broadsea to deploy the full OHDSI technology stack, which includes the OHDSI CDM and a web API to provide REST services. By sending the concept name and namespace (vocabulary identifier) to the vocabulary search service we were able to get the representation of the concept from the OHDSI repository. Mappings were made to address slight differences in the code system names. For example, where VSAC stores SNOMEDCT and RXNORM; OHDSI stores SNOMED and RxNorm respectively. Using the concept set service, we first defined the name for a concept set container. Once the container has been created and an identification number has been established, a second web service call is made to populate the concept set with the list of concepts received from the vocabulary search.

The utility was verified by migrating 40 value sets into our OHDSI instance. Two members of the study team manually verified that the concepts matched within the value sets across both systems.

Discussion
Our utility to retrieve a VSAC value set based on its OID and migrate it to an OHDSI repository for use in OHDSI tools will allow users of OHDSI to easily adopt established value set definitions from any CTS2-compliant repository. In future work, we plan to use the value set migration as part of a larger implementation to allow algorithms created by the PhEMA authoring tool to be run against an OHDSI installation to retrieve patient cohorts.

References
1. PhEMA project: http://informatics.mayo.edu/phema/index.php/Main_Page
2. CTS2 specification: http://www.omg.org/spec/CTS2/1.0/
3. OHDSI program: https://www.ohdsi.org/
Antimicrobial Susceptibility Data Visualization Tool

Eugene Kim, MD\textsuperscript{1,2}, Larry A. Nathanson, MD\textsuperscript{1,2}, Evan Leventhal, MD PhD\textsuperscript{1,2}, Steven Horng, MD MMSc\textsuperscript{1,2}

\textsuperscript{1}Division of Clinical Informatics, Beth Israel Deaconess Medical Center, Boston, MA
\textsuperscript{2}Harvard Medical School, Boston, MA

Introduction
Increasing levels of antibiotic resistance remain a significant source of morbidity and mortality while the development of new antibiotics lags\textsuperscript{1}. The appropriate use of antimicrobial medications is critical and depends on multiple factors, such as expert recommendations, antibiograms (i.e. summaries of antibiotic resistance patterns within a population), and the patient’s clinical condition. Culture results are numerous, and patients with complex medical histories may have a wealth of previous culture data available to further tailor drug selection. There were 3,475 unique isolates grown from cultures with more than 1,464 resistant strains found at our institution from 7/1/2016 to 12/31/2016. Clinicians may find it difficult to assimilate this data into their decision-making as electronic health records (EHRs) typically spread this information over multiple pages, requiring several clicks to access each result, and do not facilitate comparison of results or analysis for trends. Our goal is to assist antibiotic selection by enabling easy comprehension of the entirety of a patient's microbiological history through an intuitive visualization tool.

Methods
Patient microbiology data were retrieved using the Caché development environment [InterSystems Corp., www.intersystems.com], the system with which the EHR at Beth Israel Deaconess Medical Center was written. Web programming tools, JavaScript, HTML, CSS, and jQuery [www.jquery.com], were used by the authors to create a graphical representation of culture results that was integrated into an EHR display. An overview of the culture sensitivity data across each antibiotic was organized in a table by test date, with the sensitivity results (i.e. 'sensitive', 'intermediate', 'resistant') color-coded. Color-coding was also used to reflect the highest level of resistance encountered for each individual test and antibiotic. Clinicians were able to click to "drill down" to reveal further information about the culture as needed.

Conclusion
We developed a novel visualization to summarize large amounts of culture data in a clinically meaningful and intuitive manner. Such overviews of data can be especially helpful in an environment such as the emergency department, where clinicians are often task-saturated and decisions need to be made quickly. Further study is planned to assess the usability and utility of the tool. Future enhancements could also assist clinicians with antibiotic selection using clinical decision support tools and by incorporating additional patient data from the EHR.

References

Abstract
Increasing levels of antibiotic resistance remain a significant source of morbidity and mortality. Use of past culture data can result in better choices of antibiotic agents, however this information is often cumbersome to access within an EHR. We developed a microbiology data visualization tool to quickly discern susceptibility patterns for individual patients and facilitate optimal antibiotic selection.
STD MIG-PGx Implementation in Clinical Trials Using Legacy PGx DB

Eun-Young Kim, MD, PhD1, Su-Yeon Park, BS2, Na-Yul Kim, MS2
Eun-Young Cha, MS2, Jae-Gook Shin, MD, PhD1,2
1Department of Clinical Pharmacology, 2Department of Pharmacology and PharmacoGenomics Research Center, Inje University College of Medicine, Busan, Korea

It has been increasing the need of pharmacogenomics data in clinical trials, which led to introduce Study Data Tabulation Model Implementation Guide for Pharmacogenomics / Genetics (SDTMIG-PGx) as a data standard. In this study, we developed an auto-converter that converts data in legacy PGx DB to standardized data format according to the guideline. This converter may be contributed to improve quality of clinical trials and to reduce time and cost of clinical trials.

Introduction
Data standardization is very important in clinical trials especially as electronic data capture technologies has been developed. A standards developing organization, Clinical Data Interchange Standards Consortium (CDISC, https://www.cdisc.org), issued Study Data Tabulation Model (SDTM) for clinical trial data standardization. The Unite States Food and Drug Administration (FDA) has asked clinical trial data submission as a SDTM format in product application process. Recently SDTM Implementation Guide for Pharmacogenomics/Genetics (SDTMIG-PGx) was introduced to guide the organization, structure, and format of gene-related data to submit into regulatory authority such as FDA (2015). Therefore we developed STD MIG-PGx implementation tool using our legacy PGx DB (http://pgrc.inje.ac.kr/) which is not compatible to SDTMIG-PGx guide.

Materials and Methods
The data standardization process is demonstrated on Figure 1. The legacy DB includes subject, sample, genetic test results, and gene information. The comparison and mapping specification between STD MIG-PGx domains and our legacy data variables were performed. The auto-converter was developed based on the work using PHP (http://php.net) and MySQL (http://mysql.com) for API. We tested it using previous clinical trial data with PGx test results.

Results
The mapping specification after comparison between data variables in our legacy PGx DB and SDTMIG-PGx domains was successfully completed. The auto-converter based on the above work was developed and it applied to generate the standard format of PGx datasets using genotype results of clinical trial subjects. It included CYP2C9 and CYP2D6 gene-related data (Figure 2).

Figure 2. The auto-converter of PGx data and the extracted standardized datasets.

Conclusion
In this study, we developed an auto-converter for SDTMIG-PGx implementation using our legacy PGx DB. It would be a useful tool for improving clinical trial data quality and strengthening competitiveness in clinical trial industry.
Guiding Principles for the Duke Connected Care Predictive Modeling Pilot

Eugenie Komives, MD\(^1\), Shelley A. Rusincovitch\(^1\), John Paat, MD\(^1\), Lawrence Carin, PhD\(^2\), Daniel Costello, MPA\(^1\), Michael Gao\(^2\), Bradley G. Hammill, DrPH\(^1\), Ricardo Henao, PhD\(^2\), N. Benjamin Neely, MS\(^1\), Ursula Rogers\(^1\), Devdutta Sangvai, MD, MBA\(^1\), Mary Schilder, MA\(^1\), Erich S. Huang, MD, PhD\(^1\)

\(^1\)Duke Health, Durham, NC; \(^2\)Duke University, Durham, NC

Introduction
Predictive modeling in healthcare is a complex undertaking. It benefits from cross-disciplinary perspectives that reflect the multi-faceted nature of risk and potential interventions. However, it can be challenging to foster these multiple viewpoints to recognize common goals and shared purposes.

Methods
In the first stages of developing a clinically-useful machine learning-based predictive model for hospital admissions, Duke Connected Care, an accountable care organization (ACO) managing a Managed Shared Savings population, assembled a diverse stakeholder group to develop a set of guiding principles that were finalized on December 21, 2016. The stakeholders represent a cross-section of expertise, including clinical care, case management, health administration, data science, information technology, translational science, and informatics. We based this activity on the approach used by the FDA Sentinel Initiative and the National Patient-Centered Clinical Research Network (PCORnet). Our intention was to form a strategic roadmap for the pilot; allow underlying assumptions, intent, and values to be recognized; create a basis for priorities to be balanced and decisions evaluated; and enable every team member to understand the underlying context.

<table>
<thead>
<tr>
<th>Table 1.</th>
<th>Stakeholder Group Guiding Principles for the MSSP Predictive Modeling Pilot.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td><strong>Practicality is key.</strong> Elaborate and overly complex risk prediction models may not provide strong value in actual use.</td>
</tr>
<tr>
<td>2.</td>
<td>Above all, our predictive models must generate <strong>information that is actionable</strong> and closely tied to intervention.</td>
</tr>
<tr>
<td>3.</td>
<td>The advisory role of <strong>clinicians and care managers</strong> is integral in designing and improving the predictive workflow, especially within the context of the clinical care setting.</td>
</tr>
<tr>
<td>4.</td>
<td>Careful incorporation of predictive models into <strong>clinical workflows</strong> is crucial for adoption and use.</td>
</tr>
<tr>
<td>5.</td>
<td><strong>Clear and transparent documentation</strong> is a priority, especially to capture decision points along the way.</td>
</tr>
<tr>
<td>6.</td>
<td>The <strong>“why” is important</strong> in making decisions about actionability and to understand the basis of risk. “Black-box” risk prediction, where the underlying design is not transparent, is less helpful. Understanding what drives the level of risk – particularly increasing risk – is critical to care managers and providers understanding what they might be able to do to mitigate that risk.</td>
</tr>
<tr>
<td>7.</td>
<td>Any given model <strong>may involve trade-offs</strong> (for example, more timely/less accurate vs. less timely/more accurate). Therefore, we will be open to simultaneous development of multiple models to provide multiple options where different models may be optimized for timing of deployment within clinical workflows.</td>
</tr>
<tr>
<td>8.</td>
<td>We will seek to <strong>learn from the work that already exists</strong> in predictive modeling, including published and proprietary models, but be open-minded to the new possibilities in this pilot. We will also consider leveraging and/or modifying existing tools where they exist and where permissible, especially where efficiencies can be gained.</td>
</tr>
<tr>
<td>9.</td>
<td>The <strong>lessons learned through model development</strong> may be as important as the models themselves, including new paradigms of engagement and collaboration.</td>
</tr>
<tr>
<td>10.</td>
<td><strong>Data privacy and security</strong> are critically important.</td>
</tr>
<tr>
<td>11.</td>
<td>We need to understand measures of <strong>value, quality, and cost</strong> that will allow us to monitor performance and gauge success of this predictive modeling process. Tracking and observing <strong>performance of risk prediction over time</strong> is important. It is also important to monitor the <strong>performance of interventions</strong> associated with the risk predictions.</td>
</tr>
</tbody>
</table>

Conclusion
The development of guiding principles has been a valuable foundation for our work in predictive modeling for the Medicare Shared Savings Program (MSSP) population served by Duke Connected Care. Although these specific principles reflect the uniqueness of the pilot and collaboration of the stakeholders, we anticipate that others can benefit from the value of developing guiding principles as a generalizable activity and phase of development.
Machine Learning Models to Predict Readmission for Patients with Cirrhosis
Jejo D. Koola1,3, Aize Cao2, Guanhua Chen2, Amy M. Perkins2, Samuel B. Ho3,4, Sharon E. Davis2, Michael E. Matheny1,2
1Tennessee Valley Healthcare System VA, Nashville, TN; 2Vanderbilt University, Nashville, TN; 3University of California at San Diego Medical Center; 4Dept. of Veterans Affairs, San Diego, CA

BACKGROUND: Patients with cirrhosis, a late stage of chronic liver disease, are at increased risk of hospitalization and hospital readmission. Although several studies have looked at models for predicting readmission for patients with cirrhosis, they are limited by small sample sizes, limited candidate predictor variables, and limited evaluation of discrimination and calibration.

METHODS: We analyzed a retrospective cohort of patients hospitalized nationwide in the Department of Veterans Affairs between 2005 and 2013. We included patients (age ≥ 18) with a diagnosis of cirrhosis hospitalized for any cause, and excluded hospitalizations where the patient died during the index hospitalization, was discharged against medical advice, or was transferred to another non-VA hospital. We identified patients readmitted to any VA facility within 30 days for any cause. We included 254 candidate predictor variables consisting of demographics, laboratory values, comorbidities, healthcare utilization, and medications. We evaluated four models: penalized logistic regression using Least Absolute Shrinkage and Selection Operator (LASSO), gradient boosting, random forest, and weighted random forest. Each model was run on 10 bootstrap resampled datasets. We evaluated discrimination by the area under the receiver operating characteristic curve (AUC) and the calibration by the estimated calibration index (ECI) and the observed-to-expected ratio (O/E). In addition, we ran the models on the full dataset to estimate odds ratios and variable importance.

RESULTS: There were 193,674 index hospital admissions comprising 72,737 patients. Of those, there were 43,674 readmissions within 30 days, for a readmission rate of 22.6%. On average patients were 61 +/- 9 years old, 73.6% Caucasian, and with an admission MELD score of 12 +/- 5. Etiologies of cirrhosis were 28.3% alcoholic, 18.3% viral, 33.9% alcoholic and viral, and 13.4% non-alcoholic fatty liver disease. Gradient boosting outperformed the other models in terms of discrimination and calibration (Table 1, AUC: 0.705 ± 0.003, ECI: 0.007 ± 0.002). Table 2 lists the ten variables with highest and lowest odds ratios for readmission.

DISCUSSION: Hospitals need tools to identify cirrhotic patients at high risk for readmission in order to efficiently target interventions. These data indicate that patients with ascites are at higher risk for readmission, particularly if they require albumin infusions. Prior high readmission rates and psychiatric/substance use diagnoses also were significant risk factors. In contrast, use of hospice and nursing home services were significantly associated with a reduced risk for readmission. Gradient boosting outperformed other models in terms of discrimination and significantly based on calibration. These data are useful for further efforts to develop and implement population and case management strategies to help reduce early readmission in patients with cirrhosis.

Table 1: Performance characteristics.

<table>
<thead>
<tr>
<th>Model</th>
<th>AUC</th>
<th>ECI</th>
<th>O/E</th>
</tr>
</thead>
<tbody>
<tr>
<td>LASSO</td>
<td>0.678 ± 0.003</td>
<td>6.53 ± 0.090</td>
<td>0.47 ± 0.002</td>
</tr>
<tr>
<td>GB</td>
<td>0.705 ± 0.003</td>
<td>0.007 ± 0.002</td>
<td>1.00 ± 0.003</td>
</tr>
<tr>
<td>RF</td>
<td>0.643 ± 0.024</td>
<td>42.5 ± 18.5</td>
<td>0.273 ± 0.057</td>
</tr>
<tr>
<td>WRF</td>
<td>0.634 ± 0.036</td>
<td>43.1 ± 18.0</td>
<td>0.271 ± 0.057</td>
</tr>
</tbody>
</table>

Values are mean ± SD. RF: Random Forest; WRF: Weighted Random Forest; GB: Gradient Boosting

Table 2: Variables with ten highest and ten lowest odds ratios for readmission based on LASSO.

<table>
<thead>
<tr>
<th>Variable Name</th>
<th>Odds Ratio</th>
<th>Variable Name</th>
<th>Odds Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Home Albumin Infusion</td>
<td>1.44</td>
<td>Inpatient Norepinephrine</td>
<td>0.38</td>
</tr>
<tr>
<td>Admission Within Prior 30d</td>
<td>1.42</td>
<td>Inpatient Vasopressin</td>
<td>0.43</td>
</tr>
<tr>
<td>History of Paracentesis</td>
<td>1.18</td>
<td>Discharge to Hospice</td>
<td>0.61</td>
</tr>
<tr>
<td>Inpatient Albumin Infusion</td>
<td>1.17</td>
<td>Discharge Info. Missing</td>
<td>0.67</td>
</tr>
<tr>
<td>Total #Inpatient Visits Past Year</td>
<td>1.12</td>
<td>Discharge to Nursing Home</td>
<td>0.81</td>
</tr>
<tr>
<td>History of Psychiatric Illness</td>
<td>1.09</td>
<td>Discharge from Spinal Cord Injury</td>
<td>0.84</td>
</tr>
<tr>
<td>Home Glucocorticoids</td>
<td>1.08</td>
<td>Discharge from Surgery</td>
<td>0.88</td>
</tr>
<tr>
<td>History of Drug Abuse</td>
<td>1.08</td>
<td>Admission for any Surgical Reason</td>
<td>0.88</td>
</tr>
<tr>
<td>Inpatient Benzodiazepines</td>
<td>1.07</td>
<td>History of HCC</td>
<td>0.89</td>
</tr>
<tr>
<td>Home Lactulose</td>
<td>1.07</td>
<td>Inpatient Cephalexin (Gen. 1)</td>
<td>0.90</td>
</tr>
</tbody>
</table>
Combining Social and Behavioral Determinants of Health Questionnaires to Inform Precision Medicine in a Community Health Center Clinic
Taneya Y. Koonce, MSLS, MPH\(^1\), Sheila V. Kusnoor, PhD\(^1\), Suzanne T. Hurley, RN, APRN-BC\(^2\), Kalonji M. McClellan, MA\(^1\), Nunzia B. Giuse, MD, FACMI\(^1,3\)

\(^1\)Center for Knowledge Management, Vanderbilt University Medical Center, Nashville, TN; \(^2\)University Community Health Services, Nashville, TN; \(^3\)Departments of Biomedical Informatics and Medicine, Vanderbilt University Medical Center, Nashville, TN

**Purpose:** Precision medicine is becoming more pervasive and points to a more holistic approach to individualized patient care. We are conducting a series of detailed studies to understand how to best administer a questionnaire to assess social determinants of health prior to their routine incorporation in electronic health records (EHRs). Using an online sample, we previously evaluated a set of measures recommended by the National Academy of Medicine (NAM) for assessing health-related social and behavioral needs\(^1,2\). Here, we explore in-person administration of a hybrid instrument to collect social and behavioral determinants of health in a community clinic setting, to gain further insight into data collection methods and to guide standards for EHR implementation.

**Methods:** For this pilot study, a convenience sample of patients ages 18 and older was recruited at the Vine Hill Community Clinic, a health center providing medical care for Nashville’s underserved population. Patients were asked to complete a social and behavioral determinants of health questionnaire developed specifically for the community clinic population, called the Protocol for Responding to and Assessing Patient Assets, Risks, and Experiences (PRAPARE)\(^3\) in combination with a subset of items from the NAM questionnaire which were not addressed in PRAPARE, and the Federal Reserve Board’s Survey of Household Economics and Decisionmaking (SHED)\(^4\).

**Results:** One hundred participants completed the study. The median age was 38 (interquartile range = 26.5-56.5; n=100), and the gender distribution was 67% female, 30% male, and 3% transgender (n=100). The median questionnaire completion time was 11 minutes. The questions that were most commonly unanswered included those about household income and race, which were not answered by 40% and 7% of participants, respectively. Results from descriptive analyses indicated that the most prevalent social and behavioral needs were in the domains of physical activity, financial and material security, education, and social integration and support.

**Conclusion:** This study supports the feasibility of administering a hybrid social determinants of health instrument in the community clinic setting and provides insight on what questions are most likely to be answered. Our data suggest that inferences can be made based on patterns of unanswered questions, and data interdependencies may be used to overcome voids in data collection. The findings regarding unanswered questions also suggest an opportunity for future research on better approaches for assessment. Use of a hybrid questionnaire, incorporating elements of PRAPARE, NAM, and SHED, allows providers to obtain a more complete assessment of patients’ social and behavioral health factors, which is a necessary component for truly achieving precision medicine. The results from this study will help to establish standards for routinely collecting information about social and behavioral determinants of health via electronic health record systems. Future efforts are planned to study how to best connect patients with appropriate public health and community-based resources and evaluate their effectiveness in addressing identified social and behavioral needs.

**References**
Combined approach to Lyme disease surveillance using text mining and structured data elements available in Electronic Health Record

Rajesh Koralkar, MSHI¹, Elham Sagheh Hossein Pour, MS¹, Jennifer Meece, PhD¹, Anna Schotthoefer, PhD¹

¹Marshfield Clinic Research Foundation, Marshfield, WI;

Introduction:
Lyme disease and other tick-borne diseases have been on the rise in Wisconsin (1). Early diagnosis and treatment of Lyme disease patients are currently the most effective ways of preventing severe illness and development of secondary sequelae, such as Lyme arthritis and neurologic abnormalities (3). However, diagnosis of Lyme disease is made difficult by the non-specific symptoms found in early stages of the disease and the absence of a confirmatory laboratory test. Collecting data on these patients, in the form of a Lyme disease registry, is needed to better understand the disease. One of the difficulties in developing such a registry is identifying patients most likely to have the disease. To achieve this goal we developed an electronic surveillance tool which incorporates regular expression (Regex) text mining and structured data elements available in electronic health records (EHR) to identify Lyme disease cases.

Methods: This study was approved by the Marshfield Clinic Institutional Review Board. The overall aim of this work is to develop a Lyme disease registry. Patients who visited the Marshfield Clinic Health System from January 2014–December 2016 were included in the study. We screened the electronic health record clinical narratives for terms and structured data elements listed in Table 1. We used a stepwise approach to evaluate the accuracy of adding additional elements to the screening tool. We manually validated 10% of the patient data from each group and calculated the algorithm’s predictive positive value for detecting patients with Lyme disease.

Results: We identified 2086 patients who had positive mention of Regex terms; of these, 691 patients were tested for Lyme serology, 463 had ICD code in EHR and 1279 received Lyme specific treatment. Based on these findings, patients were divided into 4 groups as shown in Table 2. Group A had .35 PPV, group B had .65 PPV, and group C and D achieved .95 PPV.

Discussion: Using multiple data sources, we were able to develop a highly predictive EHR-based phenotyping algorithm to identify patients with Lyme disease. Our results are similar to the experience of others in that the more phenotyping approaches and data sources used tended to increase PPV [3]. Regex alone achieved 35% accuracy as we did not consider negation and location of text in patient history section. Adding laboratory test data improved the PPV and finally combining Regex with treatment and ICD diagnosis code achieved the best results. Regex text mining is easy to setup with minimal infrastructure.

Conclusion: This is an example of a combined text and structured data mining application for identifying Lyme disease patients with minimal infrastructural development.

<table>
<thead>
<tr>
<th>Table 1. Study elements</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample Regex terms</td>
</tr>
<tr>
<td>“Erythema migrans”</td>
</tr>
<tr>
<td>“Embedded tick”</td>
</tr>
<tr>
<td>“Bull’s eye rash”</td>
</tr>
</tbody>
</table>

| Table 2. Groups based on Positive study elements |
|-----------------|-----------------|-----------------|-----------------|-----------------|
| Groups          | Lyme serology   | ICD Diagnosis   | Treatment       | Total Documents |
| A. 1 element present | 0               | 0               | 0               | 484             |
| B. Any 2 elements +  | 288             | 18              | 605             | 911             |
| C. Any 3 elements +  | 263             | 305             | 534             | 551             |
| D. All 4 elements +  | 140             | 140             | 140             | 140             |
| **Total**        | **893**         | **802**         | **1510**        | **2504**        |

References

2065
Evaluation and Adoption of Security Standards to Improve Information Security in Healthcare Administration Systems

Prashansa Rao1, MS and Gunes Koru1, PhD
University of Maryland Baltimore County

Abstract Information security has become crucial in healthcare information systems as high-profile data breaches have increased in the last decade. Unfortunately, the lack of adherence to current information security standards and practices is an important contributing factor. This research identified a number of information security standards and guidelines whose application proved to be useful in determining a large number of security vulnerabilities in a real-life healthcare administration system. This approach and the lessons learned could contribute to further minimizing security breaches in various healthcare administration systems.

Introduction Preserving the security of information stored in the healthcare information systems is becoming an increasingly important yet a challenging target as the utilization, size, and complexity of these systems constantly and substantially increase [1, 2]. This research investigated the relevance of various generic standards and guidelines which can be used to improve the information security in large-scale healthcare administration systems and evaluated the applicability of these security standards in identifying security vulnerabilities in a specific healthcare administration system.

Methodology In this research, we first performed an extensive literature review of the current generic information security standards to healthcare administration systems. The identified security guidelines (e.g., vulnerability scanning) were applied to a real-life healthcare administration system. This system is used by Medicaid providers to attest to the meaningful use of their electronic health records systems. The system maintains meaningful use data for the Medicaid providers in one of the states in the US. Based on the identified guidelings, a prioritized security evaluation plan was prepared to evaluate the system for different vulnerabilities such as Uniform Resource Locator (URL) manipulation, session hijacking, and cross-site scripting (XSS).

Results Four major information security standards, namely, Centers for Medicare and Medicaid Services (CMS) Information Security Acceptable Risk Safeguards, International Organization for Standardization and International Electrotechnical Commission’s (ISO/IEC) 27000 Series, National Institute of Standards and Technology (NIST) Special Publication (SP) 800 series of Computer Security Publication, and Health Insurance Portability and Accountability Act (HIPPA), were identified through the literature review. Twelve security tasks were established which should be followed to secure sensitive information. Application of these in identifying a large number of security problems. Almost 1,000 security warning were found using various software code analyzers which included warnings of potentially malicious code and SQL injection vulnerabilities. The prioritized security evaluation revealed that the healthcare information system was highly vulnerable and susceptible to attacks. During evaluation for URL manipulation vulnerabilities, approximately 1,600 URLs were found to be leaking highly sensitive data, such as names and social security numbers. Lack of documentation of procedures and policies followed for the security increased the overall complexity and aggravated the difficulty of security evaluation and the effort it required.


2066
The Accuracy of Monitoring Stress from Wearable Devices

Maciej Kos, MA, MS1, Christine M. Gordon, MPH2, Xuan Li, MFA3, Iman Khaghani-Far, Ph.D.4, Misha Pavel, Ph.D.5, and Holly B. Jimison, Ph.D., FACMI6

123456College of Computer and Information Science & Bouvé College of Health Sciences, Northeastern University, Boston, MA, United States

Introduction

Continuous monitoring of behaviors as well as physiological and mental states is increasingly considered to be a key prerequisite for optimizing health interventions. The recent influx of affordable wearable sensors may enable continuous and unobtrusive assessment of individuals’ health in ambulatory settings. If such sensors provide data of sufficient quality, we will be better equipped to help individuals improve or maintain quality of life.

Stress has important implications for a wide variety of health conditions1. Hence, there is a clear need for new tools to monitor stress in real time to provide tailored and timely interventions. Although feasible in a clinical setting, ambulatory stress monitoring (ASM) still proves challenging. Even short instruments, like the 10-item Perceived Stress Scale, would be too burdensome for patients to complete multiple times a day. Of great promise in ASM are heart rate variability (HRV) and electrodermal activity (EDA). While numerous studies demonstrated that HRV and EDA are biomarkers of stress and emotion regulation2, relatively few have attempted to use such data obtained from wrist sensors in ASM. As the efficacy of this data capture method hinges upon the quality of obtained measurements, the focus of this poster is on assessing the quality of data collected by two wrist sensors and proposing new algorithms to improve it enough for these sensors to be viable sources of ASM data.

Study design

Physiological data was collected in a laboratory setting using Firstbeat Bodyguard 2 (FB) (interbeat intervals, or IBI), Microsoft Band 2 (MB) (IBI, EDA), Empatica E4 (IBI, EDA), and J&J Engineering I-330-C2+ instrument (EDA). IBI were collected using pulse photoplethysmography sensors (MB and E4) and ECG chest electrodes (FB). Eight participants were led through cycles of relaxation (listening to relaxing music in a dark, quiet room) and stress (viewing International Affective Picture System photos, performing Stroop and mental arithmetic tasks, and engaging in physical exercise) to elicit physiological changes.

Results

Our preliminary results suggest that despite a significant degree of agreement between HRV data collected by wrist sensors and an ECG sensor in a laboratory, there is a relatively high level of dependency of the inferences on context and activity levels. To attenuate these effects, we developed an algorithm based on generative-model that uses singular spectrum analysis (SSA) of the RR interval sequence to separate the low-frequency components of HRV from the disturbances due movements as recorded by the accelerometers. SSA method generates a Toeplitz matrix from the RR sequences and then computes four principal components with the highest eigenvalues. We will describe this analytic process that makes it possible to fuse the accelerometry and the photo plethysmographic data to improve data accuracy. This poster will describe these novel approaches for extracting more precise stress information from noisy bio-signals. We will discuss the advantages of the model-based algorithms in capturing individual-level trends and differences in HRV and EDA over traditional population-based approaches.

Conclusion

Data collected by wearable sensors in a laboratory setting are of high quality, but new algorithms are required to make them useful for ambulatory stress monitoring.

References

An Exploratory Analysis of the Excessive Administration of Seasonal Influenza Vaccinations in Minnesota

Samuel J. Kozloff, MD¹, Miriam Muscoplat, MPH², Zeke J. McKinney, MD, MHI, MPH³
¹New York University, New York, New York, ²Minnesota Department of Health, ³HealthPartners Occupational and Environmental Medicine, Saint Paul, MN

Introduction
Seasonal influenza illness is associated with significant morbidity, mortality and concomitant healthcare spending. The influenza vaccine is known to reduce these undesired outcomes. The complicated logistics of manufacturing and distributing influenza vaccine coupled with variable patient demand can lead to occasional shortages or delivery delays. Excessive influenza vaccine administration may increase healthcare cost and decrease immunization availability. The Minnesota (MN) Department of Health maintains an immunization information system, the MN Immunization Information Connection (MIIC), which includes data submitted by 4,972 vaccination providers, despite no legal requirement except for pharmacists administering vaccines. This research adds to the sparse literature investigating the prevalence of excessive influenza vaccination and identifying targets for intervention.

Methods
The MIIC was queried to generate a dataset of influenza vaccines administered between July 1st, 2009 and June 30th, 2016 to individuals of all ages. Influenza seasons were defined as starting July 1st and ending on June 30th of the next year. Influenza vaccine appropriateness criteria were based on historical Advisory Committee for Immunizations Practices’ (ACIP) guidelines.¹ Inappropriate administration of an influenza vaccine was defined as receiving more than one vaccine per influenza season for individuals nine years and older, and as receiving more than two vaccines in a given influenza season or having multiple seasons in which two or more vaccines were administered for those between six months and eight years of age. All vaccines given to those under six months of age were considered inappropriate. Cost estimates were obtained using Archived CDC Vaccine Price Lists.² An algorithm was created to accurately classify vaccines as appropriately or inappropriately administered, allowing for analysis of excessive vaccination patterns that was otherwise not easily possible with the MIIC data.

Results
The cohort comprised 4,137,453 individuals and 13,339,584 immunizations. 278,700 individuals received inappropriate influenza vaccines (6.56% of all individuals). A total of 332,041 inappropriate vaccines were administered (2.43% of all vaccines). Of individuals who received inappropriate vaccinations since 2009, 84.9% received only one dose more than the number recommended. These influenza vaccines were reported to MIIC by 2,009 different organizations. Ten clinics accounted for 30.3% of all inappropriate vaccinations. Individuals who received multiple inappropriate vaccinations in an influenza season more commonly received these vaccines from multiple clinics. The age group that received the most inappropriate vaccinations was six-month-old infants, followed by middle-aged adults (57 years). The cost of inappropriate administration was estimated to be $4,213,469.

Conclusions
Inappropriate administration of influenza vaccines is not an insignificant issue in MN. Most instances of excessive administration involved one inappropriate dose, and most cases occurred when vaccination was administered from multiple clinics in a given influenza season. The study results suggest populations for informatics-based interventions, including clinics and age groups with the most frequent inappropriate administrations. A known limitation of this data is primarily voluntary data submission to MIIC. More reliable reporting of vaccine administration to MIIC and better incorporation of MIIC into existing clinical workflows could reduce the incidence of vaccine overutilization, thus reducing unnecessary healthcare spending, decreasing the potential for adverse vaccine reactions, and potentially maintaining local availability of immunizations.

References
Introduction

Behavior and lifestyle habits play an important role in defining long-term health outcomes of an individual. This is especially true for an individual burdened by a chronic condition, whose behaviors may directly impact the management of the condition. The recent rise in popularity of digital health devices, activity trackers and apps has created the opportunity to quantify behavioral patterns with the ultimate goal of identifying those that are associated with specific health outcomes. In this study, we consider a population of patients with digital activity trackers who self-reported as diabetic. Research has shown that individuals with type 1 or type 2 diabetes have a higher risk of developing mental health illnesses (MHI) such as depression and anxiety than individuals without diabetes. However, many individuals with diabetes experience undiagnosed MHIs due to inadequate screening or diagnosis. The goal of this study is to understand whether there are significant digital activity tracker-based behavioral characteristics that can help identify individuals with MHI symptoms in a population of individuals with diabetes. Early identification of MHI comorbidities may enable interventions to prevent poor diabetes management and progression of MHI.

Methods

Members of an online health community that use activity trackers and apps were invited to participate in a survey that included self-reports of diagnoses and symptoms, ratings of perceived overall health status, physical health, and mental health, and characteristics such as age and gender. We used the members' activity information recorded from activity trackers and apps over the past 12 months to compute per-individual daily features that indicate the tracking rate, consistency, and intensity of step counts and sleep duration. When comparing digital activity tracker-based behavioral characteristics between individuals with diabetes and individuals with diabetes and MHI, we conducted nonparametric tests and t-tests, and examined the false discovery rate.

Results

A total of 1,330 members reported a diagnosis of diabetes, and 336 (25.3%) of those individuals reported having MHI symptoms in the last year. Of those who reported MHI symptoms, 259 (77.1%) reported having anxiety symptoms and 237 (70.5%) reported having depression symptoms. Among individuals diagnosed with diabetes, those with self-reported symptoms of a MHI walk on average 1,469 steps per day less than individuals without MHI symptoms (p<0.001). Those who reported MHI symptoms had significantly lower frequency of tracking steps than individuals without symptoms (p<0.001). These individuals also had significantly less frequent days with high activity (more than 10,000 steps/day) and more frequent days with low activity (fewer than 500 steps/day) compared to individuals with no reported MHI symptoms (p<0.01). Despite similar mean sleep lengths between both groups, individuals experiencing MHI symptoms slept for longer durations (more than 9 hours/day) significantly more frequently than individuals without symptoms (p<0.01).

Conclusion

Our preliminary results indicate that digital activity tracker data can help identify behavioral traits associated with self-reported symptoms of MHI in a population of individuals with self-reported diabetes. Traditionally, clinicians use self-reported data to screen and measure the progression of MHI. Digital activity tracker-based behavioral data may enrich and supplement this self-reported data, and can potentially help detect MHI and track MHI symptoms. Further research should focus on detecting behavioral markers of onset of and progression of MHI symptoms in individuals with diabetes as soon as they begin to develop.

References

Mapping U.S. FDA National Drug Codes to Anatomical-Therapeutic-Chemical Classes using RxNorm

Fabrício S. P. Kury, MD¹, Olivier Bodenreider, MD, PhD¹
¹U.S. National Library of Medicine, U.S. National Institutes of Health, Bethesda, MD, USA

Background

The U.S. Food and Drug Administration National Drug Codes (NDCs) are the official identifiers for drug products in the U.S. and therefore commonly appear in large drug prescription datasets from this country. To analyze these datasets from a clinical perspective, it is often useful to abstract away from details such as manufacturer, dosage form and packaging information, and to group prescriptions by drug classes. The Anatomical-Therapeutic-Chemical¹ drug classification system provides a 5-level aggregation scheme for drugs. However, neither WHO nor FDA provides a map from NDC to ATC. Here we demonstrate how to use the U.S. National Library of Medicine RxNorm² application programming interface (API) to automatically create such a map. We apply this by mapping 71,309 NDCs found in 1.015 billion Medicare Part D claims from 2006 to 2013, and 134,580 NDCs found in a commercial all-payer claims dataset from Partners Healthcare© from 2011 to 2012.

Methods and Results

We produced an R script and made it freely available under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International license in the GitHub repository at https://github.com/fabkury/ndc_map. The script uses caching and parallel processing to greatly improve efficiency. The script ingests a list of NDCs, queries the RxNorm API to obtain their RxNorm Concept Unique Identifiers (RxCUIs), then queries the API again to obtain the fourth-level ATC class or classes (ATC-4) of each RxCUI. Therefore, our mapping can fail at two stages: when NDCs are not recognized by RxNorm, and when an RxNorm drug has no associated ATC-4 class. Tables 1 and 2 provide statistics about the mapping of the two datasets under investigation.

<table>
<thead>
<tr>
<th>Table 1. Results of the mapping process.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mapped prescriptions: 55,565 (77.9%)</td>
</tr>
<tr>
<td>Mapped NDCs: 55,565 (77.9%)</td>
</tr>
<tr>
<td>NDCs with no RxCUI: 10,413 (14.6%)</td>
</tr>
<tr>
<td>RxCUI but no ATC-4: 5,331 (7.5%)</td>
</tr>
<tr>
<td>Unique NDC—ATC-4: 114,069</td>
</tr>
<tr>
<td>Unique ATC-4 classes: 517 (58.62%)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 2. Tallies of ATC-4 classes per NDC.</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATC-4 per NDC</td>
</tr>
<tr>
<td>0</td>
</tr>
<tr>
<td>1 to 2</td>
</tr>
<tr>
<td>3 to 8</td>
</tr>
<tr>
<td>9 to 21</td>
</tr>
<tr>
<td>≥ 22</td>
</tr>
</tbody>
</table>

Discussion and Conclusion

Medicare claims contained a smaller set of NDCs which were proportionally better covered by RxNorm. In the Medicare dataset, only 77.9% of the NDCs were mapped to at least one ATC class, but those NDCs represented 97.6% of all claims. From an informal analysis of the Medicare data, the majority of the non-mapped claims were for over-the-counter medications or non-drug supplies (e.g. syringes, gauze). In both datasets, individual NDCs were ambiguously associated with up to 21 ATC-4 classes for two reasons: some drugs have more than one active ingredient with different ATC classes; and ATC can provide multiple classes for the same ingredient depending on its usage. For example, miconazole has 6 ATC level 4 classes across oral use, gynecological use, otological use, and so on. If not addressed, the NDC-to-ATC ambiguity can create issues in analyzing prescription datasets. Conclusion: Our approach leveraging RxNorm provides a simple and scalable solution for mapping NDCs to ATC classes. The ambiguity of some mappings is the biggest concern for researchers using the NDC-to-ATC map.

References

Rapid Low-Cost Usability Engineering: Can We Demonstrate the Cost-Effectiveness of the Method for Ensuring Usability of Health IT?

Andre W. Kushniruk, MSc, PhD, FACMI1, Elizabeth M. Borycki, RN, MN, PhD1

1School of Health Information Science, University of Victoria, Canada

Introduction

Worldwide there are increasing reports about the poor usability of health information technology (HIT) being deployed. This has included concerns about electronic health record systems (EHRs), decision support systems, mobile applications and other types of HIT applications. To address this, we describe an approach we have developed known as rapid low-cost usability engineering and present a cost-benefit analysis case for its application.

Methods

The authors have been involved in the development and refinement of an approach to usability engineering in healthcare they have termed rapid low-cost usability engineering. The approach argues that studies can be conducted in-situ, at low cost and rapidly in real clinical settings. The approach typically involves recording the providers’ interactions with systems in a real or simulated environment. Screen recordings of all computer interactions are captured and integrated with audio transcripts to facilitate rapid coding of the data using coding schemes that can be adjusted to the application being tested. The protocols are coded to identify and statistically summarize usability issues and previous studies have shown that the method can be applied rapidly. In order to determine the cost-effectiveness of the approach, data from studies were re-analyzed where total costs of applying the method were calculated (e.g. costs of equipment, participant remuneration, costs of analysis etc.) in order to document the costs of applying the approach. After analyzing the usability data for usability problems, estimates were made (using tables of software costs) to determine the cost of fixing identified usability problems, in order to estimate tangible benefits of applying the approach to catch usability problems. The cost-benefit ratio was then calculated.

Results

From analyses of the results from several usability studies where the rapid low-cost usability engineering approach was applied (and the costs and benefits calculated as described above), it was determined that benefits outweighed costs for applying the approach in all cases, with percent cost savings ranging from 36.5% to over 200%. Furthermore, the total cost for conducting the studies was typically well under $10,000 USD per study. In addition, the methodology also detected errors that would not be considered usability problems, including programming bugs, safety issues and technology-induced errors (that would have otherwise not have been detected prior to widespread implementation). The approach was therefore found to be useful in detection of both usability problems and technology-induced error.

Discussion

We have developed an integrated approach to detect serious usability problems that can serve as a safety net for catching problems. In order to provide evidence that the method is cost-effective we conducted analyses of several usability evaluations to determine the benefits of the approach. Future work will include plans for knowledge translation and wider dissemination of the method through educational initiatives and design of training programs.

References

Reach and Impact of an EHR Pain Care Clinical Decision Support Program
Steven Labkoff, MD, FACP, FACMI 1, Christopher Bond, PhD1, Lee Kallenbach, MPH, PhD2, Nam Nguyen, MS2, Shruti Gangadhar, MS2, Dan O’Brien2
1Purdue Pharmaceuticals, Stamford, CT; 2Practice Fusion, San Francisco, CA

Introduction
Pain care is one of the top reasons for ambulatory office visits in the US. Assessment of pain is a critical step to providing good pain management. To further both general and chronic pain care management, a manufacturer of pain medications engaged an EHR vendor to implement a clinical decision support (CDS) program aimed at pain care.

Methods
Based on current evidence-based clinical guidelines three pain care-related CDS alerts were deployed July 5, 2016 on the Practice Fusion EHR platform (PF-EHR)1. The PF-EHR is cloud-based with practices in all 50 states facilitating over 5M office visits per month representing 6.7% of US ambulatory care. CDS alerts are displayed on the PF-EHR in real-time when a provider opens a patient chart for editing that meets specified logic. The program includes three CDS alerts: 1) document a pain score in the vital signs flowsheet view for adult patients without chronic pain; 2) complete a brief pain inventory (BPI) assessment for patients with chronic pain; and 3) document a pain care plan for patients with a recent pain score or BPI. The CDS program consisted of matched test and control practices with data evaluated for a six-month period prior to program initiation. Data for the first six months of the program are reported.

Results
Alerts were displayed to 103,717 health care professionals (HCP) during 24,820,820 office visits for 8,355,683 patients at 18,896 practices. Most alerts were for entry of pain scores with the remainder for completing the BPI or for documenting a patient pain care plan. Response to alerts was sustained throughout the 6-month period. (Figure 1).

Figure 1. Pain Care CDS Program Alert Responses by 2-Week Intervals.

Impact of the program is noted by an increase in pain score documentation for an existing EHR data element and by number of BPIs and care plans entered for new EHR data elements. Compared to control practices, test practices entered pain scores, BPIs, and care plans 1.2, 341, and 777 times as often, resulting in 123,662, 38,218, and 29,466 patients with responses attributable to the program, respectively.

Conclusion
Provider acceptance of pain care plan prompts on the PF-EHR platform was positive with completion of prompted actions sustained over six months. CDS provides a method for increased documentation of pain-related care data. Further work will evaluate the impact of this additional data on patient level care and clinical outcomes.

References

Abstract is entered directly into the web site form for a Poster presentation (50-75 words). Abstract:
Assessment of pain is critical to providing good pain management. Based on evidence-based clinical guidelines three pain care clinical decision support alerts were deployed on the Practice Fusion cloud-based EHR platform. The program design consisted of matched test and control practices. Provider acceptance of pain care prompts was positive and sustained over six months with completion of prompted actions including pain scores, brief pain inventories, and care plans higher in test than control practices.
Substance Use Extraction in Cancer Patients' Clinical Notes

William Lane, MS¹, Spencer Morris, MS¹, Emily Silgard, MS¹
¹Fred Hutchinson Cancer Research Center, Seattle, WA

Introduction

The development of software systems that can reliably extract clinical events from free-text clinical notes could prove of immense value to clinicians, health care management personnel, as well as medical researchers. The task is a challenging one, and much work has been done around extracting smoking status information, particularly in the context of the 2006 i2b2 shared task¹. The intervening years have seen much success for researchers applying NLP to smoking status detection. This work expands on those efforts by attempting to extract not only smoking status, but also alcohol use status as well as the relevant attributes of the particular substance use. Extracting substance quantities like “pack-years” in smoking², for example, are important for categorizing more fine-grained exposure.

Data and System Structure

The data set consists of 1,450 clinical notes annotated on the sentence level for alcohol and tobacco status (Current, Former, User, Non), as well as for each substance type’s associated attributes (e.g. amount, duration, quit date, and time ago quit) belonging to patients treated at the Fred Hutch/University of Washington Cancer Consortium between 2010 and 2015. Our system takes a batch of clinic notes as input and first filters sentences for alcohol or tobacco-related references according to an extensive library of regex patterns. We implicitly label sentences not containing any hits as an “unknown” status for the particular substance type. The remaining sentences are fed to a linear SVM and classified as alcohol or tobacco related. All sentences classified as relevant to tobacco or alcohol use are put into a multi-class linear SVM where lexical features are used to classify the substance use status. The set of sentences is then passed to the attribute extraction module, where attributes related to the substance use event and status are extracted using a CRF. The final module uses another SVM to match attributes with their appropriate substance event and status, using a combination of lexical and contextual features. Document-level status values were determined by organizing status values into a hierarchy. Attributes were selected on a document level by selecting the attribute value for a particular field which had the highest confidence score output by the CRF calculation.

Results and Discussion

<table>
<thead>
<tr>
<th>Tobacco F1 Scores</th>
<th>Alcohol F1 Scores</th>
<th>Tobacco F1 Scores</th>
<th>Alcohol F1 Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>Status</td>
<td>84.8%</td>
<td>Status</td>
<td>84.0%</td>
</tr>
<tr>
<td>Quit Time Ago</td>
<td>79.3%</td>
<td>Quit Time Ago</td>
<td>22.4%</td>
</tr>
<tr>
<td>Duration</td>
<td>78.3%</td>
<td>Duration</td>
<td>30.0%</td>
</tr>
<tr>
<td>Quit Date</td>
<td>80.2%</td>
<td>Quit Date</td>
<td>52.2%</td>
</tr>
<tr>
<td>Amount</td>
<td>73.4%</td>
<td>Amount</td>
<td>63.4%</td>
</tr>
<tr>
<td>Type</td>
<td>66.0%</td>
<td>Type</td>
<td>N/A</td>
</tr>
</tbody>
</table>

*Results of status classification and attribute extraction in 5-fold cross validation

The performance of the system varied depending on the attribute being extracted, and some attributes' performance suffered greatly due to overall sparsity in the training set. For example, it is far less common for a clinician to report on alcohol cessation dates than to report on tobacco cessation dates, and the performance gap between those two figures demonstrates the limitation of a system driven purely by supervised learning. Error analysis on the system output showed that the CRF component was actually quite effective at extracting general amounts and dates, but that the final numbers were significantly diminished by the final SVM classification step which assigns attributes to substance events. In future work, we plan to enhance the machine learning approach with domain-specific and situational logic to mitigate the effect of a single weak link in the process propagating errors to subsequent layers.

References

New MetaMap Features for Processing Numerical Tables
François-Michel Lang, MSE, Dina Demner-Fushman, MD, PhD
US National Library of Medicine, Bethesda, MD

Abstract
A recent project using MetaMap to identify adverse reactions in drug labels encountered two serious problems: long runtimes and low precision. Analysis revealed that tables containing many numbers caused both problems; we present two MetaMap enhancements that substantially improved MetaMap's runtime and precision with no loss in recall.

Introduction
The US National Library of Medicine (NLM) collaborated with the FDA’s Center for Drug Evaluation and Research on an evaluation of 476 Structured Product Labels (SPL) sections provided by FDA in XML format. NLM's contribution consisted of (1) Creating a Gold Standard via manual annotation of the 476 label sections to identify adverse reactions (ARs); (2) Running MetaMap1 on the ASCII text extracted from the 476 sections, restricting results to UMLS concepts having Semantic Type(s) in the Disorder Semantic Group2; and (3) Evaluating MetaMap's performance relative to the Gold Standard. This poster describes two new optional MetaMap features that greatly improve MetaMap's runtime and precision.

Methods
SPLs contain many numerical tables, such as the following small example:

| Rash | 1 0 6 | 0 0 |
| Pruritus | 2 0 5 | 0 0 |
| Alopecia | 10 0 0 | 5 0 |

- **Runtime problem:** MetaMap’s phrase chunker discovers no internal syntactic structure in such text, because it is simply a noun pile with interspersed numbers. Consequently many tables are analyzed as a single long phrase, requiring extensive processing time. To solve this problem, we modified the chunker to force a phrase break after encountering two consecutive numerical tokens. The much shorter resulting phrases led to far shorter runtimes.

- **Precision problem:** MetaMap is a named-entity recognition (NER) application that maps text to UMLS concepts; however, many numerical strings, such as those in the above table, appear in the UMLS, but contribute only noise, thereby severely degrading precision. We substantially improved precision via the ability to ignore concepts (1) containing no alphabetic character and (2) consisting of ≥50% numeric characters.

Results
- **Runtime improvement:** We observed an overall 2.2x speedup; 9 files showed >10x speedup, and 2 files containing very long tables improved by >50x.

- **Precision improvement:** We observed an increase of 35% in precision (from 31% to 42%), with no loss of recall. 17 files showed a doubling of precision, and one a quadrupling. Even so, MetaMap's precision on this project was far lower than normal, because although a vast majority (>97.5%) of ARs identified are in CUIs with a Disorder SemType, many disorder concepts are not ARs: E.g., in “Do not administer to patients with infections”, infections is a disorder concept, but not an AR. These results demonstrate that AR recognition needs processing beyond NER (e.g., semantic analysis)—although NER could be a good first step as it provides good recall. An ancillary benefit of this enhancement was an improvement in the performance of NLM’s Medical Text Indexer (MTI)3, which should result in ~1,000 additional correct MeSH recommendations each year.

Conclusions
These new features show that special-purpose processing in MetaMap can substantially improve the processing of tables, regardless of whether created via screen scraping or automated extraction from markup such as XML.

Acknowledgements
This work was supported by the intramural research program of the NIH, U. S. National Library of Medicine.

References
Leveraging Clinical Decision Support Tools to Improve Early Recognition and Treatment of Pediatric Sepsis

Michael E. Lang, MD, MPH\textsuperscript{1}, Anna Keller, RD\textsuperscript{1}, Robert S. Heidersbach, MD\textsuperscript{1}, Heidi Flori, MD\textsuperscript{2}

\textsuperscript{1}UCSF Benioff Children’s Hospital Oakland, CA; \textsuperscript{2}C.S. Mott Children’s Hospital Michigan, Ann Arbor, MI

Abstract

Septic shock is an important cause of morbidity and mortality for pediatric patients. As part of a quality improvement project, we implemented clinical decision support tools (CDS) to help clinicians recognize signs of sepsis and to direct them to follow our sepsis clinical practice guideline (CPG). While our outcomes were not statistically significant, length of stay (18.2 to 15.6 days) and mortality (13.4% to 7.1%) trended towards improvement post-CDS intervention.

Introduction

Worldwide, sepsis is responsible for greater than 1,000,000 deaths per year and is one of the top ten causes of death in children in the U.S. (mortality ranges from 4-11%)\textsuperscript{1}. Studies have shown that early recognition and treatment of sepsis improve mortality rates\textsuperscript{2}. We embarked on a five-year process improvement project, funded by U.S. Army Medical Research and Material Command Grant W81-XWH-10-01-0682, to determine if CDS tools in our EHR can help guide clinical practice to improve sepsis outcomes.

Methods

Our project was divided into three phases: Phase 1, Baseline Assessment (7/2010 – 12/2011); Phase 2: CPG Education (1/2012 – 11/2013); Phase 3: CDS Intervention (11/2013 – 12/2015). We used the International Consensus Conference on Pediatric Sepsis\textsuperscript{3} as a basis for our CPG and study inclusion criteria. Patients admitted to the UCSF Children’s Hospital Oakland PICU for severe sepsis/septic shock were enrolled in our study. Clinicians (physicians, mid-level providers, and nurses) received education on the new CPG (Phase 2). In Phase 3, we created an alert that appears when clinicians open the chart of a patient meeting criteria for severe sepsis/septic shock. The alert presents ordering providers an order set modeled after the CPG, thus encouraging them to adhere to the CPG. We obtained feedback from clinicians throughout all three phases, leading to a rebuild of the alert in 7/2015.

Results

For each of the three phases, the numbers of patients meeting severe sepsis criteria were 50, 67, and 56, respectively. Process measures and clinical outcomes were determined in each phase. None of the results were statistically significant; however, our data showed a trend towards improvement in length of stay (18.6, 18.2, and 15.6 days) and mortality (10%, 13.4%, and 7.1%) through the three phases, respectively. Clinicians reported too many false positives for the alert. The alert discrete triggering parameters and display were adjusted in 7/2015, which led to anecdotal reports of improvement.

Conclusion

Although our process improvement project failed to improve sepsis outcomes, our post-CDS intervention length of stay and mortality were lower than pre-intervention.

References

Understanding Ovarian Cancer Symptoms from Patients’ Perspective: A Topic Modeling Approach

Young Ji Lee, PhD, RN1,2, Albert Park, PhD4, Mary Roberge, BSN, RN1, Heidi Donovan, PhD, RN1,3

1School of Nursing; 2Dept. of Biomedical Informatics; 3Dept. of Obstetrics, Gynecology & Reproductive Sciences, School of Medicine, Univ. of Pittsburgh, PA; 4Dept. of Biomedical Informatics, Univ. of Utah, Salt Lake City, UT

Introduction: Ovarian cancer is the fourth leading cause of cancer-related deaths in women in the United States1. Patients with ovarian cancer often experience multiple symptoms that adversely affect their quality-of-life2. To manage cancer- and treatment-related symptoms more effectively, we must understand symptoms from the perspectives of the patients. Social media has created new opportunities for gaining understanding of the patient experience through not only large volumes of data, but also the nature of information that is distinct from data obtained in clinical settings1. However, to date, no studies have utilized data gathered from social media data in this way to understand symptoms among ovarian cancer patients. The purpose of this study is to understand major topics of communication (e.g., social, informational, management advice) related to the symptoms frequently discussed on social media by women and family members affected by ovarian cancer.

Materials and Methods: The data was collected from the online forum of a peer-support community sponsored by the American Cancer Society: Cancer Survivors Network (http://csn.cancer.org). For this study, we used posts from January, 2006 to March, 2016. To identify symptoms discussed in the online forum, we developed a dictionary that includes a list of ovarian cancer symptoms using consumer health vocabulary. After pre-processing data, we selected the top 5 most frequently discussed symptoms and extracted and analyzed postings that mentioned any of those five symptoms. Then, we conducted Latent Dirichlet Allocation using R.

Results: In total, we used 50,626 postings made from 1,947 users. 52% of users (n=1,009) and 25% of postings (n=12,802) mentioned at least one symptom. Among 28 symptoms, pain, nausea, anxiety, fatigue and skin rash were the top five most frequently discussed symptoms in the forum (n=4536, 1296, 967, 878, 657 respectively). Topics of the postings differed across the symptoms. Our analysis resulted in 20 topics that could be further classified into 11 categories: social and spiritual support, communication with doctors, decision about treatment, fear of recurrence, stories about diagnosis and treatment, returning to normal life, complications, adjusting to treatment, chemo-related side effects, symptom management, and miscellaneous.

Figure 1. The distribution of topics in the pain postings.

Chemotherapy-related side effects were most likely to refer to fatigue, nausea, and rash while discussions of spiritual and emotional support were likely to refer to anxiety, and discussions of diagnosis and treatment were often referred to pain.

Conclusion: Text mining has shed new light on health research by identifying useful information from unstructured text. Our results showed that users with different symptoms addressed different topics in the forum. Our next study will address the influence of additional factors (e.g., age, cancer stage, time) on topic discussion to refine our approach and reflect contextual factors.

References
Meeting Common Health-Related Needs Through a Pediatric Inpatient Engagement Consultation Service

Authors: Daniel J. Lee MD, Robert M. Cronin MD, Jamie Robinson MD, Kim Unertl PhD, Shilo Anders PhD, Katherine Kelly BS, Gretchen P. Jackson MD PhD

1Departments of Urology, 2Biomedical Informatics, 3Internal Medicine, 4Pediatrics, 5Surgery, 6Anesthesiology, 7Pediatric Neurosurgery, Vanderbilt University Medical Center

Background
Engaged patients actively involved in their health care have greater satisfaction with care, decreased costs, and improved outcomes. Patients are increasingly using technologies for their health care. Access to electronic health records, online resources, and other technologies may promote engagement by providing tools to review health information and participate in their care. Hospitalizations offer opportunities to engage patients and caregivers in their child’s care. In 2014, we developed a Pediatric Patient and Family Engagement Consultation Service at the Monroe Carell Jr. Children’s Hospital at Vanderbilt to assess the health-related needs of the caregivers of hospitalized patients and to promote engagement and meet needs with educational or technology interventions. Consultations focus on supporting families with new diagnoses and chronic illnesses using internet-based or mobile resources, the My Health at Vanderbilt (MHAV) patient portal, or other technologies to assist with education, disease monitoring, behavior change, or emotional support.

Methods
We sought to describe the most common needs identified and the technology recommendations provided by this service. We examined 22 consultations from a retrospective and a prospective study. Both studies were approved by the Vanderbilt University Institutional Review Board. The analysis aimed to determine the most common health-related needs identified in the pediatric hospital setting. Each health-related need was classified according to a validated taxonomy of consumer health needs, and the technology recommendations for each need were described.

Results

NEED #1: Need for Provider Contact Information. NEED TYPE: Logistical-Contact/Communication. A health-related need expressed by nearly every parent seen by the consult service is the need for a means to contact multiple healthcare providers for follow up and ongoing care. RECOMMENDATIONS: Parents were registered for the MHAV patient portal and a link to the child’s health information was established. New providers were encouraged to send messages with contact information to the parent through the portal prior to discharge. Extensive time was spent teaching parents to navigate the MHAV patient portal so that they gained experience with messaging providers, checking pertinent test results, and looking up or changing appointments.

NEED #2: Recognizing Common, Benign Versus Concerning Symptoms. NEED TYPE: Informational-Problem-Clinical Presentation. Another common need was distinguishing normal or benign symptoms of infancy or childhood (e.g., baby spit ups) from concerning signs or symptoms of the patient’s medical problems. Parents often had difficulty identifying trustworthy resources online and were often left confused by the breadth of information from a Google search. RECOMMENDATIONS: The consult service educated the parents on how to identify trustworthy sources, such as the American Academy of Pediatrics website, and helped provide easy access to sites by bookmarking them on their mobile devices.

NEED #3: Remembering Questions for Rounds. NEED TYPE: Logistical-Contact/Communication. Another common need was remembering and keeping track of questions for hospital rounds or the answers provided by the care team. RECOMMENDATIONS: For this need, the consult team recommended utilizing NotePad apps or voice recording features present on nearly all the smartphones and tablets. Most parents used such devices frequently, but had not thought of utilizing these tools for health.

Conclusions
The parents of hospitalized patients share several common health-related needs that can be addressed by technologies. An inpatient Engagement Consultation Service can identify and meet needs and promote engagement with health information technologies.
Predictive Modeling for the Prevention and Surveillance of Adolescent Substance Abuse and its Implications for Nursing

Mikyoung Lee, PhD, RN1, Dawn Murphy, MS, RN, CPNP1, Suzanne Gillson, MSN, RN2, Leolin Castillo, PMHNP, RN1 & Anila Simon, MSN, RN1
1Texas Woman’s University College of Nursing, Denton, TX, USA; 2Houston Baptist University, Houston, TX, USA

Introduction
Substance abuse in adolescents is a significant population health concern and a risk factor for other health problems and mortality in adolescents1. Predictive analytics have served as a useful tool for healthcare decision-making, and the value of such a process is more recognized with emerging big data science2. It is valuable to review significant predictors related to substance abuse in adolescents so that nurses can transfer the knowledge into actions for prevention and management of substance abuse and relevant health issues. This systematic literature review aimed to identify what predictive analytics have been used and what predictors have been statistically significant to predict the onset and trajectory of substance abuse and subsequent health issues in adolescents.

Methods
The databases of CINAHL, PubMed, and Child Development & Adolescent Studies were searched with the following use of Boolean operations and keywords: ['Substance use' OR 'Substance abuse'] AND ['adolescent*' OR 'Teen*' OR 'Youth*'] AND ['nurs*' OR 'health*'] AND ['predict*' OR 'forecast*' OR 'modeling*']. The search was limited to peer-reviewed articles published in English between 2012 and 2017. 131 articles were found. The duplicates (n=39) were excluded. Studies irrelevant to the study aim, qualitative studies, editorial/commentary, and review articles were excluded after a quick view of the abstracts; 46 articles were reviewed using a pre-determined review table.

Results
Common data sources were national data sets, such as the National Longitudinal Study of Adolescent Health, National Survey of Drug Use and Health, followed by research data, school-level data, and state-level data. Sample size ranged from 154 to 126,764 participants. The major predictive analytic methods were logistic regression analysis, structural equation model, and latent growth curve modeling. Popular predictors with statistical significance for substance abuse were age, gender, family dynamics, peer substance use, and peer networks. Early onset or prior experience with substance abuse significantly predicted later substance abuse. The common health issue related to substance abuse was mental health, including depression and self-esteem. An interesting predictor identified was cyber-bullying.

Conclusion
This literature review identified the multifactorial complexity of substance abuse in adolescents and the strong need for multifactorial approaches to them. With the identified predictors, nurse clinicians, educators, and researchers can develop programs targeted adolescent substance abuse. Health education should be directed to adolescents, their peer groups, and other significant relations. It is important for nurses to assess adolescents with regard to their network of social interactions and engagements. As social media use becomes more prevalent in adolescents, the assessment and interventions with peer networks using social media will be a promising strategy. This review recognized the importance of assessment and intervention of substance use and relevant mental health issues in the early adolescence to prevent sustained or worsen substance abuse in later adolescence or young adulthood. Nurses should also consider family dynamics while designing and implementing preventive and therapeutic programs for adolescent substance abuse. This literature review also identified a lack of nursing-focused studies in relation to predictive analytics and substance abuse in adolescents. It is recommended for the nursing profession to apply predictive modeling to data from public and nursing-driven sources to create substance abuse prevention/surveillance programs for adolescents.

References
Analyzing Patient Data Entry Patterns of Free Text based Family History Disease Concepts

Jaehoon Lee, PhD¹², Nathan C. Hulse, PhD¹², Stacey Knight, PhD¹²
¹Intermountain Healthcare, Murray, UT; ²School of Medicine, University of Utah, Salt Lake City, UT

Abstract

In this study we used a patient-facing family health history (FHH) tool to analyze free-text based reporting of diseases to understand common patterns of data entry of patients. From 6,132 patients, we retrieved 28,982 reported disease of which 2,094 (7.2%) were free-text entries. The key findings from manual analysis of the free-texts were 1) significant number of typos (13.7%) 2) and differences in data entry patterns by age group and gender.

Background

FHH reported by patients has been recognized as valuable information for genetic risk assessment and patient-facing FHH tools that capture these data in a non-clinical setting are desirable. However the data quality of these FHH reported diseases may be questionable due to the patient’s lack of clinical knowledge, especially with free-text reporting. Understanding the free-text entry patterns in FHH collection tools will help enhance existing FHH data model, and improve user interface of the tools to capture correct FHH.

Method

OurFamilyHealth is a web-based FHH tool developed at Intermountain Healthcare and allows patients to enter their FHH interactively with a graphic pedigree drawing tool. Patients can enter diseases for family members by two methods; one is to choose from predefined list of 552 diseases that are linked to standard terminologies; the other is to enter the diseases of family members as free-text. Using descriptive statistics, we analyzed the entry pattern from 6,132 FHH records captured through OurFamilyHealth from 2009 to 2016.

Results

From the 6,132 FHH records there were 28,982 reported diseases of which 2,094 (7.2%) were free-text based entries. The free-text reporting was most often done by female patients and by patients between the ages of age 30-50 years old. The most frequently entered free-text disease were Lupus (44, 2.1%); Endometriosis (40, 1.9%); Skin cancer (39, 1.9%); High blood pressure (29, 1.4%); and Hypothyroidism (27, 1.3%). Of the free-text entries 287 (13.7%) contained typos and 1,225 (58.5%) of the entries matched to a Systematized Nomenclature of Medicine (SNOMED) term. There are differences in free-text reporting by sex and age (Table 1), with rates of typos increasing with age and matching to SNOMED concept tending to be higher in the younger age groups.

Table 1. Descriptive statistics of free text based FHH disease entry by gender and age group.

<table>
<thead>
<tr>
<th>Age group</th>
<th>Male (#patient: 166, #disease: 13244, #freetext: 391)</th>
<th>Female (#patient: 592, #disease: 15738, #freetext: 1683)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>#Entry</td>
<td>#Typo</td>
</tr>
<tr>
<td>~30</td>
<td>32</td>
<td>3 (9%)</td>
</tr>
<tr>
<td>31~40</td>
<td>154</td>
<td>20 (3%)</td>
</tr>
<tr>
<td>41~50</td>
<td>97</td>
<td>12 (12%)</td>
</tr>
<tr>
<td>51~60</td>
<td>48</td>
<td>13 (27%)</td>
</tr>
<tr>
<td>61~</td>
<td>60</td>
<td>13 (22%)</td>
</tr>
</tbody>
</table>

Conclusion

Among patients using a web-based FHH tool, we found that females and younger patients use free-text based entry more actively and with less typos. More than half the free-text entered diseases could be mapped to the standard terminology. In future research we will enhance our disease list by adopting frequently entered disease concepts and improve our tool to support patients with user interfaces to prevent or autocorrect typos.

References

Spanish Text Simplification Using Term Familiarity: Applying Principles from English Text Simplification

Gondy Leroy, PhD1, Brianda Armenta Navarrete1, Sonia Colina, PhD1, David Kauchak, PhD2
1University of Arizona, Tucson, Arizona; 2Pomona College, Pomona, California

Introduction
Health literacy is important for patients; text simplification is one tool to help facilitate the comprehension of information provided to them. Previously, we have developed and validated several features for simplifying English text. We report here on one such feature Term Familiarity, and how it applies to Spanish text simplification.

Simplifying Spanish Text Using the Term Familiarity Algorithm
We manually simplified four Spanish texts based on Term Familiarity (Table 1). We identified difficult words based on low term familiarity (low frequency in the LexEsp corpus) and either 1) replaced them with a higher frequency synonym or, when none was available, 2) added an explanation of the word. We integrated short explanations in the text, but added longer explanations between parentheses. We adjusted gender agreement as necessary.

Table 1. Text Statistics

<table>
<thead>
<tr>
<th>Topic</th>
<th>Avg. Term Familiarity (Freq. in LexEsp)</th>
<th>Terms Simplified in Easy Text</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Easy (total words)</td>
<td>Difficult (total words)</td>
</tr>
<tr>
<td>Diabetes</td>
<td>1006 (366)</td>
<td>1008 (298)</td>
</tr>
<tr>
<td>Hypertension</td>
<td>1011 (363)</td>
<td>1012 (311)</td>
</tr>
<tr>
<td>Asma</td>
<td>1175 (293)</td>
<td>1162 (296)</td>
</tr>
<tr>
<td>Tabaquismo</td>
<td>632 (463)</td>
<td>717 (604)</td>
</tr>
<tr>
<td>Average</td>
<td>956</td>
<td>987</td>
</tr>
</tbody>
</table>

Pilot Study and Results

Study Design: We recruited six native Spanish speakers without a medical background. Each read two easy and two difficult texts (Table 1) that were randomly assigned and in random order. On average, twelve (8+4) words were simplified in the texts. For example “En este artículo se repasan …” was replaced with “En este texto se revisan”. We measured perceived difficulty with a Likert scale (lower is easier), actual difficulty (accuracy) with 10 multiple-choice questions per text (after reading the text), and total reading time.

Table 2. Pilot Study Results

<table>
<thead>
<tr>
<th>Topic</th>
<th>Perceived Difficulty (1 = Very Easy, 5 = Very Difficult)</th>
<th>Actual Difficulty (Accuracy)</th>
<th>Reading Time (Minutes)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Easy</td>
<td>Difficult</td>
<td>Easy</td>
</tr>
<tr>
<td>Diabetes</td>
<td>3.0</td>
<td>3.0</td>
<td>76</td>
</tr>
<tr>
<td>Hypertension</td>
<td>3.3</td>
<td>4.0</td>
<td>85</td>
</tr>
<tr>
<td>Asma</td>
<td>3.7</td>
<td>3.5</td>
<td>61</td>
</tr>
<tr>
<td>Tabaquismo</td>
<td>2.0</td>
<td>3.0</td>
<td>82</td>
</tr>
<tr>
<td>Average</td>
<td>3.0</td>
<td>3.4</td>
<td>77</td>
</tr>
</tbody>
</table>

Conclusion and Lessons Learned: Similar to English, term familiarity can be used as the guiding principle to reduce difficulty of Spanish text. While encouraging, there were important lessons learned. In Spanish, increasing overall term familiarity is difficult, i.e. changing one term requires often adjusting other terms for gender and, as a result, the average term familiarity sometimes decreased. Furthermore, the lexicon has enormous impact. In prior studies, we used a smaller lexicon and were unable to simplify text. Finally, adding explanations may weaken improvements in term familiarity. Our approach requires more fine-tuning and the best version will be included in our future online text simplification tool.

Acknowledgments: Research reported in this publication was supported by the National Library Of Medicine of the National Institutes of Health under Award Number R01LM011975. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References
A Pilot Study on the Coverage of ICD and SNOMED CT of Problem List Terminologies Used in a Large Healthcare Institute in China

Yuxi. Li, MD 1, Jie. Jiang, MD1, Yan. Zhang, MD1, Jianping. Li, MD1, Yong. Huo, MD1
1Cardiology Department, Peking University First Hospital, Beijing, China

Introduction

Standardized problem list terminology (PLT) plays an important role in medical informatics while in China the only available standard terminology system is the ICD-9-CM Chinese Edition. Furthermore, most healthcare institute do not have local PLTs, besides synonyms and homonyms are very common in Chinese, which result in many different words for the same meaning during daily clinical practices. Previous studies1, 2 have shown SNOMED CT CORE Subset may be the best terminology in clinical health records. We performed this pilot study to evaluate the coverage of ICD and SNOMED CT CORE Subset and identify a subset of concepts based on standard terminologies that occur with high frequency in problem list data to facilitate the standardization of PLTs.

Methods and Results

We analyzed 165,959 structuralized problem list items without standardization from electronic health records database including 17,437 unique terms since 2011 August to 2016 July among patients discharged from the Cardiology Department of Peking University First Hospital. Figure 1 shows the percentage of unique terms (vertical axis) required to cover a certain percentage of total usage (horizontal axis), which shows the distribution of a small proportion of commonly used terms and a long tail of rarely used ones.

Firstly, we sorted all the unique terms according to the frequency. Secondly, we mapped items with exact match to ICD-CM Chinese Edition, which covers 32.61% of all terms. Finally, data scientist designed regular expressions using Python re Module for each term based on the rules provided by physicians to standardize and map terms into ICD and SNOMED CT CORE Subset, a teamwork revision was performed based on full list of distinct original items matched each expression. As shown in figure 2, even though there is no standard Chinese SNOMED CT available, compared with ICD, it covers those standard Chinese terms used in the hospital quite well (54.03% vs. 44.60%, p<0.0001).

Conclusion

SNOMED CT Core Subset is better than ICD-9-CM for covering PLTs for clinical needs in China.

References

2. Fung KW, Xu J. An exploration of the properties of the CORE problem list subset and how it facilitates the implementation of SNOMED CT. J Am Med Inform Assoc. 2015 May;22(3):649-58.
End-to-End solution to improve user experience of tranSMART: From data curation to advanced data analytics

Haiqing Li PhD, Eric Jiang, Ajay Shah PhD, MBA
Research Informatics and Systems, Center for Research Informatics and Data Science, Beckman Research Institute, City of Hope, Duarte, CA

Introduction

tranSMART is a web-based open source translational and clinical research platform that integrates OMICS data with clinical data and biological knowledge[1]. Once the data are curated and loaded into tranSMART data warehouse, tranSMART simplifies the data management and data analysis through the graphical user interfaces. However, researchers still face two barriers to fully utilize this system. First, curating external clinical and omics data for tranSMART is time-consuming. Second, users are limited to the analysis tools provided by tranSMART. In this work, we developed three systems to seamless user experience from data curation to external analytical features integration.

Method

We designed and implemented three components to improve the tranSMART user experience by breaking down the two barriers mentioned previously. Figure 1 shows our implementation architecture (the three implementations are highlighted in blue). First, we designed a guided data curation tool with graphic user interface. It can guide the user to convert the external data to tranSMART friendly format step by step. The system generates the all raw data file, mapping data file, and/or reference file in right formats with right names. And those data files are ready to be loaded into tranSMART data warehouse. Second, we implemented a customized ETL package to integrate over 115k de-identified patients data from our i2b2[2] cohort identification system with tranSMART. The patient data includes complete demographic information, ICD-9 diagnosis data, ICD-O tumor registration data, and sample status. The data is updated weekly with automatic dataaudit. Third, we designed and implemented a web services plug-in – “tranSMART Connector”, which extends tranSMART analysis capacity. Instead of re-implementation of the existing analysis features from other analytical systems, tranSMART Connector links the tranSMART with the external system. Researchers still use tranSMART’s data explorer feature to generate a cohort for analysis. Once the cohort is selected, the information about dataset for analysis is seamlessly transferred from tranSMART to external analytical system with a unique data identifier. With this solution, we successfully add machine learning analysis feature to tranSMART by link it with our machine learning analysis system, SPIRIT ML [3].

Results:

We designed and implemented a guided graphic data curation tool which simply the complex data curation procedure. This feature speedups the data collection and avoids user operation errors introduced by copy-and-past. The researcher can add the data they are interested in themselves without the assistance from data curators. The customized ETL demonstrates the integration solution for i2b2 and tranSMART. Finally, the “tranSMART connector” allow end users to hook up tranSMART with an external analytical system for advanced analysis features without re-implement them in tranSMART. We demonstrate two use cases to show the end-to-end user experience enhancement in tranSMART using these three features. The first one is to conduct a biomarker discovery study for Pediatric Brain Cancer by re-using public dataset from GEO. We retrieved the clinical information and gene expression dataset from GEO, and curate them using our guided curation tool. Then load it to tranSMART for Biomarker analysis using tranSMART advanced workflow. The scientific researchers can perform all the data management and analysis procedures by themselves. The second use case is to conduct a breast cancer recurrence factor study in tranSMART using Machine Learning features that are implemented in an external analytical system.

Conclusion

By adding these three systems, tranSMART can provide an end-to-end translation research solution for end users. It powers the scientific researchers the independence to utilize the data integration and analysis capacity of tranSMART. The Scientific researcher can have full control of tranSMART, from the sources of data to analysis tools.

References

Correlation Between End-User Experience and Clinical Effectiveness of a Clinical Decision Support Tool for the Management of Uncontrolled Hypertension

Victor S. Lin, MD, MS1, Andrew J. Holdaway, MD1, Neris M. Nieves-Robbins, MD1, Robert C. Marshall, MD, MPH, MISM1, Keith D. Solveson, MS1, Robert D. Highley, MD2, MPH, Leilani A. Siaki, PhD1
1Madigan Army Medical Center, Tacoma, WA; 2Analytics4Medicine, Kirkland, WA

Abstract
Clinical decision support (CDS) systems can greatly enhance patient care quality, but their success may depend upon positive user experience. We will present the results of an ongoing pilot study investigating the correlation between the end-user experience and the clinical outcomes of a web-based CDS system designed to aid clinicians in the management of uncontrolled hypertension.

Background
Clinical decision support (CDS) systems can greatly enhance the quality of patient care by assisting clinicians with the assimilation of clinical information and adherence to clinical practice guidelines. The successful adoption of CDS systems is generally associated with a positive end-user experience. A web-based CDS tool has been developed by Analytics4Medicine (Kirkland, WA) to aid clinicians in the management of uncontrolled hypertension. This tool analyzes and categorizes hypertension based primarily on the patients’ blood levels of renin and aldosterone. Approximately 15% of patients with hypertension have hyperaldosteronism, which only responds to aldosterone antagonist medications. These are not included as first-line agents in the current Joint National Committee (JNC 8) guidelines. Personalized recommendations for appropriate medication treatment, based on the sub-category of hypertension, are then generated by the CDS tool. This web-based tool has both provider and patient-facing portals. A funded, Institution Review Board-approved pilot study entitled Technology Assisted Management of Uncontrolled Hypertension (TEAM-HTN) is currently ongoing to assess the clinical effectiveness of this CDS tool, as well as the end-user experience with this tool.

Methodology
In this prospective within-subject pilot study, 160 patients with uncontrolled hypertension, and who are currently on three or more medications, will be enrolled along with 16 providers. Patient data, including laboratory results for renin and aldosterone levels, as well as demographical data, are entered into the CDS tool at the beginning of the study. The patients are provided a digital home blood pressure monitor to collect blood pressure readings and enter those readings into the CDS website. The providers will then manage the patients’ hypertension with the aid of the CDS tool. The rate of blood pressure control will be assessed at the end of the six-month study period. In addition, provider satisfaction with the CDS tool will be assessed using pre/post questionnaires evaluating end-user perception of feasibility, efficacy, and usability. Each questionnaire incorporates validated questions from the Technology Acceptance Model (TAM). The post-questionnaire will also incorporate questions from the validated End User Computing Satisfaction (EUCS) model. Patients will complete a post-questionnaire consisting of TAM and EUCS questions.

Results
Preliminary data obtained during CDS development showed that 85% (n=73) of patients achieved blood pressure control on at least 70% of their blood pressure readings, and on fewer medications. This is a significant improvement compared to the 2014 National Health and Nutrition Examination Survey (NHANES) results demonstrating that only 53% of hypertensive patients were controlled. Additional results from this ongoing study are expected by October, 2017. The hypertension control rates will be correlated with the end-user experience as measured by the pre/post questionnaires.

Conclusions
Uncontrolled hypertension is a significant clinical problem with devastating consequences. Personalizing hypertension management with this CDS tool may positively impact morbidity and mortality, and reduce the cost of care. End-user satisfaction with CDS systems is expected to play a significant role in achieving these outcomes.
Integrating Smart Pump Infusion Data with Electronic Health Record Data to Analyze Medication Administration Alerts

Todd Lingren, Yizhao Ni, Eric Hall, Matthew Leonard, Eric Kirkendall*, Kristin Melton*
Cincinnati Children’s Hospital Medical Center, Cincinnati, OH *Equal contribution

Introduction
Smart pumps are designed to deliver continuous medication infusions in hospital settings like the neonatal intensive care units (NICUs). Unlike standard infusion pumps, smart pumps are programmed with standardized medication libraries, rules, and alerts to facilitate administration of appropriate drug concentrations and infusion rates\(^1\). Although the pumps may prevent some potential errors\(^2\), they are still vulnerable to technological and nursing behavioral factors as well as data entry errors\(^3\). Furthermore, drug administrations are communicated wirelessly to local databases and the lack of smart pump integration with the electronic health record (EHR) hampers medication safety efforts\(^4\). As part of a larger study on NICU medication safety\(^5\), we seek to evaluate smart pump errors, perform alert statistics, quantify high rates of errors, and identify potential causes. Here we describe initial efforts to match medication order and medication administration records (MARs) in the EHR with smart pump records (SPRs).

Methods
We selected all EHR medication data and SPRs from the Cincinnati Children’s Hospital Medical Center (CCHMC) NICU in 2016. We identified medications delivered continuously including parenteral nutrition, lipid, intravenous fluids, narcotics (morphine and fentanyl), insulin, milrinone, dobutamine, vasopressin, dopamine, and epinephrine resulting in a data set of 19,614 medication orders, 154,459 MARs and 437,474 SPRs. MARs are initiated at the bedside by scanning a patient and medication barcode, but there is no similar integration for manually-programmed smart pumps. SPRs have a manually-entered patient id field, which typically captures the patient medical record number (MRN), but no EHR medication order field. We developed algorithms to link corresponding medication orders, MARs, and SPRs. First, we linked records using matching patient ids. Next, exact orders were linked with MARs and SPRs using medication name, time, and date of administration.

Results
Smart pump data contained 486 unique patient ids. Of these, 345 matched MRNs from EHR data and 31 matched encounter ids. Non-matched ids included 1) seemingly random numbers (e.g. 0, 999999999), 2) numbers matching the format of MRNs, but not matching MRNs in the EHR data set, 3) alpha characters (e.g. A, C, last names), and 4) one typo (PP was added to an MRN). MRN and encounter id-matched records accounted for 74.5% (325,909/437,474) of SPRs. Of the 325,909 records matched using patient id, 11.8% (38,544) matched exact MARs (medication name and time, within 10 minutes); a further 42.9% (140,103) matched MARs with the same medication on the same day. Only 8.3% of alert events in the SPRs were matched (2,911/34,953). Overrides to CCHMC’s pre-programmed medication library complicate matching: 95,763 (21.9%) of SPRs had medication name Unknown and 87,563 SPRs (20.0%) with Unknown as a patient id. Of these, 19,330 (4.4% overall) have both Unknown as an id and a medication name.

Conclusion
We described algorithm development to match SPRs with EHR medication information for the purpose of analyzing errors and alert events. Matching non-integrated medication administration information remains a challenge, given data entry errors and standard pump programming overrides. Further work includes physician chart review to develop more robust rules in order to refine the matching algorithms.

References
Using Machine Learning to Predict Hospital Acquired Sepsis

Xinran Liu, MD1, Greg McKelvey, MD2, James Marquardt, MS3,
Rosemary Grant, BSN, RN, CPHQ4, David Carlbom, MD5
1, 4, 5 University of Washington, Seattle, WA; 2, 3 KenSci, Seattle, WA

Problem addressed

Hospital acquired sepsis (HAS) accounts for 11-15% of all cases of sepsis. HAS is associated with higher mortality (19.2% versus 8.6%), costs ($38,400 versus $8,800), and length of stay (17 days versus 6 days), compared to community acquired sepsis1. At Harborview Medical Center (HMC), there are about 270 cases of HAS per year, with about 70 deaths/year (26% mortality). HMC has its own sepsis screening algorithm (SePSIS), which looks for trends and absolute values of temperature, heart rate, respiratory rate, white blood count, systolic blood pressure, lactate, shock index, bicarb, and fractional of inspired oxygen. While this algorithm is very sensitive (95%+), it is not very specific (~13%), which leads to a high false positive rate (low positive predictive value). There are about 2000 triggers from this algorithm per month, which leads to alert fatigue for both bedside nurses and providers.

Aim

Our goal is to create a predictive model that can identify hospitalized patients who are likely to develop sepsis. This will allow these patients to receive treatment earlier, which could decrease mortality rates as it has been shown that every 1 hour delay in starting antibiotics leads to 7.6% decrease in survival.3 Specifically, we hope to create a novel predictive model built on top of the current SePSIS screening algorithm. This model would essentially act as a “filter” to the SePSIS screening algorithm, and would be designed specifically for the HMC patient population. This should decrease the number of false positives, which will also decrease alert fatigue.

Methods

This will be a retrospective study in which clinical data extracted from the electronic health record will be used to train a machine learning model to predict hospitalized patients who are at risk for developing HAS. The population will be patients 18 or older who are admitted to HMC in a 5 year retrospective timespan. Time spent in the ICU will be excluded. Four years of data will be used to train the model, and 1 year of data will be used for validation. The model will be trained to predict sepsis based on the sepsis-3 criteria, which is suspected infection and signs of organ dysfunction. During the training phase, data elements (i.e. temperature, heart rate) for the model will be selected bases off literature review and clinical input. Data elements will also undergo feature engineering to create more novel features to be used in the model. Examples of feature engineering include monotonically decreasing/increasing values, adjustments for conditions or medications that affect vitals or labs (i.e. adjusting HR for patients on beta blockers), polynomial expansions, threshold discovery, categorical variable grouping, etc.

We propose to run our data through multiple different machine learning models such as boosting models, bagging models, neural networks, and regression based models, etc. Once training is complete, we will then validate the models on the validation data set. Our outcome metrics will be sensitivity, specificity, positive predictive value of identifying patients who develop sepsis in the validation set. We will work with the data science company KenSci to create and run machine learning models on this data and expect to have preliminary results before November of 2017. The KenSci platform has already been used to create predictive models on topics such as: surgical outcomes, inpatient deterioration, patient mortality, sepsis in the outpatient setting, readmissions, etc.

References

Number of Citations of Structured and Unstructured Abstracts in PubMed

Fang Liu, MS; Paul Fontelo, MD, MPH
National Library of Medicine, Bethesda, MD

Abstract

Structured abstracts are more informative so some may rely solely on abstracts to inform decisions. We wanted to know whether articles with structured abstracts would be cited more often than unstructured abstracts. We reviewed citations data for 11 terms in PubMed. We found that unstructured abstracts outnumber structured abstracts by more than 2:1. However, we found no significant difference in the number of citation between structured and unstructured abstract.

Introduction

Because structured abstracts are labeled with tags, readers can discover key elements in papers more easily. In recent years, around 30% of newly added publications in PubMed were structured abstracts. Our goal in this study is to find if there is a difference in the number of citations between publications with structured abstracts and those with unstructured abstracts. Although PubMed has a filter for structured abstracts, it does not provide the number of citations for journal articles but PubMed Central (PMC) does.

Methods

We selected 11 queries from the askMEDLINE database. Search algorithms were developed to retrieve and process data through E-Utilities, an API tool for PubMed and PMC. We searched PubMed with an abstract-type filter (structured abstract versus unstructured abstracts) and a publication-year filter (yearly from 2007 to 2011) for each term to allow adequate time for articles to be cited. We retrieved the PubMed ID (PMID) for each query with different abstract types yearly then searched PMC database to retrieve the number of cited-by publications in PMC.

Results

From 2007 to 2011, the total number of no-abstract, structured abstracts and unstructured abstract was 642992, 980996, and 2465968 respectively. Table 1 shows the yearly average citations for structured (S) and unstructured (U) abstracts. Figure 1 shows the average number of citations for all the 11 search terms. The average number of citations for structured and unstructured abstracts were 14.9 and 16.6 in 2007, and 9.5 and 10.5 for 2011, respectively.

Table 1. Average number of citations in PMC.

<table>
<thead>
<tr>
<th>Term</th>
<th>2007</th>
<th>2008</th>
<th>2009</th>
<th>2010</th>
<th>2011</th>
</tr>
</thead>
<tbody>
<tr>
<td>checkpoint inhibitors</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>imatinib mesylate</td>
<td>5.3</td>
<td>13.2</td>
<td>14.2</td>
<td>11.5</td>
<td>14.0</td>
</tr>
<tr>
<td>atorvastatin</td>
<td>12.9</td>
<td>7.2</td>
<td>8.5</td>
<td>8.2</td>
<td>6.1</td>
</tr>
<tr>
<td>type 2 diabetes</td>
<td>16.3</td>
<td>19.8</td>
<td>16.3</td>
<td>14.3</td>
<td>14.7</td>
</tr>
<tr>
<td>malaria</td>
<td>18.3</td>
<td>15.6</td>
<td>18.6</td>
<td>14.8</td>
<td>18.1</td>
</tr>
<tr>
<td>dengue</td>
<td>14.4</td>
<td>19.5</td>
<td>12.9</td>
<td>14.9</td>
<td>16.5</td>
</tr>
<tr>
<td>clostridium difficile</td>
<td>23.4</td>
<td>14.3</td>
<td>14.9</td>
<td>16.5</td>
<td>12.6</td>
</tr>
<tr>
<td>cholera</td>
<td>6.9</td>
<td>12.9</td>
<td>11.7</td>
<td>11.2</td>
<td>12.3</td>
</tr>
<tr>
<td>crohn's disease</td>
<td>12.4</td>
<td>19.8</td>
<td>11.9</td>
<td>15.8</td>
<td>10.2</td>
</tr>
<tr>
<td>reflux esophagitis</td>
<td>4.6</td>
<td>3.2</td>
<td>3.6</td>
<td>4.8</td>
<td>3.7</td>
</tr>
<tr>
<td>proton pump inhibitors</td>
<td>7.6</td>
<td>4.7</td>
<td>9.6</td>
<td>5.3</td>
<td>7.2</td>
</tr>
</tbody>
</table>

Conclusion

Our data does not seem to support the hypothesis that structured would be cited more often than unstructured abstract. However, we discovered that unstructured abstracts outnumber structured abstracts by a ratio of 2.5:1. Although, no difference was found in the absolute number of average citations, the almost equal average number of citations may favor structured abstracts despite the higher number of unstructured abstracts. However, it could also simply mean that structured and unstructured abstracts have equal chances of being cited in future publications.

References

Detection of Weight Data-Entry Errors

Jing Liu, MS\(^1\), Laura E Brown, PhD\(^1\), Judith W Dexheimer, PhD\(^2\), S Andrew Spooner, MD\(^2\)
\(^1\)Michigan Technological University, Houghton, MI; \(^2\)Cincinnati Children’s Hospital, Cincinnati, OH

Introduction

Errors in weight data are recorded into patient’s Electronic Health Records at a small, but not insignificant rate. For pediatrics, where weight-based dosing may be used, the importance of recognizing a weight data-entry error is critical to avoid over- or under-dosing a patient. Also, pediatric patients are at high risks for adverse drug events due to the rapid weight change over a short period, therefore, they may be more vulnerable to the effects of these errors\(^1\). In this work, we examine methods to predict a patient’s weight given their historical data. This is the first step in the aim to build an automated system to identify, alert, and mitigate weight data-entry errors in real-time practice settings.

Methods & Results

Patient weight data was collected over a three-year period Jan. 1, 2013 – Dec. 31, 2015. The analysis examines a subset of 1525 patients (643 female and 882 male), aged 0 to 36 months, who have at least 10 historical weight data points. The weight data is pre-processed using the lambda, mu, and sigma (LMS) power transformation\(^2\). The methods use the records of weights, age, and gender for prediction. For each patient, all historical data except the most recent weight is used to train the methods and tested on the last data entry. The first prediction method is a percentile classification algorithm (Perc.), where the Chebyshev distance between the training data and percentile growth curves\(^3\) classify the patient into a percentile rank whose smoothed LMS parameters predict the weight. Double exponential smoothing regression (DE) is applied so as to take both individual patients’ trend and the level into account, discounting less recent data. A non-linear least square regression model (NLS) matching expected growth curve functions is trained and tested. Lastly, due to the significant bias of outliers, a robust regression method (RR) is used. The mean absolute prediction error measuring the absolute distance between predicted and actual weight is shown below (Table 1). Overall, the double exponential smoothing has the lowest prediction error. Each method has strengths on specific data patterns, e.g., percentile classification works well when the data follows a percentile rank; double exponential smoothing works well when there is a regularly sampled, recent data available; robust regression is good at removing outliers. Large prediction error, e.g., males age 12-24 months for the NLS and RR methods, can be attributed to extreme errors in weight entry in the training data; we plan to address in future work. Some data patterns such as recent data paucity, extreme overweight/underweight patients, sawtooth patterns, etc. are challenging for all methods\(^4\).

Table 1. Overall Mean Absolute Prediction Error (kg) by Gender and Age

<table>
<thead>
<tr>
<th></th>
<th>Age (mo.)</th>
<th>Perc.</th>
<th>DE</th>
<th>NLS</th>
<th>RR</th>
<th>Age (mo.)</th>
<th>Perc.</th>
<th>DE</th>
<th>NLS</th>
<th>RR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>0-12</td>
<td>0.459</td>
<td>0.284</td>
<td>0.412</td>
<td>0.425</td>
<td>0-12</td>
<td>0.472</td>
<td>0.311</td>
<td>0.520</td>
<td>0.496</td>
</tr>
<tr>
<td></td>
<td>12-24</td>
<td>0.528</td>
<td>0.394</td>
<td>0.439</td>
<td>0.438</td>
<td>12-24</td>
<td>0.679</td>
<td>0.521</td>
<td>2.311</td>
<td>2.295</td>
</tr>
<tr>
<td></td>
<td>24-36</td>
<td>0.619</td>
<td>0.438</td>
<td>0.546</td>
<td>0.550</td>
<td>24-36</td>
<td>0.667</td>
<td>0.492</td>
<td>0.592</td>
<td>0.483</td>
</tr>
<tr>
<td></td>
<td>0-36</td>
<td>0.509</td>
<td>0.343</td>
<td>0.491</td>
<td>0.494</td>
<td>0-36</td>
<td>0.536</td>
<td>0.369</td>
<td>0.889</td>
<td>0.861</td>
</tr>
<tr>
<td>Male</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Conclusion

This work is the first step to create automated alerting system for pediatric weight monitoring. In the future, we plan to explore other variables embedded in patients’ data such as, department, diagnoses, and medications, and create ensemble models by integrating and leveraging the strengths of each model mentioned above to improve predictions.

References


2087
Building an Obesity and Cancer Semantic Web Knowledge Base

Juan Antonio Lossio-Ventura, PhD¹; William Hogan, MD, MS¹; François Modave, PhD¹; Amanda Hicks, PhD¹; Yi Guo, PhD¹; Zhe He, PhD²; Mirela Vasconcelos, BS¹; Jiang Bian, PhD¹
¹University of Florida, Gainesville, FL; ²Florida State University, Tallahassee, FL

Introduction: Obesity is associated with increased risks of cancer, as well as a wide range of other chronic diseases. On the other hand, access to health information activates patient participation and improves their health outcomes. However, existing online information related to obesity and cancer, is heterogeneous and often presented as free-text data. A formal knowledge representation and a semantic web knowledge base could help to manage and deliver quality health information. Ontologically-structured knowledge base construction can be achieved using natural language processing (NLP) and machine learning (ML) methods. Nevertheless, current ontologies describing obesity, cancer and related entities are not designed to guide automatic knowledge base construction from heterogeneous information sources. In this work, we developed a novel framework (see Figure 1) to build an obesity and cancer knowledge base from PubMed abstracts.

Methods: Our framework consists of four main steps. First, we developed an approach for named-entity recognition (NER) to extract biomedical entities from abstracts of scholarly articles. Our method leverages both linguistic and statistical characteristics of biomedical entities. Compared to existing methods, our approach can automate the knowledge base construction process with more fine-grained entities (e.g. be able to identify "type 2 diabetes" rather than simply "diabetes" from a sentence), and new entities¹ (i.e., entities that do not exist in current terminology services such as those in the Unified Medical Language System). Second, we used a similar statistic- and linguistic-based method to extract predicates from the text. Third, we developed a binary classifier to detect whether a pair of biomedical entities and a predicate form a meaningful relationship (i.e., relation detection). Fourth, we built a multi-class classification model to identify the specific relation between the pair of entities, and mapped the predicate to 12 classes in the relation ontology (RO). Overall, our framework is capable of extracting semantic triple statements (i.e., the subject-predicate-object expressions, such as “obesity is a risk factor of cancer”) from free-text PubMed abstracts.

Results: We evaluated the four steps in the framework separately. First, we used two datasets to evaluate the proposed NER approach: (1) a set of 214 manually annotated sentences extracted from 23 PubMed abstracts related to obesity and cancer; and (2) the well-known NCBI disease corpus. We compared our method with several state-of-the-art biomedical NER applications, including MetaMap and BioPortal annotator. Our method achieved the best performance with an F-measure of 90.1%. Second, we compared the performance of our predicate extraction method to a state-of-the-art application, Reverb, obtaining a better precision (62.5%). Third, our relation detection method obtained an F-measure of 98.6%. Finally, the multi-class classification model achieved an F-measure of 85.3% when we classified the specific relations to RO classes. The two classification models are built with 51 statistical, lexical and semantic features. We also experimented with a wide range of well-known supervised learning algorithms in these classification tasks, and found that the Random Forests model performs the best for both tasks.

Conclusions: We have developed a framework for building an obesity and cancer knowledge base. The steps in the proposed framework are important building blocks towards automating construction of an ontologically-structured knowledge base from large, free-text datasets. We presented an improved biomedical NER model and a predicate recognition model incorporating both linguistic and statistical characteristics of the text. Moreover, we used two supervised learning methods to: (1) predict if two biomedical entities and a predicate in a sentence are related, and then (2) classify the kind of relationship between them. Our initial testing of the proposed framework on a small, manually-curated gold-standard dataset shows promising results.

References
Characterization of Information Collected from Decision Support Request Forms in Academic Medical Centers

Jean F. Louis¹, MIM, Guilherme Del Fiol¹, MD, PhD, Rachel L. Richesson², PhD
¹Biomedical Informatics, University of Utah, Salt Lake City, UT, ²Duke University School of Nursing, Durham, NC

Introduction
In academic medical centers, requests for new Clinical Decision Support (CDS) tools can involve a lengthy and sometimes inefficient process of soliciting functional and data requirements from clinicians. Characterizing the type of information that different organizations collect in their CDS request forms can help in developing a more effective and productive CDS request system. We analyzed the information collected from the CDS request forms of three academic medical centers in order to create a set of categories of information collected in all three forms.

Methods
Three informatics experts (the authors) independently created categories for each question in two of the three CDS request forms. Next, discrepancies were reconciled through consensus, yielding a master set of categories along with definitions. Last, the three experts assigned the categories to the questions on the third form. Collectively, the experts have extensive experience in CDS, terminologies and standards, public health, and information management.

Results
We evaluated 78 questions from the three CDS request forms (34, 17, and 27 questions for forms 1, 2, and 3 respectively), which were assigned to 11 categories. The table lists the categories with definitions along with the number of questions per category in each of the three forms.

<table>
<thead>
<tr>
<th>Information Categories</th>
<th>Definition</th>
<th>F1</th>
<th>F2</th>
<th>F3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Requester</td>
<td>Information about the individual or group who submitted the CDS request.</td>
<td>1</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>Request Metadata</td>
<td>Information about the request used by the decision support committee to keep track of the status and communication related to the request.</td>
<td>0</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>Rationale</td>
<td>Reasons justifying the benefit or need for the CDS.</td>
<td>2</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>Supporter</td>
<td>Individuals or groups that have endorsed the proposed CDS and/or will support its implementation.</td>
<td>1</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Target audience</td>
<td>Intended care settings, users or recipients of the CDS intervention.</td>
<td>1</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Target patient population</td>
<td>The broad population of patients who are intended to benefit from the CDS implementation.</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>CDS Logic</td>
<td>User actions, patient conditions, or events that should trigger the execution of the CDS intervention.</td>
<td>7</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>CDS behavior / workflow</td>
<td>Considerations regarding the behavior, content, and user interface as well as integration with the clinical workflow.</td>
<td>9</td>
<td>3</td>
<td>8</td>
</tr>
<tr>
<td>Effort</td>
<td>Estimate of the resources required to complete and implement the CDS request.</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Implementation</td>
<td>Considerations regarding the CDS implementation, such as need for training, awareness and buy-in of CDS recommendations among target audience.</td>
<td>4</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Evaluation and monitoring</td>
<td>Criteria, measures, procedures, or activities that are available or planned and that will be needed to evaluate and/or continuously monitor the effect of the CDS intervention.</td>
<td>6</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Discussion
As organizations update their CDS request forms, the collection of information from these categories can improve how organizations prioritize CDS requests and define requirements. Our research enhances our knowledge of the CDS request process while working toward a standard platform to submit CDS requests that can be shared and reused across organizations. This project will help streamline communication between experts and CDS requesters. It may reduce time spent in CDS meetings by soliciting CDS functional and data requirements from requesters before meeting with CDS committees. Ultimately, we plan to implement a logic-based CDS request form in REDCap to be used across institutions and evaluate additional institutions as part of our future work.

Acknowledgement: Supported by grants T15LM007124-20 and R15 LM012335-01A1 from the National Library of Medicine.
Enhancing LexSynonym Features in the Lexical Tools

Chris J. Lu, PhD1,2, Destinee Tormey1,2, Lynn McCreedy, PhD1,2 and Allen C. Browne1
1National Library of Medicine, Bethesda, MD 2Medical Science & Computing, LLC, Rockville, MD

Introduction

Concept mapping is vital to natural language processing (NLP) for bioinformatics. Query expansion using synonyms for subterm substitutions is an effective technique to increase recall when no direct concept mapping can be found through normalization. For example, no concept can be found by direct mapping through normalization if the source vocabulary is “calcaneal fracture”. By substituting the subterm “calcaneal” for its synonym, “heel bone”, the matched UMLS concept [C0281926, fracture of calcaneus] is found. In this example, “calcaneal” and “heel bone” are element synonyms while “heel bone fracture” is the expanded term. The effectiveness of this technique relies on the completeness and quality of both the element synonyms and the UMLS synonym thesaurus in the query expansion pipeline of UMLS concept mapping (Figure 1).

Figure 1. Query Expansion Pipeline of UMLS Metathesaurus Concept Mapping

Enhanced LexSynonym Features in the Lexical Tools

A systematic approach to LexSynonym acquisition from the UMLS Metathesaurus and the Lexicon was developed to meet two necessary properties, commutativity and transitivity, for quality element synonyms [1]. LexSynonyms are synonymous (multi)words in the Lexicon, such as “calcaneal” and “heel bone” in the above example. 190,844 LexSynonyms are acquired in the Lexicon.2017. They are integrated in the Lexical Tools (2017) for synonym retrieval. The synonym flow component (-f:y) is enhanced to include synonyms, POS and source information. A sophisticated algorithm is implemented in the recursive synonym flow component (-f:r) to preserve precision based on the source types. In addition, the synonym source option (-ks) is added to restrict the results by source type (CUI, EUI, NLP), or any combination of the above. These new features provide needed information for downstream NLP processing. For example, the size of the fruitful variants flow component (-f:v, using recursive synonym features), that is used in various NLP projects (Sophia, MMTx, Custom Taxonomy Builder) for better recall, has been increased from 8.9M (2016) to 9.9M (2017) due to the enhancement of LexSynonyms.

Test, Results and Conclusion

The UMLS-CORE project assigned CUI(s) to terms (13,076) that are within the top 95% usage and mappable to SNOMED CT [2]. 2,755 of these terms (with 2,756 CUIs) that do not have direct mapped concepts in UMLS.2016AB are used to test the performance of LexSynonyms for query expansion. Three different synonym sets are tested through the Sub-Term Mapping Tools (STMT) [3]. The results show improvements in recall (4.97%) and F1 (0.05) from both the 2016 and 2017 LexSynonyms with the STMT synonym set (Table 1) [1]. Precision is slightly increased (0.26%) due to the high quality of LexSynonym 2017 (71.04% precision in this test). The set of LexSynonyms and enhanced features are distributed in the Lexical Tools.2017 with UMLS by NLM via an Open Source License agreement. Improvements in performance can be anticipated for NLP applications that use these enhanced features of LexSynonyms.

Table 1. Test Result for LexSynonyms 2016-17

<table>
<thead>
<tr>
<th>Synonym Set</th>
<th>TP</th>
<th>FP</th>
<th>Rel.</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>STMT + 2016</td>
<td>691</td>
<td>358</td>
<td>2,756</td>
<td>65.87%</td>
<td>25.07%</td>
<td>0.3632</td>
</tr>
<tr>
<td>STMT + 2017</td>
<td>828</td>
<td>424</td>
<td>2,756</td>
<td>66.13%</td>
<td>30.04%</td>
<td>0.4132</td>
</tr>
<tr>
<td>2017</td>
<td>287</td>
<td>117</td>
<td>2,756</td>
<td>71.04%</td>
<td>10.41%</td>
<td>0.1816</td>
</tr>
</tbody>
</table>

*T: True, F: False, P: Positive, Rel.: Relevant

References

2. KW Fung and J Xu, An exploration of the properties of the CORE problem list subset and how it facilitates the implementation of SNOMED CT, JAMIA 2015, 22: 649-658.
Current Practices in Integrating Social and Behavioral Data into Patient Care at a VAMC

Cherie Luckhurst, PhD, MS¹, Michael Weiner, MD, MPH¹-³

¹VA HSR&D Center for Health Information and Communication, Indianapolis, IN; ²Indiana University School of Medicine, Indianapolis, IN; ³Regenstrief Institute Inc., Indianapolis, IN

Study Rationale
Research conducted over the last several decades has firmly established that social and behavioral (S&B) factors influence physical health. Recently, the Institute of Medicine recommended that such data be collected from every patient, and it proposed that S&B data will aid in recognizing, diagnosing, and treating medical problems¹. However, some VA care providers do not see the usefulness of such data in their clinical practices. As a result, these data are rarely collected and, when collected, they are used inconsistently²; a mechanism has not yet been developed to integrate such data into clinical workflow and decision-making³.⁴ Before developing such a mechanism, however, the principles of informatics direct us to examine current practices.

Design
This study uses mixed methods to understand current use of S&B data in the diagnosis and treatment of medical issues at the Roudebush VAMC. Personnel in many VHA clinics, including primary care, social work, and mental health/psychology will be recruited to participate. We will interview 24 clinicians who serve in a variety of roles (e.g., primary care providers, nurse practitioners, care-coordinators, technicians). We will also administer a short survey that includes demographic and practice-focused questions that explore a care provider’s perspective on the usefulness of S&B data in his or her medical practice. Quantitative data will be analyzed using descriptive statistics; qualitative interviews will be analyzed inductively based on the grounded theory approach.

Timeline
Interviews of 24 clinicians are being conducted in March and April of 2017. Thematic analysis will be conducted in May and June 2017, with poster preparation to follow.

Outcome
The work described here is the first step toward development of new procedures to collect and use S&B data across services and departments in a VA medical center. We hope, through a multi-disciplinary collaboration, to fulfill the IOM’s promise of improved diagnosis and treatment of medical problems by incorporating psychosocial data into clinical practice, thereby bridging the current gap between research and practice.

References
Computer Vision-based Approach to Maintain Independent Living for Seniors

1Zelun Luo, 1Alisha Rege, 1Guido Pusiol, 2Arnold Milstein, 1Li Fei-Fei, 2N. Lance Downing
1Stanford Vision Lab, Stanford University, 2Department of Medicine, Stanford University

Introduction

As the first baby-boomers begin to reach retirement age, the U.S. has begun to experience a shift in the age demographics of its population that has significant implications for Medicare spending and the federal budget. Thus far, computer vision has not had significant application in healthcare but holds significant potential for automation and augmentation in the monitoring and recognition of health and healthcare behaviors1. In this work, we introduce a computer vision-based system to allow seniors to live independently at home. Depth and thermal sensors are used as they hold advantages over wearable sensors and RGB cameras: they do not require daily wearing, provide richer information, avoid personal identification, and operate in different lighting conditions. Our prototypes are lightweight and easy to install, requiring only power source and internet access. We introduce a new open-source dataset: Thermset2, which contains 214 hours of thermal video collected at senior bedrooms. A convolutional neural network (CNN) is used to detect five clinical states: background, people present, standing, sitting, and sleeping. These states are important on their own and necessary to identify more complex activities that are more likely to be actionable by caregivers.

Methods

We trained a CNN-based feed forward network on Thermset (Figure 2) to detect whether an activity occurs in a video frame. Our network is inspired in AlexNet3, and it consists of five convolutional layers. Each layer is followed by a ReLU nonlinearity layer, a batch-normalization layer, and a dropout layer. The output is the multi-label classification of the target activities, obtained with a softmax function. At training stage, 21,960 frames (4,392 from each category) were used. For evaluation, we do not include any frame of a scene that appears in the training set.

Results

Figure 1 shows the confusion matrix for the algorithm. The matrix shows that an average 76% correct classification rate on the five aforementioned clinical states. We were only recently able to begin collecting data from depth sensors so do not have results to present here. As we add data diversity by collecting more data and using depth data, we expect a much more robust system in the future.

Conclusion

This work shows the viability of using thermal signals for monitoring elderly citizen, and can potentially allow for them to receive the care that they require from the comfort of home. In our future work, we plan to train a model that recognizes a wider range of clinical states, including day and night reversals, eating, falls, slowed movements, unstable transfers, front door loitering, chair and bed immobility, urinary frequency, restlessness, and fever. By detecting these clinical states, we aim to identify elder patients who are at risk for requiring long-term care, and to provide feedback to caregivers that would support their safe and independent living.

References

Meaningful Use Stage 3 Safety-Enhanced Design: Third Time’s a Charm?
New Certification Criteria, Insights and Lessons Learned

Andrea Mahnke MS¹, Gabrielle Falkey BS², Elise Ewers BS²
¹ Marshfield Clinic Research Foundation, Marshfield, WI
² Marshfield Clinic Information Services, Marshfield, WI

Introduction

The Office of the National Coordinator (ONC) for health information technology (HIT) is driving implementation of electronic health record systems with advance decision support in order to improve and monitor care quality and increase patient safety¹. The health care environment is dynamic and complex and therefore HIT can be successful only if systems are designed taking human technology interaction in mind and integrating them well into clinical workflows. The ONC has again required EHR developers to conduct summative usability evaluations and submit results as part of Stage 3 certification. The 2015 edition (170.315) for Stage 3 certification includes the same test criteria as the 2014 edition, but also introduces several new test criteria including additional CPOE requirements, demographics, problem list and implantable device list. Test results of all systems will be made public on the ONC website. This will allow both consumers and developers to review and assess the usability of systems to help steer purchasing and design revisions to other systems. Experience from two previous rounds of certification helps to prepare us for Stage 3. We discuss methods, results, lessons learned and proposed changes based on findings.

Methods

MCIS Clinicals V6.2, the Marshfield Clinic home-grown EHR, was used for the summative validation testing and resided on dedicated testing environment. Physicians, physician assistants, nurses and medical assistants were recruited as they represent user groups of the EHR. Recruited participants had a mix of backgrounds and demographic characteristics. Participants were scheduled for 90 minute sessions and all completed the same set of tasks. Two non-patient seeing staff was recruited to perform two administrative tasks. A number of tasks were constructed that would be realistic and representative of the kinds of activities a user might do with this EHR. Tasks matched requirements for ambulatory certification as identified by the Test Procedure for §170.315(g)(3) Safety-enhanced design certification criteria. Measures assessed system effectiveness, efficiency and satisfaction. They also served to identify usability issues that, if left unaddressed, could result in errors that could potentially lead to patient harm².

Results & Conclusion

Pending data collection and analysis scheduled for summer 2017. Reported results will include: task success, task failures, task deviations, task time, task difficulty rating, System Usability Scale score and, usability issues and errors that could potentially lead to patient harm.

References

Repurposing Software Built for Large Clinical Trials for Use in Small Trials
Nilla Majahalme, BSE1; Stephen J. Miller, MPH1; Andrew Zimolzak, MD, MMSc1,2; Jason L. Vassy, MD, MPH, SM1,3
1VA Boston Healthcare System, Boston, MA; 2Boston University School of Medicine, Boston, MA; 3Harvard Medical School, Boston, MA

Statement of Problem
Point-of-care (POC) clinical trials, which can enable a learning healthcare system, rely on the integration of electronic health record (EHR) data with other study data. Data must be presented to the study team in a usable manner, especially when the data drive study decision logic for patient enrollment, outcome monitoring, or safety events. This integration is often resource-intensive and not readily available for small, budget-constrained studies. We recently launched a POC randomized trial called the I-PICC (Integrating Pharmacogenetics in Clinical Care) Study. The pharmacogenetic association between the SLCO1B1 gene and simvastatin myopathy is well validated, but the clinical value of its use in patient care is unknown. The I-PICC Study will enroll 400 statin-naive patients with elevated cardiovascular disease risk and determine whether clinical use of SLCO1B1 testing improves patient outcomes. The limited budget of the I-PICC Study does not support extensive custom informatics development.

System Purpose
We leveraged existing components originally built for the 13,500-enrollee Diuretic Comparison Project (DCP) to support the I-PICC Study. DCP is supported by full back-end automation featuring web services to process data on demand and scheduled execution of complex jobs. For the resource-limited I-PICC Study, we used the same front-end components (ProjectFlow and workflows) but created a simpler back-end using SQL stored procedures to consume EHR data and manage decision logic. ProjectFlow is an open-source workflow application originally developed in collaboration between the Massachusetts Veterans Epidemiology Research and Information Center and the VA Informatics and Computing Infrastructure. The browser-based application connects to a database server, enabling but not requiring real-time access to data from the EHR and other sources. Data displayed in ProjectFlow are decoupled from the workflow itself, allowing not only the reuse of workflow definitions but also full customization of data displayed to the end user. The application tracks workflow sequences, prompting users to complete a next step when available. Workflows are represented by XML files built using a third-party Business Process Model and Notation (BPMN) 2.0 workflow editor. ProjectFlow applies a user interface (UI) to the workflow and uses a BPMN 2.0 engine to track progress of “clinical elements” (e.g., patients) in workflows. Data are populated into ProjectFlow through a database view, with one view driving each type of clinical element. The view is queried on each page load, giving the user a real-time view of the data. Over time, a view can be modified, with the constraints of maintaining a unique ID for each row and never removing existing data fields. The UI can be modified to display only relevant data.

ProjectFlow stores user-input data in the database in nearly human-readable form, enabling easy data retrieval from the system by any SQL programmer. For the I-PICC Study, we built a simple SQL back-end by defining 11 stored procedures to retrieve and manipulate data entered into ProjectFlow. The stored procedures can be invoked either manually or by scheduled execution on the server. We implemented a state machine of patient progress through the study by testing for conditions found in the EHR and ProjectFlow data. Patients meeting certain criteria (for example, consent call complete and an eligible lab found in the EHR) were allowed to proceed to the next state. ProjectFlow was configured to invoke a specific workflow for each patient in a particular state. Thus, new workflows are invoked and displayed to study personnel as patients progress through the study.

Seven workflows were built for I-PICC: 1) recording provider consent, 2) tracking patient mailings and consent, 3) asking providers to order the genetic test, 4) reminding providers about orders if necessary, 5) sending the lab specimen, 6) reporting genetic results, and 7) withdrawing a patient from the study. Development effort was ~3 days to build workflows using a 3rd party open source BPMN editor and ~1 month to create SQL stored procedures.

In summary, we have reused “trialware” components from a large funded study and implemented a simpler back-end to automate a state machine and decision logic for a resource-limited trial, simulating real-time data updating. ProjectFlow is being applied to other projects including the VA Precision Oncology Project. We demonstrated that if trialware is built in a modular fashion, it can be scaled for use in both small and large trials.
Optimizing Parameters of word2vec for a Text Classification Task

Vincent Major, ME1, Alisa Surkis, PhD2, Yindalon Aphinyanaphongs, MD PhD1
1Department of Population Health, NYU School of Medicine, New York, USA;
2Health Sciences Library, NYU School of Medicine, New York, USA.

Introduction

Machine learning text classification models often rely on empirical parameter optimization to identify classifier settings that are highly performant (for example, support vector machines optimize the C parameter). word2vec1,2, a recent method that uses neural networks to build word embeddings, has several key parameters. However, prior works select word2vec parameters against an internal measure of semantic similarity rather than an external measure of utility, such as performance in a classification task. Word embeddings are vectorized representations of words such that similar meaning words such as “strong” and “powerful” are in the same general Euclidian space. Representing words as vectors allows text classification models to learn concepts, rather than simply word occurrences. In theory, using embeddings results in improved model generalizability. In practice, incorporating word embeddings have demonstrated state of the art performances. In this poster, we optimize the utility of word embeddings to classify MEDLINE articles into their translational stage of T0, T1/T2, or T3/T43. One metric for success of a translational research grant is how quickly a concept or method develops from basic science through animal and human studies and into clinical practice. A model that could classify translational stage could be used to map research trajectories and evaluate return on investment.

Methods

Word embeddings were constructed using titles and abstracts from 10.5 million MEDLINE articles published between January 2000 and December 2016. The following word2vec parameters were optimized in a naïve grid search: model = {continuous bag-of-words, skip-gram}, optimization = {hierarchical softmax, negative sampling}, dimension = {50, 100, 150, 200, 300}, and context window lengths = {4, 5, 6, 8, 10}. These word embeddings are then used with a state of the art classification algorithm called fastText4 to learn and predict on a manually labeled dataset (updated from our prior work3) consisting of 542 manually labeled MEDLINE abstracts classified into 3 categories: T0 (n=281), T1/T2 (n=109), and T3/T4 (n=152). We assess predictive performance using the area under the receiver operating characteristic (AUC) in 5-fold cross validation.

Results

The baseline performance from our prior work3 (using a support vector machine) in T0, T1/T2, and T3/T3 classes are: AUC(%) [range] = 94 [93-96], 84 [76-98], and 92 [90-94]. Many of the 100 parameter combinations result in AUCs comparable to or marginally greater than the baseline.

Discussion

The comparable performance to baseline suggests that learning text semantics in a large dataset results in models that are comparable but likely more generalizable. Across the 3 classes, the skip-gram model outperformed the continuous bag of words model, as previously reported5,6, with hierarchical sampling marginally outperforming negative sampling. The number of dimensions and context window length do affect classification performance however; the effect size is small everywhere except the extremes. These data suggest that expensive optimization may not be necessary for some tasks. However, for performance critical tasks the expense may be warranted.

References

Guideline Based Recommended Therapies in PAD Patients

Identified by NLP: A Quality Assessment Project

Vishnu Priya Mallipeddi, MBBS, Naveed Afzal, PhD, Iftikhar J. Kullo, MD, Hongfang Liu, PhD, Rajeev Chaudhry, MBBS, MPH, Adelaide Arruda-Olson, MD, PhD

Mayo Clinic, Rochester, MN, USA

Introduction
Patients with peripheral artery disease (PAD) constitute a high-risk population. Evidence-based management strategies for secondary prevention have been recommended by PAD clinical practice guidelines. However, prior studies have demonstrated a substantial gap between guideline recommendations and clinical practice with low proportion of PAD patients using these strategies. We previously developed a natural language processing (NLP) algorithm for identification of PAD cases from narrative clinical notes, which had a higher accuracy than billing code algorithms. We applied this NLP algorithm for PAD case identification to conduct a quality assessment project to evaluate the use of 4 distinct guideline-recommended risk modification strategies in PAD patients from the community.

Methods
We applied a NLP based PAD algorithm to a cohort of patients evaluated at the Employee Health Clinic, Mayo Clinic, Rochester. For validation, we compared the NLP results with the gold standard manual abstraction of the medical records by two abstractors. Baseline demographic characteristics and use of each one of the four guideline-recommended strategies within 6 months of PAD diagnosis were also manually abstracted from EHRs. These strategies were: abstinence from smoking, use of moderate or high intensity statin therapy, use of aspirin and use of angiotensin-converting inhibitors (ACEI) or Angiotensin Receptor Blockers (ARB).

Results
There were 73 patients (26% women, mean age 72.87 ± 11.4 years). The NLP algorithm had positive predictive value (PPV) of 97% and sensitivity 99% for PAD case identification, compared with the gold standard. Guideline recommended therapy within 6 months of the PAD diagnosis included the use of Aspirin in 63 (86%), statin medications in 47 (64%), ACEI or ARBs in 35 (48%) and abstinence from smoking in 49 (67%) patients. The figure shows distribution of combined use of guideline recommended strategies for PAD risk modification.

Conclusions
In this pilot study, an NLP algorithm applied to clinical notes had excellent PPV and may be used to support quality assessment projects. The use of guideline recommended therapies in PAD patients remains limited. Our results indicate a need to develop automated clinical decision support tool to facilitate further implementation of guideline-recommended risk modification strategies for PAD patients. As a next step, we are planning to conduct a provider survey to assess potential need for clinical decision tool to facilitated discussion of guideline-recommended risk modification strategies in patients with PAD at the point of care.


Jason V. Mancuso, BS¹, Peter Short ¹, Edmund Reineks, MD ², Elena A. Manilich, PhD¹

¹John Carroll University, University Heights, Ohio, USA; ²Cleveland Clinic Foundation, Cleveland, Ohio, USA

Abstract

Automated laboratories analyze thousands of blood samples every day. Some samples may become contaminated due to improper collection, producing incorrect results. Identification of such samples is critical but often requires a labor intensive expert review. We present a set of machine learning algorithms that efficiently analyzes Big Laboratory Data in Hadoop and Apache Spark. We compared performance of each model using 13,945 blood samples. Our best model, Random Forest, achieves 86.5% sensitivity and 99.9% specificity.

Background: The study objective was to evaluate the accuracy, validity, and clinical usefulness of detecting laboratory errors with machine learning algorithms in Hadoop distributed computing systems. We present a machine learning approach that can be integrated into a high-throughput automated laboratory system.

Materials and Methods: Test results were extracted from an automated laboratory system for 13,945 blood samples, of which 194 were contaminated. A rule-based classifier and the clinical judgment of physicians were used to identify contaminated samples. A basic metabolic panel with eight commonly measured analytes was incorporated into our models. We compared performance of four algorithms: a decision tree, a random forest, an elastic net logistic regression, and an artificial neural network. We split each dataset into training, validation and test sets (60-20-20). All models were tuned with Bayesian optimization, selecting hyper parameter settings with the highest AUC for the validation set. We also used a larger synthetic data set (2,013,945 samples) to demonstrate scalability. These samples were synthesized by a generative model, specifically a Gaussian mixture model with 7318 components. All experiments were performed on a four-node 2.2 GHz quad-core Hadoop cluster.

Results: Each model included eight analytes from a basic metabolic panel: blood urea nitrogen, calcium, chloride, bicarbonate, creatinine, glucose, potassium, and sodium. When validating the performance of four models on the test data set, specificity was above 99% across all models. Our final Random Forest models showed best discrimination, with areas under the receiver-operating-characteristic curve of more than 0.90. Random Forest model showed the highest sensitivity of 86.4%. Comparing the experimental results with several state-of-the-art methods, such as rule-based and decision tree methods, revealed that the proposed method outperforms previous methods.

Conclusions: Machine learning models based on a basic metabolic panel can be used to accurately detect contaminated samples and assist in clinical interpretation of results. Machine learning algorithms combined with clinical judgment could be applied in daily practice. Our work is the first to apply machine learning at this scale for detecting contaminated cases.

Figure 1. Performance of machine learning algorithms, receiver-operating-characteristics curves, for 2,789 samples in the test dataset (A) and for 2,2013,945 synthetic data points (B).
Implementing Data Event Triggers with Standard-based Document Exchange to Improve Timely Communication for Referring Providers

Darren K. Mann, BS¹, Shan He, PhD¹, Jennifer Nelson, RN¹, Gary Anderson¹, Sidney N. Thornton, PhD¹,²
¹Intermountain Healthcare, Salt Lake City, UT; ²University of Utah, Salt Lake City, UT

Introduction
Communication between primary care physicians (PCPs) and specialists for referrals and consultations is often incomplete due to unreliable communication methods and time constraints. Referring providers need to receive consultation reports and test results back from specialists in order to provide optimal coordinated care. However, the common practice for sending results to referring providers is through phone, fax or secure email, which is resource-intensive and error-prone. A study reported that 80% of specialists said they “always” or “most of the time” send consultation results to the referring PCP, but only 62% of PCPs said they received such information. At Intermountain Healthcare we repurposed an automated electronic data delivery system to link referrals to the creation and transmission of an electrophysiology report from the Heart Institute to the referring providers.

Methods
To enable automated electrophysiology report delivery, we implemented clinical event based triggers in our Electronic Health Record (EHR) system. The trigger was fired when an electrophysiology report was finalized and stored to the EHR by a heart rhythm specialist. We implemented an HL7 Consolidated Clinical Document Architecture (C-CDA) document generator that embedded the electrophysiology report in the Assessment and Plan section of the progress note template after the trigger was fired. The C-CDA document was then automatically routed to the referring provider via the Direct protocol. A dashboard report from the transmission audit records facilitates the measurement of timely and complete communication of the referrals.

Results
Using trigger-based document transmissions, the Heart Institute sends finalized electrophysiology reports immediately to referring providers without any manual processing. Previously, clinical staff spent 4 hours per day facilitating the transmissions. Automating this process also increased the percentage of results that were successfully sent back to the referring providers. The Direct receipt acknowledgment increased the reliability of the consultation results arriving at the referring provider.

Discussion
The advantage of using Direct to exchange documents among healthcare providers is the ease of configuring the providers’ Direct email address. However, without a national provider directory each organization has to collect and maintain the Direct email addresses for its Direct exchange partners. As a result, the effectiveness of our implemented solution for the Heart Institute relies on the information of the referring provider collected during registration and their Direct email address existing in the local provider directory. For future work, we plan to accommodate other industry standard delivery methods preferred by referring providers such as Cross-Enterprise Document Reliable Interchange (XDR) and FHIR. Currently we embed the free-text electrophysiology report in the C-CDA document with discrete patient demographic information. We plan to improve the semantic interoperability of our document when additional standard-based exchange specifications become available that support additional structured and coded data elements for an electrophysiology report.

Conclusion
Using automated and standard-based document exchange improves the efficiency, effectiveness, and security of the communication between referring providers and specialists. Our implementation for the Intermountain Heart Institute can be generalized to other clinical use cases that require timely and reliable patient data exchange.

References
Evaluation of the Surveillance, Epidemiology, and End Results Data Management System (SEER*DMS) to support efforts in enhancing U.S cancer surveillance

Marina Z. Matatova, MA1, Paul A. Fearn, PhD1, Eric B. Durbin, DrPH2, Gary M. Levin3, 
Frank J. Manion, MS4, Patrick M. Mergler, MBA5

1National Cancer Institute, Rockville, MD; 2Kentucky Cancer Registry, University of Kentucky, Lexington, KY; 3University of Miami, Miami, FL; 4University of Michigan Comprehensive Cancer Center, Ann Arbor, MI; 5University Hospitals Seidman/Case Comprehensive Cancer Center, Cleveland, OH

Introduction
The Surveillance, Epidemiology, and End Results (SEER) Program at the National Cancer Institute (NCI) is an authoritative source of information on cancer incidence and survival in the United States representing approximately 30 percent of the US population. To improve on patient outcomes SEER is actively employing innovative approaches to augment and automate data acquisition and processing, such as natural language processing (NLP) techniques for data extraction. The SEER Data Management System (SEER*DMS) has been utilized since 2005 by central cancer registries to process and submit cancer data to the SEER program. To ensure that the system is optimally positioned to handle and is configured to seamlessly integrate machine-learning technologies for information processing. SEER contracted with health informatics experts (ED, PF, FM, PM) to methodically evaluate SEER*DMS.

Methods
The evaluation process began with an initial meeting with SEER program leadership to identify goals and objectives. After initial assessment, the health informatics committee met with staff from Information Management Systems (IMS) Inc., company contracted to design and develop SEER*DMS, in order to obtain an overview of the system. The committee proceeded to collect field observation data by visiting four SEER registries including, Iowa, Louisiana, Georgia and Seattle. The registry visits were apportioned among the committee members and yielded insights into central registry characteristics, operations, and registry user feedback of SEER*DMS. The findings were compiled by the committee members, analyzed, and organized into a comprehensive report. The team extracted nine key recommended areas of improvements and delivered a final summary report intended to support current and future goals of the SEER program. The program leadership reviewed each recommendation with IMS in a joint effort to align understanding of the proposed enhancements to SEER*DMS.

Results
The summary report provided nine key areas recommended for enhancement: (1) increasing automation of linking multiple tumor case records related to the same cancer diagnosis, (2) consolidating data elements from different records to create a complete summary of patient and tumor information, (3) developing an integrated web portal to support submission files from reporting facilities to the SEER registry, (4) developing an integrated report writer with functional and user friendly features that would support registry operations and research needs, (5) developing and supporting a natural language processing infrastructure within SEER*DMS to augment and enhance data extraction, (6) performing a system-wide usability assessment of SEER*DMS through a cognitive approach aimed to improve functional and ergonomic needs of the registry user community, (7) a redesign in the change control board for SEER*DMS and to improve communication between NCI, IMS, and the SEER registries, (8) improvements and updates to system architecture, performance, security, and sustainability to ensure SEER*DMS employs the latest technologies and methodologies to support the scale of nationwide operations, and (9) developing a strategic roadmap to prepare SEER*DMS for future trends in molecular/precision medicine which will produce a greater volume and variety of health information.

Conclusion
A systematic evaluation of SEER*DMS by health informatics experts yielded key areas of enhancements to the system that would support current program needs, central registry operations, and position the system to seamlessly integrate innovative approaches, such as natural language processing, to modernize U.S cancer surveillance with the aim of enabling current and future research to improve patient outcomes.
A Data-driven Framework for Sub-Typing Stem Cell Transplant Recipients at Risk for Infection

Anoop Mayampurath PhD $^{1,2}$, Jonathan Matthews ME$^3$, Samuel L Volchenboum MD, PhD, MS$^{2,4}$ and L. Nelson Sanchez-Pinto MD, MBI$^4$

$^1$ Computation Institute, $^2$ Center for Research Informatics, $^3$ Pritzker School of Medicine, $^4$ Department of Pediatrics, The University of Chicago, Chicago IL 60637

**Introduction**

Advances in hematopoietic stem cell transplant (HSCT) treatment have significantly improved outcomes for diseases that were often fatal. However, infection remains one of the top causes of mortality in HSCT recipients, especially in the early post-HSCT phase.$^{1,2}$ Because early identification of life-threatening infections can improve outcomes, identification of HSCT recipients at high risk for severe infection *prior to transplant* and consequent increased surveillance after transplant could be lifesaving. In this paper, we devise a computational paradigm that utilizes both unsupervised clustering and supervised-classification techniques to define different sub-types of HSCT recipients that are prone to different risk of infection due to pre-HSCT factors.

**Methods**

*Setting and study population:* We conducted a retrospective study of all allogeneic HSTC recipients from 2009-2015 at the University of Chicago Medicine. *Outcome and features of interest:* The primary outcome of interest was bacteremia, as identified by a positive blood culture within the first 100 days of HSCT. The features of interest are from a three-month pre-HSCT time period and included variables related to patient demographics, diagnosis, HSCT/donor factors, prior inpatient stay, prior infection and antibiotic use, and routine laboratory test results prior to transplant. *Analysis:* In order to detect sub-types of patients with differential risk *at time of HSCT for infection post-HSCT*, we built a computational framework using three successive steps. First, we performed hierarchical clustering on all pre-HSCT features using the Gower distance (since the data is of mixed-type) to detect patient groupings in the population in an unsupervised manner. Next, we compared rates of the primary outcome across groups and used chi-squared testing to determine if any observable difference was significant ($p < 0.001$). This allows us to detect sub-types of patients with differential rates of post-HSCT infection. The last step is to utilize classification and regression tree (CART) models to determine the variables that primarily drive the sub-type differentiation.

**Results**

Out of 425 patients that received allogeneic HSCT, 151 (35%) developed bacteremia infection. Hierarchical clustering of pre-HSCT variables in the allogeneic population revealed four distinct groups of patients (Figure 1), of which one group showed higher rates of infection compared to other groups (Table 1, $p < 0.001$). The direction and strength of association between each pre-HSCT feature and the grouping is currently being studied.

**Conclusion**

Our novel computational framework detected sub-types of patients with different risk of infection based on characteristics prior to transplant. This allows for patient risk stratification at transplant as well as incorporation of a baseline risk score in prediction models that use post-transplant variables. The framework can be easily applied to other scenarios in which determining the baseline risk for infection prior to high-risk medical or surgical intervention might be beneficial.

**References**


HealthPro: An integrated web application for essential health data and biological specimen collection in the Precision Medicine Initiative

Kelsey R Mayo, Ph.D.¹, Robert J Carroll, Ph.D.¹, Jason Tan, M.S.¹, Rebecca L Johnston, M.S.¹, Celecia M Scott¹, Joshua Denny, Ph.D.¹, Paul Harris, Ph.D.¹
¹Vanderbilt University Medical Center, Nashville, TN, USA

Abstract

A critical opportunity and challenge within the Precision Medicine Initiative (PMI) is standardized health data and biospecimen collection for one million or more participants across diverse national enrollment sites. The HealthPro web application was developed to facilitate data standardization and communication between healthcare provider organizations and PMI data consuming nodes. Furthermore, HealthPro integrates with MayoLINK®, the Mayo Clinic’s universal test management solution, to enable a nationally scalable system for biospecimen collection.

Introduction

One primary goal of the All of Us Research Program, formerly known as the Precision Medicine Initiative Cohort Program, is to create a broad, representative public resource to advance the health of all Americans¹. To achieve this goal, All of Us will recruit one million or more participants through healthcare provider organizations (including regional medical centers, federally qualified health centers, and Veterans Affairs) as well as through direct volunteer enrollment site networks (e.g. Walgreens, EMSI, QTC). The number and diversity of these enrollment centers presents a unique challenge to the successful execution of the All of Us Research Program. To meet this challenge, the Data and Research Center (DRC) developed HealthPro, a novel tool to securely collect and transfer baseline health data and biospecimens as part of the All of Us initial core data set.

Results and Discussion

The DRC applied an iterative, user-centered design process to prototype, develop, and refine the HealthPro web application. Initial functional requirements were elicited through collaboration with All of Us Working Groups and Principle Investigators (PIs), and trained staff (e.g. nurses, research coordinators, phlebotomists) at All of Us enrollment sites were identified as HealthPro target users. In October 2016, prospective users and PIs (N=22) rated a prototype on usefulness and usability of the following core functionalities on a 5-point Likert scale from “Strongly Disagree” to “Strongly Agree”: 1) providing standardized online data capture tools for collecting and transferring participant’s baseline health data and 2) integrating with MayoLINK® to create and manage orders for collected biospecimen. Overall, responses were positive, with HealthPro usefulness and usability ranking a median of “Agree” (4) on the Likert scale. Based on feedback from this early user testing, HealthPro’s core functionalities were updated to include: 3) providing training reminders and capturing protocol deviations in a structured manner.

From November 2016 to March 2017, the All of Us enrollment sites (N=47) ran exercises using HealthPro for baseline data collection and biospecimen collection, processing, and shipment to the All of Us Biobank. Feedback from these practice exercises catalyzed development of new functionality: 4) providing a “Work Queue” for operational data (ex. date of consent and contact information) to enable enrollment workflows. 94.7% of users (N=38) reported that the Work Queue prototype was “Very Important” (N=14) or “Absolutely Essential” (N=22) to every day operations.

Final testing of the production version of HealthPro, integrated with the production versions of MayoLINK® and the All of Us Research Program database, successfully concluded in May 2017. Since the official launch of the All of Us Research Program on May 30, 2017, HealthPro has enabled participating organizations to successfully enroll, capture baseline health data, and collect biospecimens from more than 250 participants across the nation, growing daily.

Conclusion

User-centered design and innovative integration with clinical biospecimen management software resulted in HealthPro - a scalable solution for secure, standardized health data and biospecimen collection in support of precision medicine.

References

Precision Medicine Initiative Working Group. The Precision Medicine Initiative Cohort Program – building a research foundation for 21st century medicine. Report to the Advisory Committee to the Director, NIH. 2015.

Precision Medicine Initiative, PMI, All of Us, the All of Us logo, and The Future of Health Begins With You are service marks of the U.S. Department of Health and Human Services.
Electronic Health Record Safety: Identifying Measures for Clinical Decision Support Quality

Allison B. McCoy, PhD1, Adam Wright, PhD2, Hardeep Singh, MD, MPH3
Marie Krousel-Wood, MD, MSPH1, 4, Dean F. Sittig, PhD5

1Tulane University School of Public Health and Tropical Medicine, New Orleans, LA,
2Brigham and Women's Hospital, Partners HealthCare, Harvard Medical School, Boston, MA,
3Houston Veterans Affairs Center for Innovations in Quality, Effectiveness and Safety, Michael E. DeBakey Veterans Affairs Medical Center, Baylor College of Medicine, Houston, TX, 4Tulane University School of Medicine, New Orleans, LA, 5The University of Texas School of Biomedical Informatics at Houston, Houston, TX

Introduction
Recent reports have highlighted potential challenges and unintended consequences of electronic health records (EHRs), which are implemented to improve patient safety and reduce healthcare costs. Effective measurement of potential hazards may help healthcare settings detect unsafe system conditions and implement interventions to prevent medical errors. We sought to develop measures for EHR safety based on a previously described health information technology safety (HITS) framework and the National Quality Forum (NQF) report on prioritization and identification of HIT patient safety measures, focusing on the clinical decision support (CDS) key risk area.

Methods
We used the potential measure concepts proposed by NQF as a starting point and, through a consensus process, expanded each into a list of operational measures. Study team members (AM, AW, HS, MK, DS) independently reviewed each measure for implementation feasibility (i.e., measure is easily defined and computed, and the data are easily accessible in an existing, electronically available format) and usefulness (i.e., measure will be useful in assessing EHR safety). We generated a combined score by summing the median of reviewers’ feasibility and usefulness scores. We ranked the measures by the combined score and described factors that contributed to each measure’s scores.

Results
From the 6 NQF potential measure concepts, we identified 41 operational measures. Table 1 shows the three highest and three lowest scoring measures. Most measures with low implementation feasibility were difficult to define or required chart review or clinical judgment to implement. Some measures were recurring and could be recalculated over time to observe changes in CDS use, while others were one-time measures (e.g., institutional practices), often with a binary outcome (i.e., true/false), that were not likely to change over time. There was no apparent correlation between these characteristics and the resulting ranking, or between implementation feasibility and usefulness.

Conclusion
The NQF potential measure concepts served as a good starting point for identifying measures, but many of the concepts were too broad or imprecise and required discussion to translate into one or more measures, and many of the resulting measures were neither feasible nor useful. Further research is underway to refine and better rank the measures for CDS through an improved review process, and to identify additional measures in other key risk areas, including system downtime and patient identification.

Acknowledgments: This work was supported by NLM Grant R21 LM 012448.
Engaging Bedside Nurses in the EHR Change Request Process

Jacquel J. McDonald, MSN,RN-BC, Kimberly C. Burress, MIS,BSN,RN-BC
Children’s Hospital of Philadelphia, Philadelphia, PA

Introduction

Children’s Hospital of Philadelphia, had an established well process for handling EHR change requests generated from inpatient nursing. The process began with one analyst, supporting over 2200 requestors from the Department of Nursing (inpatient) and meeting with every requestor to define workflow requirements. The workflows were subsequently review by a group of practicing bedside nurses with expertise in the EHR [Epic Nursing Clinical Champions (ENCCs)] who often suggested significant redesign. The analyst began the cyclical process of meeting with requestors and the ENCC team until final design requirement were finalized. This methodology was labor and resource intensive for the analyst, caused delays in ticket initiation to completion rates, and requestor dissatisfaction.

Goals of this project were to: 1) 10% increase in nursing satisfaction with the Epic change request process 2) 10% increase in the number of Epic requests completed monthly 3) 10 % decrease in ticket initiation to completion rates

Methods

Starting in September 2016, a nurse who submits a change request receives a follow-up email requiring the completion of a scoring tool. The weighted prioritization scoring tool requires the nurse to evaluate the request based on the IOM’s six domains of health care quality and include quality effectiveness, productivity, work effort, patient safety, financial impact, clinician, and patient scope. Each scored request is validated by the informatics nurse specialist and Epic analyst. Based on the score and level of complexity an ENCC is assigned to meet with the requestor(s) and define the clinical and business requirements. The ENCC presents the final requirements to the team for feedback and approval. With the approval process complete, the change request is transitioned to the analyst who focuses on the build.

Results

![EHR Request Data]

Figure 1. Tickets open before the model change in September and two months of post data.

Conclusion

The ENCC participation in the request process decreased time required to complete a ticket from 201.6 to 139 days, on average (30% reduction) and decreased average ticket open time from 160 to 138.2 days (13% reduction). Initial outcomes indicate faster response times to nursing requests for EHR documentation changes and reflects increase in time the Epic analyst now has on completing request. Based on our pre and post model change, we have also increased overall end user satisfaction by 5%.
SMARTenAPP: The Development of a Web-Based Application as an Antimicrobial Stewardship Intervention: A Pilot Project

Richard Medford MD FRCPC¹, Siddhartha Srivastava MSc MD FRCPC², Gerald Evans MD FRCPC³

¹ Department of Clinical Informatics, Stanford School of Medicine, Palo Alto, California, United States
² Division of General Internal Medicine, Queen’s University, Kingston, Ontario, Canada
³ Division of Infectious Diseases, Queen’s University, Kingston, Ontario, Canada

BACKGROUND: Antimicrobial stewardship programs (ASPs) are useful in promoting appropriate antimicrobial use while maximizing patient safety and minimizing cost. Many institutions have implemented such programs but have faced challenges secondary to lack of clinical resources and time and economic constraints. In the era of advancing technology and increasing use of mobile health resources, it is important to look at this avenue as a possible stewardship tool.

OBJECTIVES: To develop and evaluate the utility of a web-based application (app) as an antimicrobial stewardship intervention.

METHODS: A web-based app to assist prescribers with selection of antimicrobial therapy based on local resistance data was developed. The app utilizes four main headings to direct prescribers to the pertinent sections: Clinical Condition, Microbiology, Antibiotics and Location. In March 2014, the app was introduced on to the Internal Medicine Clinical Teaching Units (CTUs). We looked at the uptake and use of the app using web-based statistics. We then did both a comparative and “before/after” analysis of the usage (Defined Daily Doses/1000-patient days) of two targeted antibiotics: piperacillin/tazobactam (TZP) and vancomycin (VAN) on surgical and medical floors at our hospital the year before and after our intervention.

RESULTS: From March 2014 to February 2015, there were 3110 sessions, 16,442 page views, 57.7% returning users and 42.3% new users. There were 5.28 page views and 2.31 minutes per session. Comparing targeted antibiotic usage rates on medical to surgical units from our intervention period - March 2013-February 2014 vs. March 2014-February 2015: utilization of VAN was 21% lower on the medical floors (p=0.0001), while TZP showed a non-significant increase of 7% (p=0.4222). This correlated to a decrease in DDD/1000 patient days of 66 to 55 for VAN and an increase in TZP from 35 to 42 DDD/1000 patient days.

DISCUSSION: A web-based application is a practical, accessible tool for prescribers to aid in antimicrobial selection. Our data suggest a reduction in use of targeted antibiotics is achievable and sustainable, especially in low-resource settings. Future directions include evaluating the usability and impact on local antimicrobial resistance patterns over extended time periods.
Cross-species Mapping of Human and Mouse Craniofacial Structures

Jose LV Mejino Jr. MD¹, Landon T Detwiler, MS¹,², Timothy C Cox, PhD³,⁴, and James F Brinkley, MD, PhD¹,²

¹ Department of Biological Structure, ² Department of Biomedical Informatics and Medical Education, University of Washington, ³ Division of Craniofacial Medicine, Department of Pediatrics, University of Washington, ⁴ Center for Developmental Biology and Regenerative Medicine, Seattle Children’s Research Institute, Seattle, WA, USA

The Ontology of Craniofacial Development and Malformation (OCDM)¹ is an OWL application ontology for use by the NIH-sponsored FaceBase consortium, as well as by the larger craniofacial research community. The purpose of FaceBase is to collect and integrate data ranging from genomic to clinical in order to understand the causes and possible treatments of craniofacial birth defects such as cleft lip and cleft palate. The purpose of the OCDM is to provide the terminology standards, as well as the relationships (e.g. parts, developmental lineage, associated malformations), to associate and correlate disparate FaceBase data pertaining to the craniofacial complex by formally representing multi-species canonical anatomy, embryological development, malformation, and cross-species anatomy homology mappings.

Understanding development and malformation requires knowledge derived from research on model organisms, such as mice and zebrafish, that cannot directly come from human subjects. The wealth and availability of information from these organisms can be leveraged and utilized to advance research at the human level. In previous work on the OCDM we have provided comprehensive representations, not only of human, but also of mouse and zebrafish craniofacial anatomy. In the work reported here we investigated homology correspondences between human and mouse craniofacial structures, and represented them in the Craniofacial Human Mouse Mapping Ontology (CHMMO), one of several sub-ontologies of the OCDM.

Mappings include developmental entities from the zygote to perinatal structures at progressive stages of development, with emphasis on embryonic precursors of components of the musculoskeletal system of the head (muscles, bones, cartilages, skeletal ligaments and joints). Despite overtly apparent differences between mice and humans, these organisms are closely related mammals and share very similar biology. The structural similarity between the two organisms is the basis of much translational research. We initially carried out provisional mappings based on existing nomenclature, incorporating similar names, and then validated these, through consultation with domain experts and the literature, for homology evidence, either based on direct molecular studies or inferred vertebrate Bauplan. Bi-directional one-to-one and uni-directional one-to-null mappings were recorded using a Protégé mapping plugin we developed (Figure 1). Except for the very early stages of development (zygote to germ layers), mappings were largely implemented at the gross anatomical level.

These mappings (and mappings in-progress relating human to zebrafish) will facilitate ontology-based searches of FaceBase model organism data, such as gene expression studies, that can give insight into the causes of clinical conditions such as cleft lip or palate. Funding: NIH DE24417.

References
Use of an iPad Program to Improve Access to Physical Therapy Services for the Management of Chronic Low Back Pain

Brittany L. Melton, PhD, PharmD1, Zoe Lai1, Neena Sharma, PT, PhD2
1University of Kansas School of Pharmacy, Lawrence, KS; 2University of Kansas Medical Center, Kansas City, KS

Abstract
This study sought to evaluate the impact of a home-based iPad program on chronic low back pain and associated psychological factors. Participants completed a 6-week program that provided pain education and exercises. Pain, fear, depression, and other factors were assessed at baseline and completion. Participants had a significant reduction in pain and improved energy. These results can be used to expand the program and improve access to physical therapy.

Introduction
Low back pain (LBP) is one of the leading causes of disability in the United States and is estimated to cost up to $635 billion annually.1,2 While exercise is the standard of care for chronic LBP, it does not address the psychological issues associated with chronic pain which can negatively impact patient quality of life.3 Less than a third of patients see a physical therapist for chronic LBP, and a home-based program which addresses both the exercise and psychological factors of chronic pain could improve access to care while improving patient outcomes.4 The objective of this study was to evaluate an iPad program’s impact on chronic LBP patient pain and functioning.

Methodology
Participants were given an iPad pre-loaded with pain education and exercise programs which incorporated stretching, strengthening, and aerobic exercises. Each session was an hour or less, and participants were instructed to follow the session recommended activities three times a week for six weeks, to reflect traditional physical therapy sessions. After two weeks, participants had a virtual visit via the iPad with a study investigator to answer any questions and ensure exercises were done correctly. Participants completed validated surveys related to fear avoidance, disability, and pain at baseline, two weeks, and at program completion. Three months after completion, participants were contacted about their LBP and exercises. Participant questionnaire responses were assessed for changes in participant pain and associated outcomes using Wilcoxon Signed Ranks Test. The study was approved by the Institutional Review Board.

Results
Twelve pilot participants were mostly female (72.7%), with a median age of 37 years, and median duration of current pain episode of 30 months. There was a significant reduction in pain (median 8.5 vs 4, p=0.005), and increase in energy (median 1 vs 0, p=0.007) after completing the program compared to baseline. Participants generally reported that the program was easy to use, and felt it gave them access to therapy they would not otherwise have. At the three month follow-up, some participants reported still performing the exercises without the presence of the iPad and some asked if they could continue to have access to the information and exercises because of the perceived benefit.

Conclusion
A pilot iPad program to educate and guide participants in chronic LBP management was effective, while also being usable. These findings will be used to refine and expand the program to increase rural access to physical therapists.

References
Research on the Accessibility of Parks Based on GIS in 8 districts of Beijing

Yujie Meng, MD, Xiaopeng Qi, PhD, Songwang Wang, MD, Qiang Chen, MD, Rui Zhang, MD, Huan Zheng, MD
Chinese Center for Disease Control and Prevention (China CDC), Beijing, China.

Introduction
Parks are the main part of urban green infrastructure, they have recreational function and can provide the most common places for population-level physical activity. The urban planners and policy-makers widely took park area per capita etc. as the key indexes to guide urban planning. They pay few attention to spatial accessibility and utilization of parks. Good park access is associated with increased physical activity1.

Methods
We measured residential population’s potential spatial access to parks by population-weighted distance (PWD) to parks using the method developed by Zhang and colleagues2. We selected 1 km population grids as basic geographic neighborhood unit of analysis and used centroids of 1 km square grids to measure geographic location more accurately. Firstly, 1 km grid level PWD to nearest parks for the population in that grid is calculated, and then the PWD for the population in larger geographic units than 1 km grid (counties) to visit nearest parks is aggregated.

Results
We drew the distribution and per capita park area for 8 districts of Beijing (see figure 1 and 2). Darker green means more park area per capita in figure 2. The county in the edge have more park area per capita than the county in the center. Chongwen district has the highest per capita park area. We also calculated the PWDs for 8 districts of Beijing. People in the 8 districts of Beijing are expected to travel 0.85 mile on average to access their local neighborhood park. Figure 3 depicts the county level PWDs to parks in 8 districts of Beijing. Darker green means better potential spatial access to parks for local residents. Chongwen District has the best access to local neighborhood parks with PWD of 0.38 miles. Haidian District and Shijingshan District have the largest PWDs of 1.03 and 0.97 miles, respectively. In summary, Chongwen District is the best location for these two indicators.

Conclusion
This study adopted a GIS approach to aggregate potential spatial access to parks from 1 km grid level to county level. Counties in the center part of Beijing have significantly better spatial accessibility to parks than the urban-fringe counties. The result in PWD is different from per capita park area in some districts. The accessibility of parks is a great supplement for per capita park area. It will enrich the key index system for assessing parks. The study evaluates the potential spatial access to parks and environment justice in Beijing. It could be informative for health professionals and policy makers to improve the availability and accessibility of public parks.

References
An NLP Tool for Boosting Annotation Capture from Clinical Documents

Frank Meng PhD\textsuperscript{1,2}, Craig Morioka PhD\textsuperscript{2,3}, Weixia Yu PhD\textsuperscript{2}, Nilla Majalahme BSE\textsuperscript{1}, Danne C. Elbers MS\textsuperscript{1}, Sergey Goryachev MS\textsuperscript{1,4}, Nhan V. Do MD\textsuperscript{1}

\textsuperscript{1}MAVERIC, VA Boston Healthcare System, Boston, MA
\textsuperscript{2}Medical Imaging Informatics, Dept. of Radiological Sciences, UCLA, Los Angeles, CA
\textsuperscript{3}VA Hines Healthcare System, Hines, IL
\textsuperscript{4}Ariadne Labs, Boston, MA

Introduction

Data-driven machine learning systems have the potential of revolutionizing healthcare by leveraging large data sets to automatically derive the patterns and knowledge needed to accurately augment the work of clinical professionals. Though these systems may eventually be capable of surpassing human levels of performance for certain tasks, their fundamental need and limitation is the availability of large annotated data sets that can be utilized as training data for generating the required knowledge. Many domains that have seen machine learning algorithms exhibit high levels of performance, such as object recognition in images and games of strategy such as chess, have readily available training data that is straightforward to obtain with little associated costs in terms of human labor. Obtaining high quality, large volume, annotated data sets within the medical field, however, is a more challenging endeavor due to the higher labor costs of annotators because they must possess sufficient medical knowledge and training. To alleviate this bottleneck, we have developed a pattern-based NLP tool with an intuitive web-based GUI that is capable of boosting the productivity of annotators by automatically pre-annotating documents with high precision.

Methods

We leveraged an existing information extraction (IE) system that uses multiple sequence alignments (MSAs) to automatically generate extraction patterns from annotated training documents [1]. Our working hypothesis is that these patterns can be used to identify specific data elements from within free text with high precision when trained on smaller training sets. Based on this, we designed a web application that provides a case report form (CRF)-based interface to clinical users for facilitating chart review and data abstraction workflows. The user can abstract data from documents into a CRF using only a point-and-click input device. This results in the capture of both structured data values and annotations that are used to train the MSA algorithm. The backend IE system utilizes user annotations to produce extraction patterns, determines which patterns exhibit high precision levels, and then applies the high precision patterns to documents that have not yet been seen by the user.

For this preliminary study, we leveraged a corpus of 100 de-identified lung cancer screening reports from the UCLA Healthcare System and designed a CRF containing a set of salient data elements based on direct input from thoracic radiologists who have experience reading lung cancer screening cases. This study focused on the annotation of four particular data elements: patient age, patient gender, the existence of a previous comparison exam, and smoking status. The reports were ordered arbitrarily and the annotator was instructed to fill out the four chosen data elements from the corresponding CRF for each report, moving from one report to the next sequentially. As each report was annotated, the backend IE system automatically annotated reports downstream from the user’s current position. Our results showed that the system was able to automatically annotate the four data elements in new documents at a level of 95% precision and 40% recall.

Conclusion

The results of this study show the potential of using pattern-based NLP to boost the efficiency of clinical annotators. The fact that the results exhibited high precision levels was important because it reduces the amount of manual effort needed to correct the system’s errors. On the other hand, a recall of 40% enabled the system to locate almost half of the cases that the user would have had to manually annotate, saving overall time and effort.

References

Reflecting on Diabetes Self-Management Logs with Simulated, Continuous Blood Glucose Curves: A Pilot Study

Elliot G. Mitchell¹, Matthew E. Levine¹, Dave J. Albers¹, Lena Mamykina, PhD¹
¹Department of Biomedical Informatics, Columbia University, New York, NY

Introduction
Continuous Glucose Monitors (CGM) are a powerful intervention to support patients with diabetes in better managing their condition – using a CGM significantly improves glycemic control in individuals with diabetes.¹ However, CGMs can be uncomfortable to wear, and are too expensive to be standard care for type 2 diabetes.² We hypothesized that viewing computationally generated, continuous blood glucose curves can yield similar benefits to reviewing CGM data, like improved reasoning and reflection, without the cost and discomfort of a physical monitor.

Methods
Participants were 5 adults (3 female) with diabetes who had recently completed a trial of GlucOracle, a diabetes self-monitoring smartphone application for recording meal logs, with nutrition and pre/post meal blood glucose readings. We conducted a mixed-methods, within-subjects pilot study to compare the experience of reviewing blood glucose as single point measurements, or with a continuous curve. All participants reviewed summaries of their meal logs (Figure 1) with blood glucose represented as pre/post points (Figure 2), or with a continuous curve (Figure 3). The curves were generated using a physiology-based computational model, trained on each participant’s individual data.³,⁴ After reviewing a set of meals, participants also completed a prediction task with another set of meals in which they estimated blood glucose readings after their meals. Participants drew their estimate on the graph using either a single point or a continuous curve, matching the format they had just reviewed. Baseline prediction was assessed before reviewing any meals, and meals were presented in a random order.

Results and Conclusion
All five participants preferred viewing their data with simulated curves. Whether or not they had used a CGM in the past, they thought the continuous representation was intuitive, and that it provided more information than a single post-meal finger stick reading. Many participants mentioned that a single reading lacks information about what happened leading up to the reading, or which direction blood glucose is trending.

While there was no significant difference in the mean squared error of estimates between point and curve conditions, the small sample size of our pilot study limits our power for statistical inference. Counter to our hypothesis, baseline performance was actually slightly better than performance after reviewing meals. This underscores the notion that predicting blood glucose change after a meal is difficult,⁵ and suggests that reviewing individual meals may not improve estimates, and could actually make estimates worse.

References
Clinical Informatics for Precision Cancer Medicine: Computational Methods for Interpreting Individual Whole Exome and Transcriptome Tumor Sequence Data

Nathanael D. Moore, BS1,2,3,4, Brendan Reardon, BS1,2, Eliezer M. Van Allen, MD1,2,5
1Dana-Farber Cancer Institute, Boston, MA; 2Broad Institute of MIT and Harvard, Cambridge, MA; 3Howard Hughes Medical Institute, Chevy Chase, MD; 4Indiana University School of Medicine, Indianapolis, IN; 5Harvard Medical School, Boston, MA

Introduction

Precision cancer medicine seeks to improve cancer care by tailoring treatments to individual patients’ alteration profiles. Towards this end, we have developed PHIAL (Precision Heuristics for Interpreting the Alteration Landscape), a heuristic clinical interpretation algorithm, and TARGET (Tumor Alterations Relevant for GEnomics-driven Therapy), a knowledge base of clinically actionable alterations. Combined, PHIAL and TARGET can help to inform cancer treatment decisions at the point of care. Our current work involves greatly enhancing and reimplementing PHIAL and TARGET to provide individualized treatment recommendations with greater accuracy.

Methodology

PHIAL was originally designed to analyze somatic whole exome sequencing data. While it was able to extract meaningful alterations from patient tumor data, it was unable to distinguish the relative actionability of variants, did not consider somatic-germline interactions, and was limited to first-order genomic relationships (e.g., BRAF V600E → RAF/MEK inhibition). TARGET began as a catalogue of putatively actionable alterations, but lacked specificity beyond the gene level and was unable to denote the relative confidence of each clinical assertion.

We made several improvements to increase PHIAL’s predictive abilities. PHIAL was revised to predict actionability based on the presence of SNVs, indels, SCNAS, and fusions derived from transcriptome data, by matching the presence of these alterations to information stored in several alteration-action databases. PHIAL also assesses global features (e.g., context-specific mutational burden) for possible clinical relevance. TARGET was redesigned to allow representation of multiple feature types at varying levels of specificity (“assertions”) and to incorporate indicators of clinical confidence (“predictive implication levels”) for each assertion. Finally, 215 new alteration-action relationships were added to TARGET and the resource was made publicly accessible through a web portal.

Results

We applied the original (PHIAL1) and updated versions of PHIAL (PHIAL2) to a cohort of 255 patients with whole exome and transcriptome sequencing data (146 castration-resistant prostate cancer and 109 metastatic melanoma samples). PHIAL1 identified 1,342 putatively clinically actionable/biologically relevant events across the cohort, with a median of 3 events per patient (95% of patients had at least one event). PHIAL2 identified 2,508 events, with a median of 6 events per patient (98.5% of patients had at least one event). Of these events, TARGET associated 8.12% with an FDA-approved therapy and 2.09% with a clinical trial.

Conclusion

Our improvements allowed identification of a greater number of putatively actionable alterations in our patient cohort. PHIAL successfully transitioned from a variant-based to a feature-based approach, and was able to leverage the additional information recorded in TARGET to make more meaningful clinical suggestions. This strategy could inform future use of whole-exome/transcriptome sequencing in the clinical setting, and help identify clinical actions in cases where there is no clear single alteration associated with oncogenesis.

References

A Comparison of Stroke Classifiers Leveraging Hospital Billing Codes versus Natural Language Processing

Danielle L. Mowery PhD MS1,2; Brent Hill PhD3; Mingyuan Zhang MS4; Wendy W. Chapman PhD1,2; Lisa Cannon-Albright PhD MStat5; Jennifer J. Majersik MD6

1IDEAS Center, George E. Wahlen Veteran Affairs Medical Center, 2Biomedical Informatics, 3Clinical Neurosciences Center, 4University of Utah HealthCare, 5Division of Genetic Epidemiology, 6Division of Vascular Neurology; Salt Lake City, UT

Introduction

In large-scale epidemiological studies, ICD-9 billing codes can misclassify many non-stroke patients as stroke cases resulting in varied precisions (ischemic: 66-95% and hemorrhagic: 65-89%)1. More precise classification might be achieved using information extracted from the clinical notes using natural language processing (NLP) and machine learning. Therefore, we developed an NLP-powered stroke classifier and compared its performance to a classifier using ICD-9 billing codes. This is a first step toward NLP-powered stroke phenotyping to support large-scale pedigree studies leveraging the electronic health record (EHR) and the Utah Population Database.

Methods

For this Institute Review Board-approved pilot study, we randomly selected 400 patients hospitalized in the Neurocritical Care Unit at the University of Utah Hospital from 2008–2015 with hand-adjudicated discharge diagnoses based on EPIC EHR data by two domain experts. This dataset consists of three classes: 50% ischemic stroke, 25% hemorrhagic stroke, and 25% acute neurological disorders (other). We compared the performance of an NLP-powered stroke classifier (linear support vector machine) trained using n-grams (1-4 word features) from radiology notes compared to a stroke classifier trained using ICD-9 billing codes. The NLP classifier was trained with non-stop word features occurring more than twice and ranking in the top 1% using Chi-square feature selection (n=620 features). Each classifier was trained and tested using 5-fold cross-validation (average scores below).

Results

Our NLP stroke classifier produced comparable recall and promising precision for each class.

<table>
<thead>
<tr>
<th></th>
<th>F1-score</th>
<th>Recall</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICD-9 codes (primary + secondary codes)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ischemic Stroke</td>
<td>94.9</td>
<td>96.0</td>
<td>94.2</td>
</tr>
<tr>
<td>Hemorrhagic stroke</td>
<td>94.9</td>
<td>97.0</td>
<td>92.9</td>
</tr>
<tr>
<td>Other</td>
<td>94.1</td>
<td>91.3</td>
<td>97.2</td>
</tr>
<tr>
<td>N-grams</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ischemic Stroke</td>
<td>81.2</td>
<td>95.5</td>
<td>70.1</td>
</tr>
<tr>
<td>Hemorrhagic stroke</td>
<td>90.3</td>
<td>96.0</td>
<td>85.3</td>
</tr>
<tr>
<td>Other</td>
<td>87.7</td>
<td>95.3</td>
<td>81.2</td>
</tr>
</tbody>
</table>

Conclusion

The NLP stroke classifier produces precisions within known performances of epidemiological studies1, but a more complex approach is required to produce comparable ICD-9 precisions, which performed better than anticipated. Given the ICD-9-based precisions are not likely to precisely identify pre-EPIC stroke cases locally and stroke cases at other institutions, we are enhancing our NLP stroke classifier to leverage features from a stroke ontology2.

References

Developing a Biomedical Information System for the geo-spatial analysis of California Office of Statewide Health Planning and Development (OSHPD) Data among minority population in Los Angeles

Sukrit Mukherjee, PhD, Paul L. Robinson, PhD
1Charles R. Drew University of Medicine and Science, Los Angeles, CA, USA

Abstract
At Charles R. Drew University of Medicine and Science, researchers analyze large secondary dataset called OSHPD using existing ArcGIS software for geo-spatial analysis. But due to limitation in labeling capabilities, role based logins, customized screens and dependence on GIS personnel in ArcGIS, we developed an alternative web-based tool to access, retrieve and analyze OSHPD data from geo-spatial perspective contributing towards an enhanced effort in translational research for removing health disparities at Los Angeles.

Introduction
At Charles R. Drew University of Medicine and Science (CDU), researchers are involved in translational research activities involving minority population in Los Angeles with a goal of removing health disparities. At CDU’s Urban Health Institute, researchers need to study OSHPD data for clinical, epidemiological and health services research. The Institute used to have a license of an ArcGIS software for geo-spatial analysis of OSHPD data and the operation of the ArcGIS needed a GIS person on board to help researchers to run the data through ArcGIS. Our current version of ArcGIS has no labeling capabilities. The role-based security features provided in ArcGIS is also very limited. Although ArcGIS provides some good user interfaces, they are more geared towards sharing data within an organization. They are not really designed to share analysis and study with our collaborators. On top of that, there is very limited support available for developers to customize the screens and analysis algorithm as long as geo-spatial analysis is concerned. These factors justified the background of development of a biomedical information system called ‘OSHPD Geo-Spatial Analysis System’ to load, access, retrieve and analyze OSHPD data for geo-spatial analysis for the study of minority population in Los Angeles.

Methods
OSHPD Geo-Spatial Analysis System was designed and developed after an initial feasibility study which involved carefully analyzing the requirements for our geo-spatial analysis, the existing process of studying OSHPD data and the usage of the data with ArcGIS. The current research strategy in place was used to develop initial conceptualization for the new system including algorithms, functionality, featured interfaces and the expected customized behavior of the system so that researchers could operate the system themselves without turning to GIS personnel for their analysis and processing of OSHPD data.

Results
The pilot implementation of the system was designed with web-based front end designed with ASP.Net and C# and MS SQL Server as backend with clinical, epidemiological and health services researchers in mind. The users specify search criteria including demographics, diagnosis, type of care, admission types, payment, etc. There is an ETL (Extract, Transform and Load) module embedded within the system that loads multiple years of OSHPD data into an optimized SQL Server Enterprise Edition database from flat files media. An automatic parser module located next to the ETL module verifies for the correct syntax of the input data and creates appropriate database schema. The system processes the user input, performs calculations, retrieves data and generates reports either (i) to feed as user specified format to other data analyzing software e.g. SPSS or (ii) to analyze the report online. The ultimate goal of the analysis report or the downloadable file is to act as an efficient aid for the geo-spatial analysis of the minority population belonging to SPA 6 region in Los Angeles.

Conclusion
We addressed the limitations of the ArcGIS software and created a pilot version of a customized solution to access, retrieve and analyze OSHPD data for geo-spatial analysis of the minority population in SPA 6 of Los Angeles. When the system will be deployed on its full scale, this system carries potential to increase the productivity of translational research at CDU thus contributing towards an enhanced effort of removing health disparities at Los Angeles.
Global HIV Case Surveillance Using Electronic Medical Records (EMR) Data
Lisa A. Murie, MSc, Eric-Jan Manders, PhD, James Kariuki, MPH, Mahesh Swaminathan, MD, Xenophon Santas, BA
Centers for Disease Control and Prevention, Atlanta, GA, USA

Introduction
The global HIV epidemic remains a major health challenge in low and middle-income countries (LMIC). Approximately 37 million people are living with HIV globally, 46% of those eligible for treatment receive it, and an estimated 2.1 million new infections occur every year1. The United Nations Programme on HIV/AIDS (UNAIDS) set an ambitious global target that aims to diagnose 90% of all people living with HIV, 90% of whom would receive sustained treatment, and 90% of those on treatment would achieve viral suppression by 20202. To measure progress towards these goals, strategic information planning increasingly requires patient centric monitoring3, and for public health decision making, establishing routine HIV case-based surveillance (CBS) is an essential tool to describe the dynamics of the epidemic. HIV CBS establishes a longitudinal data set to track key events of public health relevance along the continuum of care. The US President’s Emergency Plan for AIDS Relief (PEPFAR) is focusing its effort where the impact of resources is maximized by targeting high burden geographic areas and high volume service delivery settings4. PEPFAR has also supported implementation of electronic medical records (EMR) systems and improved information and communications technology (ICT) infrastructure. This work lays the foundation for leveraging EMR data in the design and implementation of an HIV CBS system5.

Methods
The organization of health services, including governance and policy, are all inputs in the design of the physical architecture and information flow of a national HIV CBS system. For HIV, privacy and security concerns are particularly critical. Based on guidelines mentioned, we describe key events in the care continuum and define the trigger conditions to transmit data from service delivery to the public health level. We construct a minimum data set for the corresponding case report message, and describe the process and content standards that promote data sharing and reuse in public health. Through data mapping, we assess an EMR system’s ability to generate the minimum dataset, which could help identify and address potential gaps in data quality before implementation. We define the role of a surveillance officer - an EMR user who reports to the public health level; and we discuss the role of ICT infrastructure as a primary enabler for electronic data exchange. Lastly, we define an evaluation criteria used to measure the HIV CBS system’s impact on public health decision making.

Results
We have defined a minimum dataset and trigger conditions for an HIV CBS system whose case reports are generated from an EMR, and documented how information flows from a health facility to a public health agency using a business process model.

Discussion
Governance and policy; standards and guidelines; workflow; and ICT infrastructure influence the design of a HIV case surveillance system. This approach leverages on EMR data and increases the value of patient data for public health decision making, and may improve program planning. A thorough documentation of this approach aims to inform a country-led process of defining a regulatory framework, which would enforce compliance to standards and guidelines for HIV case surveillance. A country that aims to implement an HIV CBS system will benefit from an evaluation design where informatics principles and practice as outlined in this work may contribute to the evidence base on how the information system impacts public health decision making.

References
Does Practice Follow Guidelines? Improving Accuracy for Comparison of Structured and Unstructured Text in Medical Records to Expectations
William J Murphy, MS, Academy of Nutrition and Dietetics, Chicago, IL; Martin M Yadrick, MBI, MS, Computrition, Inc., West Hills, CA; Rosa K Hand, MS, Academy of Nutrition and Dietetics, Chicago, IL

Background: Comparative effectiveness research on evidence-based practice guidelines must be able to distinguish whether care follows recommendations in order to attribute health outcomes to guidelines. Previously, we convened an expert panel to translate recommendations from one guideline into 6 connected steps of an expected care plan (ECPs) using a combination of the Nutrition Care Process and Terminology (NCPT) standardized language(1) and unstructured text. We created an R language(2) script to process data from Registered Dietitian Nutritionist visit documentation to determine the congruence of practiced care with ECPs. We reported previously(3) on a validation study using a blinded manual audit of a random subset of 124 standard comparisons and computed intra-class correlations (ICC, $1 - \beta = 0.8$ for an ICC of 0.6 v. 0.1). The results were acceptable overall (ICC = 0.60), but steps including natural language processing (NLP) performed poorly (ICC 0.0 – 0.38). This report describes our process of improving algorithm performance, including integration with the Unified Medical Language System (UMLS) SPECIALIST lexical tools (SLT) and Metathesaurus (MT) version “2016AA” including >100 ontologies (4).

Methodology: We created an R language interface to the SLT command line tools and the MT Web API. We considered care plan text to match an expectation if they shared ≥ 1 common result in MT queries. We used normalized strings, converted via the SLT Norm program, as MT queries. For documented care unstructured text, we also corrected spelling with SLT GSpellFind and tokenized into n-grams of length 1 – 4. We report on the "goal" step, i.e. specific behavior and health outcome targets, which is unstructured and started with an ICC of 0.00.

Results: All ECP standards for unstructured text matched to ≥ 1 MT concept, indicating relevant concepts were well represented. In sum, “goal” step ICC increased from 0.00 to 0.66 and overall ICC from 0.60 to 0.81. Stepwise improvements are listed in Table 1 to distinguish steps with generalizable NLP from study-specific audits. We recovered “goal” step records from 15 comparisons inadvertently excluded due to missing data. We refined some ECPs based on audit reviews by: excluding some broad concepts causing false positives (e.g. “percent energy intake”), expanding some narrow concepts causing false negatives (e.g. from “protein intake” to “protein”), and adding some near-synonymous standardized terms (e.g. “concentrated sweets intake” in addition to “sugar intake”).

<table>
<thead>
<tr>
<th>Algorithm Alterations</th>
<th>Alteration Type</th>
<th>Overall ICC</th>
<th>Goal step ICC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial</td>
<td></td>
<td>0.60</td>
<td>0.00</td>
</tr>
<tr>
<td>Integrate with UMLS, using n-grams of length 2 – 4 from care plan text</td>
<td>NLP</td>
<td>0.64</td>
<td>0.19</td>
</tr>
<tr>
<td>Recover inadvertently excluded entries</td>
<td>Audit</td>
<td>0.67</td>
<td>0.31</td>
</tr>
<tr>
<td>Correct spelling and include unigrams from care plan text</td>
<td>NLP</td>
<td>0.70</td>
<td>0.44</td>
</tr>
<tr>
<td>ECP standard refinement</td>
<td>Audit</td>
<td>0.78</td>
<td>0.66</td>
</tr>
</tbody>
</table>

Table 1. ICC improvements with each additional algorithm refinement

Conclusion: Using a medical lexicon and tools greatly increased the accuracy of NLP comparisons between documented care and expectations. Audit results also facilitated the correction of data processing errors. There is a risk for over-fitting due to the repeated use of validation data, but this is limited as we did not use statistical learning or modelling. We now have a tool with acceptable accuracy for the automated comparison of large quantities of structured and unstructured clinical care documentation to guideline expectations.

References
Vaccine Code Set Management Service Pilot

Stuart Myerburg, JD¹; Regina A. Cox, MPH²
¹Centers for Disease Control and Prevention, Atlanta, GA; ²Deloitte Consulting, Atlanta, GA

Abstract:

CDC’s Immunization Information Systems Support Branch (IISSB) publishes the vaccine terminology code sets used for interoperability and reporting across vaccine delivery communities. Methods to manage and distribute these data sets have not kept pace with technology. The Vaccine Code Set Management Service (VCSMS) pilot has introduced automation and advanced practices to improve data quality, management, and distribution. Results will be demonstrated through feedback and testing, along with examples of data transformation and leading practices.

Introduction:

IISSB works with Immunization Information Systems (IIS), health information management vendors, standards development organizations, and other health organizations to assist with the development of standards and tools for the exchange of immunization data between healthcare providers and IIS. As part of this effort and in order to promote interoperability, IISSB has created a number of code sets to identify and map vaccines, vaccine manufacturers, vaccine groups, and Vaccine Information Statements (VIS). Electronic health records (EHR), drug databases, and public health organizations require these important data sets for recording and tracking vaccine events. The codes and maps are available for download on the CDC website. The data sets include:

<table>
<thead>
<tr>
<th>CDC Authored Standards</th>
<th>Externally Authored Standards (used in CDC maps)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CVX – Vaccine identification codes</td>
<td>NDC – National Drug Codes (FDA)</td>
</tr>
<tr>
<td>MVX – Vaccine manufacturer identification codes</td>
<td>CPT® – Administration Codes (AMA)</td>
</tr>
<tr>
<td>VIS – Vaccine Information Statements, document IDs</td>
<td></td>
</tr>
</tbody>
</table>

The Vaccine Code Sets Management Service (VCSMS) pilot was established to explore new tools for management and delivery of these immunization code sets. The pilot tools and data foundations have been implemented, and pilot constituents will be engaged from March 2017 – August 2017. The effectiveness of the pilot to meet the requirements of the participants will be reviewed based on a pre-defined approach. Lessons learned will be captured and used to inform future discussion and decisions around this service. These findings will be communicated in August.

The main objectives of the pilot are:

- Improve data management and quality – Increase automation to improve efficiency and quality, establish consistently-applied data editorial policies, and improve sourcing of external reference data used for code set maintenance and mapping.
- Expand content structure – Deliver new content data files specifically designed to aid integration with electronic systems.
- Provide distribution alternatives – Provide new configurable options for direct delivery to end user organizations.

Results will include:

- Detailed review and documentation of content changes attributable to pilot automated approaches
- Highlights of new workflow and automation practices implemented for the pilot
- Pilot participant feedback and analysis
- Summary of lessons learned
Pilot study using wearable and mobile app data in patients with the sickle cell disease to describe and predict painful vaso-occlusive crisis.
Kalindi Narine, MD\textsuperscript{1}, Fan Yang\textsuperscript{2}, Tanvi Banerjee\textsuperscript{2}, Jude Jonassaint\textsuperscript{3}, Nirmish Shah, MD\textsuperscript{1}
\textsuperscript{1}Duke University, NC; \textsuperscript{2}Wright State University, OH; \textsuperscript{3}University of Pittsburgh, PA

Abstract: Sickle cell disease (SCD) is a red blood cell disorder complicated by lifelong issues with pain. We used a novel mobile app with wearable sensor to allow analysis using machine learning to better display changes in pain and interventions. We found important correlations between pain and physiologic measurements, therefore believe this technology will provide important insight into patients hospitalized for pain with the potential to predict pain.

Introduction: SCD is typically complicated by painful vaso-occlusion, the most common cause for hospitalization. Through an innovative and interdisciplinary approach involving both clinicians and computer scientists, we have utilized a novel mobile app combined with wearable technology to improve measurements of subjective (such as pain) and objective data (such as heart rate and activity). We now describe our novel TRU-Pain system (Technology Resources to better Understand Pain) that uses statistical analysis and machine learning to map the patients’ physiological information with their pain scores as a means to assist with documentation of pain, interventions, and overall general health. Ultimately, we aim to supplement subjective reports of pain with these objective measurements to better predict pain.

Methods: Patients recorded their symptoms including pain and general health using the TRU-Pain application while wearing the wearable device up to 7 days while hospitalized. The wearable sensor (Microsoft Band 2) measured data including heart rate (HR), activity (steps), galvanic skin response (GSR), and ambient light. Pain is being determined using nursing assessments of pain using visual analog pain scores (0-10; with 0 indicating no pain and 10 indicating the highest pain level) and vital signs (HR, temperature, systolic and diastolic blood pressure). We began our analysis of patient pain data and vital signs using statistical measures including multiple imputation and correlation, as well as machine learning techniques such as logistic regression and Bayesian networks. Wearable data was preprocessed using signal processing techniques such as signal magnitude area and interpolation.

Results: We enrolled and analyzed 11 patients, 40\% females, median age 17 (range 13 to 54) who were admitted for a median 5 days (range 2 to 8). Median data collected was 18,415 (range 352-39,294) points per day with a median of 7 hours (range 6 min to 15 hours) of wearable data. Variation and missing data was found due to removal of the device to charge or the device losing charge while patient was wearing it and not realizing. The need to charge the device frequently is due to the high frequency of measurements. We found a sampling rate reduction of 10 to be reliable for data recording i.e. wearable data entries of 1 per 10 seconds.

We utilized multinomial logistic regression (MLR), one of the most popular and simple machine learning techniques, and support vector machines (SVM), a more complex and nonlinear supervised machine learning technique. MLR proved to outperform the SVM for the SCD patient data. Using clinical records, we were successfully able to identify the patients’ pain score using HR, temperature, and blood pressure data. Using MLR, we obtained a median accuracy level of 0.618 and a weighted F1 score of 0.551. Specifically, we chose weighted F1 score as it is a more generalized metric of evaluation for imbalanced classes. At an individual level, for example, for Patient 3 we were able to achieve a prediction accuracy of 80\% with a logistic regression model to predict the pain scores using multiple imputation method for missing data handling. We chose to analyze the data with machine learning techniques due to it not being temporally challenged; depending more on the number of data samples generated rather than the duration of the data collection, as well as the technique used for evaluation (such as resubstitution, cross-validation, etc.).

Conclusion: Our TRU-Pain system shows strong potential in being able to describe and predict the subjective pain in SCD patients using physiological measures. Our next steps for integration into our pain prediction model include wearable objective data (such as heart rate and activity) and patient subjective reports via our TRU-Pain mobile app. Furthermore, we are decreasing the sampling rate to help battery life and continue to enroll patients admitted in pain.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Median Accuracy</th>
<th>Median Weighted F1 Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLR</td>
<td>0.607</td>
<td>0.537</td>
</tr>
<tr>
<td>SVM</td>
<td>0.618</td>
<td>0.551</td>
</tr>
</tbody>
</table>

Figure 1. Pain prediction accuracy and weighted F1 score results
The Pulse: An Interactive Web-based Application for Tracking Clinical Quality Measure Performance

Benjamin E. Nealy, BS; Raja A. Cholan, BS; Nicholas V. Colin, MA; Bhavaya Sachdeva, MPH; Kevin G. Loftis, BS; David A. Dorr, MD, MS
Oregon Health & Science University, Portland, Oregon, United States

Introduction. The EvidenceNOW initiative provides support to thousands of primary care practices nationally across seven regional cooperatives to improve patients’ cardiovascular health by increasing adherence to key clinical quality measures (CQMs). The regional cooperative in Oregon, Washington, and Idaho is known as “Healthy Hearts Northwest” (H2N). The Pulse is an R-Shiny application used to provide timely, performance-based feedback to 197 practices within the H2N cooperative. We sought to meet information needs of the clinics by creating and revising a display incorporating data validation, benchmarks, performance over time, and changes in population.

Methods. We collect data in quarterly intervals for four CQMs using REDCap surveys. Known as the “ABCS”, the CQMs are: Aspirin when appropriate (CMS164, NQF0068), Blood pressure control (CMS165, NQF0018), Cholesterol management (CMS347, NQF pending), and Smoking cessation (CMS138, NQF0028). Raw survey data is extracted from REDCap into a MySQL database with minor processing. The Pulse Shiny application pulls data from the database in a daily refresh. Additional features, such as tables displaying past submissions, were added based on user feedback. Design and interface improvements were made using Nielsen’s Usability Heuristics, along with feedback from semi-structured qualitative user interviews.

Results. A test page for the application can be viewed at the following URL: https://octri.ohsu.edu/healthyheartsnw/?clinic=TEST. Each participating practice is assigned a unique identifying tag, which serves as a URL parameter directing users to their specific page. The page displays practice-specific process control charts, numerator and denominator bar plots for each CQM, data summary tables, and data validation messages. With an actively refreshed connection to the database, information on The Pulse is kept up to date within 24 hours. Users can display any of the four ABCS measures individually, or view all four simultaneously, using navigation buttons at the top of the application. A set of tabs allows users to switch between the process control charts, numerator and denominator bar plots, both plots side-by-side, or a CQMs summary and definitions page. The first screenshot shows the general layout of the process control chart next to the numerator and denominator bar plot; the second displays sample data validation messages. Participating practices have expressed enthusiasm for The Pulse as a data review tool. Practices' use of The Pulse varies from detailed review of the practices' data to a display used in patient waiting areas. Many facilitators working with enrolled practices use the application to share the practice’s CQM progress at site visits. Others print pages and make annotations to track quality improvement changes and historical changes in performance.

Discussion. We successfully developed and launched The Pulse to fulfill H2N-enrolled practices' need for up-to-date progress tracking and data presentation. The use of both peer and global benchmarks, changes over time, data validation, changes in submitted populations, and variations in display have led to wide use and expressed satisfaction. The technical build can easily be replicated, and the code base used for the application and the REDCap survey is available as an open resource.
Experiences Implementing a User-Centered Design Process across a Large Patient Safety Learning Laboratory

Pamela M. Neri, MS1, Anuj K Dalal, MD2,3, Theresa E. Fuller2, Dominic J. Breuer, MS4, Awatef Ergai, PhD4, David W. Bates, MD, MSc1,2, James Benneyan, PhD4

1Partners Healthcare Information Systems, Boston MA
2Brigham and Women’s Hospital, Boston, MA, 3Harvard Medical School, Boston, MA
4Northeastern Healthcare Systems Engineering Institute, Boston, MA

Problem: User-centered design (UCD) is a critical component to ensuring health information technology is adopted in complex health settings. Yet, many healthcare IT projects struggle to follow a rigorous user-centered design process because they do not have access to resources and expertise. We describe our experience of implementing a centralized UCD process as part of a large AHRQ patient safety learning laboratory initiative to design, develop, and implement novel technology (acute care patient portal, safety reporting system, safety dashboard, safety screen-savers) with the goal of making acute care safer.

Description: We assembled a core team of interdisciplinary experts with the goal of providing centralized systems engineering, human factors, and data analytic expertise to each of three projects. With assistance from the core, individual project teams worked to develop and implement patient- and provider-facing tools designed to identify, assess, and mitigate safety threats. The core shared templates and examples of user research collection forms and usability test plans that were used by each of the projects. Patients and providers on the study units were invited to participate in these methods. Core members attended individual project meetings to guide and encourage a user-centered approach. We provided usability feedback on prototypes that was used for iterative refinement, and conducted workflow observations to evaluate use of these tools during pilot testing. We assisted with content analysis of feedback gathered from these activities. We used a 2-person consensus approach to identify lessons learned.

Lessons Learned: While there was some success in soliciting input from users on the usability and design of individual tools and using UCD methods for individual projects, we encountered several challenges integrating a centralized and rigorous UCD process across the entire learning laboratory as originally envisioned. Key lessons learned include the need for 1) upfront expectation setting and standardization of the UCD approach across individual projects; 2) focused UCD educational sessions and tutorials to ensure consistent and comprehensive methods for gathering user feedback; 3) centralization of the user requirements gathering phase to specifically include requirements for how individual technological components (safety icon on bedside screensaver for patients and clinicians; dashboard event indicators on safety dashboard for clinicians) could work together as a unified system to prevent specific harms (fall, CAUTI, CLABSI). A secondary goal was to ensure we addressed the needs of all user types. We learned that different user types varied considerably with regard to their needs, mental models, literacy, and experience using technology in general, including the technological components of our intervention. Particularly challenging in this setting was gathering user input from a vulnerable patient population. Lastly, another major challenge to incorporating user-centered design across the individual projects of our safety learning laboratory was navigating the technical and design constraints of a new vendor-based enterprise EHR.

Summary: The core team of interdisciplinary experts was able to provide some centralized guidance and expertise in user-centered design across the patient safety learning lab. We learned that we could have benefited from earlier and more rigorous planning, education and requirements gathering, including identification of and planning in regards to the specific challenges of conducting this work in the healthcare environment.
Impact of Predictive Analytics on Clinical Outcomes, Clinical Decision Making, and Healthcare Organization Performance: A Systematic Review

Neris M. Nieves-Robbins, MD1, Matthew Parry, DO1, Andrew J. Holdaway, MD1, Robert C. Marshall, MD1, Fredric Wolf, PhD2, Peter Tarczy-Hornoch, MD2
1Madigan US Army Medical Center, Tacoma, WA; 2University of Washington, Seattle, WA

Introduction

Adoption and focused use of predictive analytic capabilities have been proposed as effective strategies to improve clinical practice, patient health and the delivery of healthcare services. Despite the rapid growth of the healthcare analytics market, there is a lack of integrated evidence from the peer-reviewed literature to understand the impact of predictive analytics across the various healthcare industry domains.

Purpose

The purpose of this study was to systematically review the literature for evidence of predictive analytics’ impact on clinical outcomes, clinical decision-making and healthcare organization performance.

Methods

An electronic literature search of PubMed, CINAHL, Web of Science and Business Source Complete databases was conducted for articles published in English from January 1, 2010 to January 30, 2017, after consultation with a librarian. All searches included the keywords (predictive analytics) OR (machine learning). The Business Source Complete search also contained the term AND (healthcare) to exclude results from other industries. All studies that evaluated the impact of predictive analytic capabilities used by healthcare organizations or employees to optimize performance, patient outcomes or decision-making were considered. Studies that did not refer to predictive analytics or systems with predictive capabilities, did not pertain to healthcare, and/or did not refer to systems used in actual practice were excluded. Two reviewers independently screened the titles and abstracts to assess publication eligibility. Differences were resolved by consensus and as needed, in consultation with an experienced researcher. Information was extracted from the eligible publications in accordance with Preferred Reporting Items for Systematic Reviews and Meta-analyses Protocol 2015 (PRISMA-P). Standard tools were used to assess the methodological quality and the risk of bias of the eligible publications.

Results

Database searches returned a total of 42144 articles. A total of 41382 unique titles remained after 3679 duplicate publications were removed. The titles and/or abstracts of the 41382 articles were reviewed. A total of 257 articles required a detailed abstract review. Sixteen articles were categorized as potentially eligible based solely on detailed abstract review and were retrieved for full article review. After full article review, a total of 5 out of the 16 articles were categorized as eligible. The most common reasons for article exclusion were that the articles were limited to: the conceptual domain, a computerized test environment and expert opinion. The eligible articles focused on two major areas: high acuity service over-utilization and operations management optimization. Risk of bias and methodological quality assessments of the studies showed mixed results. The designs, participants, interventions, comparators and outcomes of the eligible studies were varied. The studies focusing on operations management showed mixed results. The studies focusing on high acuity services showed reduction in over-utilization, with the caveat that the studies were of variable quality and had varying risk of bias.

Conclusions:

Our results show that applied healthcare predictive analytics research is in its early stages. Our results also indicate that there is insufficient high-quality evidence to assert that predictive analytics has significantly impacted clinical outcomes, clinical decision-making and organizational performance. Additional evidence-based information may be needed to hasten broader adoption and effectively leverage this capability to enhance all healthcare domains.
**Data Management, Sharing, and Citation Strategies in The Cancer Imaging Archive (TCIA)**

Tracy S. Nolan, BE¹, Justin S. Kirby, BS², Fred W. Prior, Ph.D.¹

¹Biomedical Informatics, University of Arkansas for Medical Sciences, Little Rock, AR
²Frederick National Laboratory for Cancer Research, Leidos Biomedical Research, Inc., Frederick, MD

**Introduction:** Big Data are expensive to collect, and complex to share

Imaging is a key component of precision medicine research. Unlike genomics and proteomics it is non-invasive, capturing the whole tumor and its spatial environment. Cancer imaging is expensive to acquire in time, cost, specialized hardware, and patient population recruitment. Sharing such data is not easy due to government regulations regarding patient privacy. Retaining key scientific information for downstream research while removing protected health information is particularly challenging. So it is extremely valuable that The Cancer Imaging Archive (TCIA)¹ ² extends global sharing and re-use of such data for secondary research aims and reproducibility studies.

**Data Management: Biomedical data are heterogeneous**

Biomedical data that have been aggregated and curated on standardized criteria into public repositories produce the widest reuse potential. Each hospital department uses their own local terminology, scan protocols, and different vendor hardware. Heterogeneity in clinical and research data acquisition pose significant curation challenges to research data centralization, compatibility, searchability, security, quality, and ease of use up- and down-stream. TCIA addresses these problems and provides centralized access through a trusted, curated repository. For example, a research question needing slides, CT scans, and drug response from a dozen different clinical trial sites can be accomplished via a single download request through TCIA.

**Sharing: The Cancer Imaging Archive (TCIA) is an open data repository federated with others**

TCIA hosts mostly DICOM³ format imaging of human cancer, often with supporting data (linked out⁴ or attached via wiki) used to generate the results that the authors who share their research have published. TCIA shares “Primary Data” (e.g. original radiological scans, clinical and imaging protocols, segmentations, software), and encourages users to recombine data (to create, cite, share new data groupings) as “Analysis Results.” These Analysis Results, created with diverse commercial and open source tools, have myriad file formats too. TCIA and other open biomedical repositories (genomics, pathology) expend considerable effort to use common terminology and research identifiers across data types. Data sharing and fusion across modalities increase overall research data quality and guide policy; helping the community distribute, test, and converge upon standards as tools mature.

**Citation: The Digital Object Identifier (DOI) as a citation tool both upstream and downstream**

Granularity of data citation through persistent DOIs in TCIA can be variably tailored to refer to articles, datasets, feature sets, or a wiki page of project details. The data authors, PubMed, TCIA, or granting agencies can use DOI tracking analytics to show that data mining is ongoing by the primary investigator and others, proving the initial grant investment was good – or that data are not being used, so resources can be optimized.

**Conclusion**

Federated, curated, public open data archives like TCIA support increased hypothesis complexity, then also the validity and value of results. Original research resources are extended by new analyses. However, it is essential to quantify the quality and completeness of source data, which DOIs can help facilitate. Dataset publishers open doors for researchers, clinicians, and the public to stretch the available technologies, with academic credit to data and process sources, to address precision medicine responsibly.

**References**

1. www.cancerimagingarchive.net
4. e.g. http://gdc.cancer.gov or http://github.com
The Depth of Historical Electronic Information Seeking by ICU Clinicians

Matthew E. Nolan, MD¹, Rodrigo Cartin-Ceba, MD, MSc², Pablo Moreno-Franco, MD³, Brian Pickering, MD¹, Vitaly Herasevich, MD, PhD¹
Mayo Clinic, Rochester, Minnesota; Mayo Clinic, Scottsdale, Arizona; Mayo Clinic, Jacksonville, Florida

Abstract
The electronic information-seeking behaviors of intensive care unit (ICU) clinicians are incompletely understood. We surveyed 234 ICU clinicians about the historical depth of their ‘electronic chart review’ when admitting new patients, receiving 155 responses (66% response rate). 2/3 of clinicians estimate reviewing data at least 3 years back, highlighting the need for thoughtful design and intelligent filtering techniques within critical care information systems to provide ready access to such large volumes of historical data.

Introduction
Electronic chart review (or “chart biopsy”) is a prevalent and important task performed by clinicians in order to build a patient narrative, but there are few studies evaluating this information-seeking behavior. Our research group is developing an electronic tool that can display longitudinal health data in a visual narrative for medical intensive care unit (ICU) patients, which requires an in-depth understanding of chart review habits.

Methods
We surveyed medical ICU clinicians at three Mayo Clinic sites, including rotating residents, fellows, attending physicians, and advanced practice providers. Among other questions, clinicians were asked to estimate how far back into the medical record they browse when admitting a new ICU patient.

Results
155/234 ICU clinicians responded (66% response rate). Sixty-six percent of clinicians reported reviewing data at least 3 years back, and over 1/3 of clinicians review data from 10 years or more (Figure: Estimated Depth of Historical Electronic Chart Review by ICU Clinicians Admitting New Patients)

Conclusion
Most ICU clinicians require ready access to the majority of the historical medical record for new patients, which emphasizes both the importance and simultaneous difficulty in providing and displaying this information succinctly. Critical care information systems require thoughtful design and intelligent filtering to address this information need.

References
The Relationship Between “Reason for Exam”, Radiologist’ Perceived Unexpected Effort, and the Time to Complete Interpretation

Daniel T. Nystrom, MS1, Heidi Kramer, PhD1, Marta E. Heilbrun, MD, MS2, and Charlene Weir, PhD, RN1

1Department of Biomedical Informatics, University of Utah, Salt Lake City, UT; 2Department of Radiology and Imaging Sciences, University of Utah, Salt Lake City, UT

Introduction
The reason for exam provides key clinical information to radiologists that support an accurate determination and efficient search through radiologic images. Despite the importance of the reason for exam, little research has been done to consider how aspects of the reason for exam impact radiologist workflow. The purpose of this pilot study is to assess the relationship between the clarity of the reason for exam, the amount of information, and the time to complete the exam, and radiologist’ interpretative effort.

Method
We observed 4 radiologists at a local academic health care system perform multiple image interpretations during 1-hour sessions and across 51 image interpretations (ranging from 8 – 22 interpretations). The amount of time to perform each interpretation was collected. In addition, radiologists were asked to make 3 judgements about each interpretation; 1) the clarity of the reason for exam using a 7-point Likert scale ranging from 1= “Poor” to 7= “Excellent.”; 2) the amount of unexpected effort they experienced while performing the exam ranging from 1= “None” to 7= “Exceptional.”; and 3) the amount of information available in the reason for exam from 1= “None” to 7= “Sufficient.” The correlation between clarity of the reason for exam, the amount of information available in the reason for exam, unexpected effort, and time to complete the exam was assessed.

Results
A weak negative correlation (r = -0.33) was observed between clarity of the reason for exam and perceived unexpected effort (Table 1). Unexpected effort was positively correlated to the amount of time spent on the exam (r = 0.52). No correlation between clarity of the reason for exam and time to complete the exam was observed (r = -0.03). The amount of information in the reason for exam was correlated to unexpected effort (r = -0.29) and the clarity of the reason for exam (r = 0.92), but not to time (r = 0.06).

Table 1. Mean ± (SD) and results of Pearson correlations for radiologist ratings and time to complete exam

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean ± (SD)</th>
<th>Variable Correlations</th>
<th>Pearson r</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clarity of reason for exam</td>
<td>4.10 ± (2.12)</td>
<td>Clarity X Effort</td>
<td>-0.33 (p&lt;.05)</td>
</tr>
<tr>
<td>Unexpected effort</td>
<td>2.00 ± (1.55)</td>
<td>Effort X Time</td>
<td>0.52 (p&lt;.001)</td>
</tr>
<tr>
<td>Time to complete exam (in seconds)</td>
<td>217.49 ± (244.94)</td>
<td>Clarity X Time</td>
<td>-0.03 (p&gt;0.05)</td>
</tr>
<tr>
<td>Amount of information</td>
<td>3.79 ± (2.30)</td>
<td>Amount of info X Effort</td>
<td>-0.29 (p&lt;0.05)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Amount of info X Time</td>
<td>0.06 (p&gt;0.05)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Amount of info X Clarity</td>
<td>0.92 (p&lt;0.01)</td>
</tr>
</tbody>
</table>

Conclusion
Results of this study indicate that as the clarity of the reason for exam and the amount of information in the exam decreases, the perceived effort to complete the exam increases. However, neither clarity of the reason for exam nor the amount of information in the exam was significantly correlated to the amount of time radiologists took to complete their interpretation. Based on these results, we are conducting further research to determine the strategies that radiologists use to regulate their workflow to compensate for poor reasons for exam.

References
Using Point of Care Systems to Deliver Patient-Centric Care: Lessons from an Australian Case Study

Louise O’Connor1, Peter Haddad 1, 2, Nilmini Wickramasinghe 1, 2

1Epworth HealthCare, Victoria, Australia, 2Faculty of Health, Deakin University, Melbourne, Victoria, Australia

Abstract
Ensuring a high level of patient-centric care is an integral consideration globally for healthcare delivery today. Evidence from the literature underscores the benefits of such an approach for enhancing the quality of care and reducing associated costs. This study investigates the use of a point of care system’s ability to deliver high value patient-centric care across a private not-for-profit tertiary healthcare group in Australia. The results show that the examined system supports patient-centric care by facilitating the fostering of good relationships between patients and their clinicians, as well as building a cultural shift towards patient-centric care as part of standard practice.

Introduction
Patient-centric care provides customised precision care to each patient’s individual needs and requirements [1]. The literature identifies several components of patient-centric care delivery, including: patient participation and involvement and the context where care is delivered [2]. This study focuses on the use of the OneView Point of Care system at an Australian not-for-profit tertiary healthcare group and its role in building a strong patient-centric healthcare system. This system is a bedside computerised information system whose terminals provide patients with a range of entertainment, education and information services, and clinicians with a range of integrated clinical applications including electronic prescribing and administration, patient results, and electronic nurse rounding. In addition, it serves non-clinical functions to cover the food and environmental needs for patients.

Methods
This exploratory multi-site, single case study [3] adopts a mixed methods approach. Data collection includes a series of 15 semi-structured interviews carried out across the various sites of the selected case over 12 months. The interviews targeted the users of the system, including clinicians (nurses and allied health), and other staff (food and environmental services). The interviews were transcribed and analysed using thematic analysis [4]. Data triangulation was achieved by site visits and hermeneutic analysis of documents, including annual reports and various case reports.

Results
The results highlight various aspects of how the system supports patient-centric care as a senior executive describes: “the point-of-care system is about creating an environment of patient-centric care, putting the staff and the doctors back where they want to be, which is in the patient's room”. In addition, as the system is at the bedside, it requires the nursing staff to visit their designated wards more often, which helps better monitoring of their patients and prevents incidents as another interviewee explains: “If staff are not in the patient's room, that's when bad things happen. That's when patients feel dissatisfied, that's when the doctor doesn't know what's going on, etc.” The system has also enabled a cultural shift so that being a patient-centric healthcare provider is now part of standard practice: “Culture is the most important thing in healthcare, because how you behave as a clinician and how you develop relationships with a patient - so your culture - is what drives your results as well”.

Conclusion
The OneView Point of Care system was found to clearly support the sustained delivery of patient-centric care. Critical aspects enabled and supported through the system include: patient participation and involvement, the relationship between the patient and the healthcare professional and the sharing of germane and pertinent information. In addition, a culture shift took place. Thus, the study is expected to highlight key aspects of enabling and supporting patient-centric care through technology implementation and adoption.

References
Validation of the Knowledge and Self-Management NOC Outcomes for Adults with Diabetes

Hyunkyoung Oh, PhD, RN1, Sue Moorhead, PhD, RN, FAAN2
1University of Wisconsin-Milwaukee, Milwaukee, WI; 2University of Iowa, Iowa City, IA

Introduction
Current health care focuses on preventive care to avoid the development of chronic diseases. Specifically, diabetes is one of the most significant chronic diseases due to its high prevalence and mortality. People with diabetes have to know how to self-manage their health conditions to promote, maintain, and restore their health. The Nursing Outcomes Classification (NOC) has been developed to evaluate and quantify the status of the patient and the effects of nursing interventions over time and across care settings1. NOC is one of recognized standardized nursing languages by American Nursing Association. The 5th edition of NOC has provided the two NOC outcomes for people with diabetes to evaluate their knowledge and self-management behaviors. First NOC outcome was Knowledge: Diabetes Management and its definition was “extent of understanding conveyed about diabetes, its treatment, and the prevention of complications”. This NOC outcome has the 36 indicators to evaluate knowledge status of patients. Second NOC outcome was Self-Management: Diabetes and its definition was “personal actions to manage diabetes, its treatment, and to prevent complications”. This NOC outcome has the 44 indicators to evaluate patients’ behaviors for self-management. However, these two NOC outcomes were not validated yet. Generally, validation of a measurement tool is required to provide trustworthy evidence2. As measurement tools, these two NOC outcomes need to be validated for accuracy, meaningfulness, and usefulness in clinical settings. The purpose of this study was to validate the knowledge and self-management NOC outcomes for adults with diabetes published in the fifth edition of NOC1.

Overview of methodology
A descriptive exploratory design using the Delphi technique was used to validate the two NOC outcomes and collect data. Two sets of experts were invited to participate in the review of the outcomes. The first expert group was related to standardized nursing languages and invited experts were members of the NANDA International or a fellow of the Center for Nursing Classification and Clinical Effectiveness. The second group was related to self-management and invited experts were members of Midwest Nursing Research Society. Descriptive statistics were used to determine the definition adequacy, clinical usefulness, and content similarity. The Outcome Content Validity (OCV) method was used to evaluate the content validity of each outcome and the indicators.

Results
A total of 16 nurse experts participated in this study. The mean experience in specialty of respondents was more than 20 years. Each definition adequacy of the two NOC outcomes was over 4.0 (perfect score is 5.0). This means that each definition was adequate to describe the outcomes. Clinical usefulness was also over 4.0. and this rating indicates that the measurement scales were easy to use in the clinical settings. The range of content validity of the two NOC outcomes was from 0.89 to 0.92 (perfect score is 1.0) using average OCV scores. According to OCV method3, over 0.80 means that this outcome is critical. The indicators of two NOC outcomes should evaluate consistent diabetic knowledge and behaviors. The invited experts reported that the content of this outcome pair was very similar each other (mean=4.0, SD=0.516).

Conclusion
The two NOC outcomes focused on knowledge and self-management for diabetic patients were validated, and the results provided strong validity. By using validated NOC outcomes, nurses who take care of patients with diabetes can evaluate patient outcomes effectively and determine the effect of nursing interventions accurately. Additionally, nursing students can understand and utilize the outcomes to specific populations based on linguistic accuracy of the validated outcomes. In order to obtain more accurate reliability and validity, these NOC outcomes should be revalidated after applying them to patients in the various clinical settings.

References
Parsing MetaMap Files in Hadoop

Amy L. Olex, M.S., Alberto Cano, PhD, Bridget McInnes, PhD
Virginia Commonwealth University, Richmond, Virginia, USA

1 Problem Description
Processing and interpreting text is natural to humans but extraordinarily difficult for machines. Much of how a person interprets texts is based on prior knowledge of concepts and associations between concepts. The UMLS::Association Perl package aids in interpreting biomedical texts by quantifying levels of association between UMLS (Unified Medical Language System) concepts. The UMLS Metathesaurus contains over 2.5 million names representing over 900,000 concepts identified by Concept Unique Identifiers (CUIs). MetaMap parses biomedical and clinical texts and links them to UMLS CUIs, and this output is processed by UMLS::Association to collect CUI bigram frequencies and quantify CUI associations. The collection of CUI bigram frequencies provides a background of knowledge from which to extract associations between biomedical concepts for further processing. However, parsing this information is time consuming, as there are hundreds of large (several gigabytes) MetaMap files for each processed text corpus.

UMLS::Association relies on the collection of CUI bigram frequencies identified by the CUICollector module, which parses the nested MetaMap structure one utterance at a time. As the CUI bigrams are parsed they are stored in a nested hash of hashes that is periodically dumped to a MySQL database. This serial implementation has several limitations: 1) Perl code is serial and not parallelizable due to Perl’s limitations in sharing nested hashes across threads for synchronization. 2) The bigram hash can become very large, very fast and fills up program memory, which forces the program to periodically write results to a MySQL database, introducing additional latency.

2 Purpose of Project
CUICollectorMapReduce takes advantage of the Hadoop MapReduce framework to overcome serial implementation limits. MapReduce algorithms are well suited to parsing text and counting occurrences. In addition, the Hadoop environment reads and writes all results directly to disk thereby resolving the memory issue. MapReduce was chosen over other distributed processing technologies (e.g. SPARK, Flink) due to system limitations in memory, a large data set size, and the batch nature of the problem that requires accessing the data only once. CUICollectorMapReduce has two modes: cui and article. In cui mode MetaMap output is parsed directly, one utterance at a time, and duplicates the results of the serial implementation. In article mode CUICollector concatenates utterances to allow the processing of an entire PubMed citation, which includes collection of CUI bigrams that cross utterances.

3 Results
Evaluation was conducted on the MetaMap 2015 MEDLINE Baseline, a 132GB compressed dataset of biomedical citations. The efficient implementation in Hadoop resulted in a 28x speedup in cui mode, reducing run time from 229 to 8 hours on a 4-core, single-node Hadoop system. Analysis of Hadoop output in comparison to that of the serial implementation revealed equivalent results. Required database operations in the serial implementation account for the majority of the serial running time. To more directly compare just the parsing speedup of Hadoop, all database operations were removed from the serial program, resulting in no program output. Hadoop still processed the data faster with a speedup of 2.8x when run on the full MetaMap 2015 MEDLINE Baseline compared to the modified serial implementation.

4 Contributions
The contributions of CUICollectorMapReduce are: 1) Parallel processing of MetaMap files with seamless scalability. 2) Eliminating the need to periodically write to a MySQL database by utilizing Hadoop’s framework where all intermediate and final results are written and read directly from disk. 3) Algorithm improvements allow for a full PubMed citation to be processed at once instead of one utterance at a time. In conclusion, implementing CUICollector in the Hadoop MapReduce environment provided significant speedup as well as additional flexibility in the type of data being processed. The Hadoop implementation also allows parsing of larger datasets that were not feasible using the serial module, thus, opening up additional avenues of research.
Medication Errors Following a Home-Grown Electronic Health Record Implementation: Identifying and Overcoming Unanticipated Challenges

Waldo Ortega, MD1, Carolina Ibañez, RN1, Marcelo Lopetegui, MD, MS1,2
1 Departamento de Informática Biomédica, Facultad de Medicina, Clínica Alemana Universidad del Desarrollo, Santiago, Chile
2 Centro de Innovación e Investigación en Informática Biomédica, Facultad de Medicina, Clínica Alemana Universidad del Desarrollo, Santiago, Chile

Introduction
Medication errors are one of the most important topics regarding patient care safety, reaching up to 1,000,000 serious medication errors per year in American hospitals, and producing 7,000 deaths annually. Although the introduction of computerized physician order entry (CPOE) has decreased medication errors, there is sparse evidence suggesting it could add new adverse events. Fifteen years after the publication of “To err is human”, The National Patient Safety Foundation (NPSF) reports that not significant changes have been made since.

In an attempt to improve patient safety and quality of care, our institution decided to replace its aging commercial electronic health record with a home-grown hospital information system, with modules specifically developed to support a safer medication process management. This study aims to identify and describe the impact of the implementation of this new system, and report the remaining gaps and opportunities to improve patient safety.

Methods
We analyzed a self-report registry of medication errors maintained by nurses at a large private hospital in Chile in order to identify existing errors, their source and moment of occurrence. Then, we assessed the perceived impact of the new hospital information system on medication errors by conducting semi-structured interviews with key stakeholders using the Reason Model. We performed content analysis of the interviews, focusing on errors solved and potential new errors introduced by the system. This information was contrasted between the different interviewees and with published evidence.

Results
The new EHR solved most of the problems identified at baseline. However, it introduced opportunities for new unexpected errors. Among newly introduced opportunity for errors, we identified two main categories: 1. Errors related to the use of the Chilean standardized pharmacological nomenclature, which differs from international standards; and 2. errors related to changes in clinical workflows. Most of them occurred during prescription or medication administration. The knowledge gathered is being used to develop updates to the involved modules and maintain a continuous patient safety assessment cycle.

Discussion
Medication errors evaluation in the era of EHRs requires a systematic approach. International guidelines and recommendations should be followed by all major EHR vendors as well as home-grown solutions in order to avoid repeating mistakes that have been already solved elsewhere. We recognize the value of adopting international instruments to conduct continuous internal assessments and benchmarking against similar healthcare institutions in Latin America, and reporting opportunities to clinicians. (Leapfrog evaluation, Medication Safety. http://www.leapfroggroup.org/).

References.
Information and Communication Technologies Adoption by Healthcare Professionals in Brazil
Cristina Lucia F. Ortolani, PhD¹; Fabio Senne¹; Alexandre Barbosa, PhD²; Heimar F. Marin, PhD, FACMI³
¹ Universidade Federal de São Paulo, Brazil ² Universidade Paulista, São Paulo, Brazil, ³Regional Center for Studies on the Development of the Information Society – Cetic.br, São Paulo, Brazil.

Abstract
The range of possible applications of ICTs in the healthcare system is enormous. The technology has progressed significantly and ICT implementation is expected to result in higher quality and more efficient healthcare system. The main objective of this research is to share information about ICT adoption and use by health care professionals in Brazil according to OECD indicators, considering the current stage of technological infrastructure implementation, including the availability of ICT-based services in Brazilian healthcare facilities.

Introduction
Information and communication technologies (ICTs) are defined as digital and analogue technologies that facilitate the capturing, processing, storage and exchange of information via electronic communication¹. ICTs used in the health sector have well-known advantages as they have the potential to improve information management and access to health services. They are also effective for cost containment and promotion of education among patients and health professionals⁴. However, implementation of ICTs remains difficult and involves changes at different levels: patients, healthcare organizations and healthcare professionals. The aim of this study is to understand the stage of ICTs adoption in Brazilian healthcare facilities and their appropriation by healthcare professionals.

Methods
This descriptive study, in which the researcher interacted with the participants, was dedicated to collect the necessary information through surveys. The sample was comprised by 3,582 health facilities in 350 Brazilian cities selected by probability-proportional-to-size sampling, using the square root of the number of professionals in each facility. Data collection was performed through two structured questionnaires, one applied to administrative professionals (n=1,685), and another to healthcare professionals (n=1,484 physicians and n=2,696 nurses). The physicians and nurses answered questions about their own routines as healthcare professionals, their profiles, as well as ICT access, use and appropriation.

Results
Among those with computer access at the healthcare facilities, 59% of nurses used computers daily when seeing patients, a percentage that reached 70% among physicians. Regarding computer use in other work activities, meaning those that do not involve direct contact with the patients, percentages were higher: 84% of physicians and 75% of nurses with computer access used them daily. The main problem with the use of electronic information by professionals when accessing the patient electronic clinical data was the lack of availability at the facilities.

Conclusion
Health professionals referred to registered data and used system functionalities when those were available at the facilities. This corroborates the idea that professionals’ compliance to the use of technological solutions is not a hindrance to implementation. Instead, the unavailability of technological applications and functionalities in facilities is still one of main challenges to adopting ICT-based solutions.

References

2127
Sepsis Risk Stratification among the CMS Oncology Care Payment Model Population

Benjamin E Orwoll, MD1, Konrad Dobbertin, MPH1, Michael Savin, MD1, David A. Dorr, MD1, Peter Graven, PhD1
1Oregon Health & Science University, Portland, Oregon, USA

Introduction

In 2016 the Centers for Medicare and Medicaid Services (CMS) initiated the Oncology Care Model (OCM), an alternative payment model for chemotherapy care. Participating practices are incentivized to provide more efficient, high quality coordinated care at reduced cost, with shared financial risk. Incentives are based on patient outcomes and in overall Medicare costs during each 6-month episode. Historical claims data provided by CMS on patients from our institution's practice identified sepsis admissions, both at our institution and outside facilities, as a major source of potentially avoidable cost. Early identification and monitoring of patients at high risk for this condition could lead to better patient health and lower care costs. Community acquired sepsis risk scores exist, but may not represent this high-risk population. We sought to develop a predictive model to identify OCM patients at high risk for sepsis hospitalizations.

Methods

Patients were identified by CMS as those who would have been assigned to our institution between 1/2012-6/2015. We used Medicare claims data from that period provided by CMS to identify sepsis hospitalizations at any hospital based on MS-DRG and ICD-9 codes for sepsis, septicemia, and septic shock. Data from the local EHR including demographics, problem lists, medical history, labs, and medication orders were evaluated for inclusion in the predictive model, based on literature and clinical plausibility. We used survival-time models with Weibull distribution and stepwise elimination (threshold \( p \geq 0.05 \)) to calculate the sepsis hospitalization risk (SHR).

Results

1501 patients had care episodes during the study period. Characteristics are shown in Table 1. There were 123 sepsis hospitalizations (8%), with only 41% of those occurring at our institution. Mortality was 41% for the entire sample, but 74% of patients with sepsis died by the end of the study period. We identified 12 significant factors (Table 2) that formed the risk prediction model; while some of these were static, such as cancer type, others were time varying. For prediction of sepsis within 30 days our model had an area under the receiver-operating-characteristic curve of 0.73 (95% CI 0.70-0.76)

Conclusions

We developed a risk prediction model for SHR among high risk chemotherapy patients in the OCM. Our model uses data that is available up to date within the local EHR to predict events at all locations. We plan to validate the model using future CMS claims data, which are only available retrospectively. The model could potentially be used to identify high risk subgroups who might benefit from more intensive monitoring and care coordination, and a 30 day outlook provides a manageable time frame and cohort size. The contribution of both static and time-varying predictors suggests that monitoring and decision support programs may need both initial and ongoing evaluations for those at highest risk.

References


Table 1

<table>
<thead>
<tr>
<th>Sepsis</th>
<th>Yes (n=123)</th>
<th>No (n=1378)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age &gt;70 (%)</td>
<td>59</td>
<td>64</td>
</tr>
<tr>
<td>Male (%)</td>
<td>58</td>
<td>48</td>
</tr>
<tr>
<td>White (%)</td>
<td>87</td>
<td>90</td>
</tr>
<tr>
<td>Top Cancer Type (%)</td>
<td>Hematologic (25)</td>
<td>Breast (25)</td>
</tr>
</tbody>
</table>

Table 2

<table>
<thead>
<tr>
<th>Static Predictor</th>
<th>Hazard Ratio* (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cancer Type</td>
<td></td>
</tr>
<tr>
<td>Bladder</td>
<td>3.2 (1.5-6.7)</td>
</tr>
<tr>
<td>ENT/Lung</td>
<td>2.6 (1.5-4.7)</td>
</tr>
<tr>
<td>GI</td>
<td>2.0 (1.2-3.5)</td>
</tr>
<tr>
<td>Hematologic</td>
<td>2.2 (1.3-3.8)</td>
</tr>
<tr>
<td>Heart Failure</td>
<td>2.1 (1.2-3.5)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Time-Varying Predictor</th>
<th>Hazard Ratio* (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hospitalization Past 60d</td>
<td>3.0 (1.7-5.4)</td>
</tr>
<tr>
<td>Lab</td>
<td></td>
</tr>
<tr>
<td>Blasts (any)</td>
<td>4.1 (1.4-12.4)</td>
</tr>
<tr>
<td>Lymphocytes (&lt;500)</td>
<td>3.0 (1.8-4.9)</td>
</tr>
<tr>
<td>Neutrophils (&gt;5000)</td>
<td>2.7 (1.9-4)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Hazard Ratio* (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anticoagulants</td>
<td>1.9 (1.2-2.9)</td>
</tr>
<tr>
<td>Antineoplastics</td>
<td>0.5 (0.3-0.7)</td>
</tr>
<tr>
<td>Immunologics</td>
<td>0.4 (0.2-0.8)</td>
</tr>
</tbody>
</table>

* all \( p \)-values <0.05
Implementing a Real-Time Location System (RTLS) at an Ambulatory Oncology Surgical Center

Aaron Ostrovsky, MPH, Alex Stevens, MA, ED, Justin McManus, Marcia Levine, MSN, RN, NE-BC, Brett Simon, MD, PhD, Peter Stetson, MD, MA, Daniel Stein, MD, PhD
Division of Health Informatics, Department of Nursing, Josie Robertson Surgery Center; Memorial Sloan Kettering, New York, NY

Introduction

Real-Time Location System (RTLS) technology, which involves the automated, collection of location-based data from a tagged entity, has been used in healthcare for a variety of use cases, such as: asset tracking, patient safety, workflow optimization, and environmental monitoring. Asset tracking has offered the greatest return on investment; however, recently the desire to focus on patient flow and workflow optimization has increased. This has been reported in the outpatient setting for monitoring room allocation or improving the efficiency of patient and staff flow. Here we present a novel use case of using RTLS to optimize care coordination and patient flow at a new ambulatory oncology surgical center, the Josie Robertson Surgery Center (JRSC).

Methods

The system implemented at the JRSC was Versus RTLS (Versus Technology, Inc., Traverse, MI). This implementation involved the installation and deployment of 768 wired-infrared (IR) sensors, 677 staff badges and 244 asset tags. The staff, patient and caregiver badges utilize both IR and radio frequency (RF) whereas the asset tags use IR and Wi-Fi locating. Interfaces with several other systems were built, including two EHRs, the nurse call system, and our in-house patient engagement system. The RTLS client was installed on 450 clinical and non-clinical workstations. Displays with live facility maps were mounted throughout the facility, so clinicians and staff could determine location of patients and care team members and monitor the status of operating and patient rooms.

Results

Since the opening of JRSC in January, 2016, we had an average of 360 unique RTLS users per month. Of that user population, 15% were perianesthesia nurses, 10% were nurse anesthetists; and the remaining 75% were other clinical and non-clinical groups. There were 7,677 patients who received an RTLS badge. Implementation of the RTLS system has resulted in reduction of error-prone documentation through the automation of clinical time stamps and has augmented communication through location-based awareness and presence-based nurse call cancellation. Bed management was enhanced by using the RTLS system for Pre-op and Post-op bed assignment. For patients and their caregivers, RTLS offered care team awareness through real-time staff and role identification, a novel patient metric for ambulation post surgery, and a family waiting view that provided updates and visual cues to family members.

Discussion

RTLS in this setting has the capability to save clinicians time, improve accuracy of data, enable care team awareness, solve unmet workflow needs, and provide new data to clinicians and patients that could impact care decisions and overall quality of the surgical experience. However, there are technology and process limitations such as badge-wearing compliance, high implementation costs, and the lack of robust native analytical tools. Future work and recommendations include integration with staff communication systems, studies to formally evaluate clinician and patient experience with RTLS, and a retrospective analysis to optimize efficiency within the perioperative space.

References

Innovative Approaches in Informatics Education

Bunyamin Ozaydin, PhD\textsuperscript{1}, Sue Feldman, RN, MEd, PhD\textsuperscript{1}, Laura J McNeill, MS\textsuperscript{1}, Amanda Dorsey, MSHI\textsuperscript{1}, Meg Bruck, MSHI\textsuperscript{1}, Eta S. Berner, EdD\textsuperscript{1}

\textsuperscript{1}University of Alabama at Birmingham, Birmingham, AL, USA

\section*{Introduction}

After revising University of Alabama at Birmingham’s (UAB) online Master of Science in Health Informatics (MSHI) curriculum to a core/track model in 2014\textsuperscript{1}, a decision was made to take relevant portions of it to offer as an online graduate-level credit certificate in clinical informatics. Because the proposed certificate courses are drawn from the existing core courses in the MSHI program, they, like the UAB AMIA 10x10 program\textsuperscript{2}, can be applied to the full master’s degree in Health Informatics. This poster describes both how the different programs interrelate as well as the innovative instructional methods that have been employed. Live demonstrations of the innovative methods will be shown.

\section*{Methods}

Development of the Graduate Programs in Health Informatics at UAB has taken an approach in which all other programs are built upon the solid foundation of our MSHI core program. The four courses in the certificate program, which include one course that is also part of the AMIA 10 x 10 program, provide basic information for someone wanting to learn about health informatics, but who may not be degree-seeking. It is important to note that while the content of the same course offered in different programs contains the same content, the delivery modality was modified to accommodate the type of student enrolling in the AMIA 10x10 program or Clinical Informatics Certificate. The courses that are included in the certificate program have been redesigned to contain a set of self-paced online modules, 2-4 units/weeks long, within specified module release and end dates. Students can apply the single AMIA 10 x10 course to the certificate or directly to the MSHI program, and all certificate courses can be used to exempt courses in the MSHI program.

For the new certificate program, we added enhancements that can be used across all programs. These enhancements include video recording of lectures using a variety of engaging technologies, and designing the videos so they can be used for other online courses. This approach consists of: (a) short videos each containing material related to a single topic, (b) avoidance of references to the specific course, module, or unit, and (c) separate content-specific introductory and/or summary videos for the specific unit to put the modular videos and the units within them in context. By converting longer lecture videos into this shorter format, we envision a modular video repository, whereby each video has a description and a set of keywords to allow instructors to reuse them in other courses where that specific content may be needed.

In order to foster more engagement than a typical video lecture, we utilized interactive strategies for online presentation of the material. For example, to provide more of an in-class instruction feeling, we used the Smart Podium technology for the instructor to interact with the content by drawing diagrams and charts and/or pointing to certain portions on them, instead of talking over a static slide. We also developed online orientation material where students answer a series of multiple-choice questions and receive personalized audio feedback for their response choice.

\section*{Results and Evaluation}

Evaluation of the content and format of the course (60\% response rate) showed that students were generally satisfied with course content and teaching methods. One of the courses used text instead of video module introductions and students indicated that they liked the video introductions better than the text. Analysis of number of plays and average duration of the videos on a technical course revealed that students accessed more often and for a longer duration videos that included hands on practice of the concepts and that were more relevant to case study assignments.

\section*{References}

Portable Precision Phenotype Algorithm for Chronic Rhinosinusitis

Jennifer A. Pacheco\textsuperscript{1}, Agnes S. Sundaresan, MD, MPH\textsuperscript{2}, Kenneth M. Borthwick\textsuperscript{2}, Sergio E. Chiarella, MD\textsuperscript{1}, David T. Coleman, MD\textsuperscript{1}, Andy Y. Wu\textsuperscript{1}, Abel N. Kho, MS, MD\textsuperscript{1} M. Geoffrey Hayes, PhD\textsuperscript{1}, Marc S. Williams, MD\textsuperscript{2} \\
\textsuperscript{1}Northwestern University, Chicago, IL; \textsuperscript{2}Geisinger Health System, Danville, PA

Introduction

Extracting patient phenotypes from the Electronic Health Record (EHR) has become very popular; however, it is difficult to extract more complex phenotypes, and more precise sub-phenotypes, especially in a format that is portable and executable across sites. Yet, it is important to be able to do so for precision medicine, as treatment can differ depending on the sub-phenotype. This is true for sub-phenotypes of chronic rhinosinusitis (CRS), with and without nasal polyps (NP), CRSwNP and CRSsNP, which are even more precise than chronic versus acute rhinosinusitis.

Methods

Previously Northwestern University (NU) published an algorithm for extracting CRS cases and controls from the EHR\textsuperscript{1}. For the electronic Medical Records and Genomics (eMERGE, gwas.org) project, we modified that algorithm to be portable and executable across sites, and more precise, by: adding International Classification of Diseases (ICD)-10 diagnosis codes, differentiating between CRSwNP and CRSsSNP, and to have more stringent controls (Figure 1). Another eMERGE site, Geisinger Health System (GHS), created a portable executable version of the phenotype algorithm, using Konstanz Information Miner Analytics Platform (knime.org), which NU was then able to easily modify to complete the new algorithm and execute. These 2 sites conducted blinded chart reviews of patients selected by the new algorithm, and a third eMERGE site is doing the same, to assess the specificity of our new algorithm.

Results and Discussion and Future Work

The new algorithm found 226 cases, 78 with NPs, and 1,485 controls at NU, out of 6,172 eMERGE subjects; plus 214 cases, 78 with NPs, and 2,057 controls at GHS, out of 12,295 eMERGE subjects. Compared to our previous algorithm, the positive predictive value (PPV) based on chart review remained relatively high overall for cases and controls, with a combined PPV of 0.87 for all cases (0.9 at NU for 200 cases reviewed, and 0.77 at GHS for 53 cases reviewed), and combined PPVs of 0.66 and 0.81 for CRSwNP and CRSsNP (both 0.7 at NU, and 0.62 and 0.85 at Geisinger), which should be sufficient to conduct a genome-wide association study of our sub-phenotypes with all eMERGE sites. However, to be more thorough, the difference in case PPVs between the 2 sites is being examined and a third eMERGE site is conducting reviews. For controls, PPV was 1.0 at NU, and 0.97 at GHS (out of 100 and 29 controls reviewed).

Conclusion

Our initial results show that it is possible to detect and automatically extract patients with precise phenotypes from the EHR, across multiple sites with different EHRs, even though it is not perfect nor simple.

References

Towards Spatial Analysis of Opioid Abuse Using Twitter

Sheetal Pandrekar, MS, Manish Valakonda, MS, Xin Chen, MS, Fusheng Wang, PhD
Stony Brook University, Stony Brook, NY

Introduction
From 2000 to 2014, the deaths from opioid overdose in United States, has seen a 200% increase in the rate. Combating opioid epidemic has become a high priority for both the U.S. government and local governments. The social networking site, Twitter, is being increasingly explored as a source for understanding the changing attitudes towards use of opioids. One advantage of Twitter over traditional surveys is the immediate accessibility of data. Analyzing opioid-related tweets has the potential to reveal patterns of opioid abuse at a large scale, and provide insight to support prevention, treatment, recovery and enforcement. In our study, we first identify keywords related to opioids by employing a 3-step procedure in order to carefully select the keywords to alleviate noise. We perform unsupervised Topic Modeling to discover the topics of discussion in the tweets. We also perform a statewide spatial analysis of opioid-related tweets in the United States.

Methods
For the 3-step procedure to select relevant keywords, we use seed keywords from the previous work. For each of the seed keywords, we perform the following steps: Step 1- Collect all slang terms for the keyword from UrbanDictionary.com; Step 2- Individually search each slang term on Twitter; and Step 3- If more than 50% of the top 20 results are relevant to opioid, include the keyword to the final list. We have finalized a list of 20 keywords. Our current dataset has a total of 19872 tweets collected with Twitter’s Streaming API, which provides access to a random sample of 1% of tweets. Initially, we manually analyze 764 tweets and classify them as relevant or irrelevant to opioids. From the complete dataset, we identify the topics using unsupervised Latent Dirichlet Allocation (LDA).

Results

LDA provides the top ten words for each identified topic. The popular topics of discussion in our dataset identified with LDA are crisis, pain, Pharma, constipation, Obamacare, Trump, treatment and bill. We map the tweets using the locations and normalize the total count by the per state population. The top 10 states identified are Oregon, Washington, Massachusetts, Ohio, New York, Pennsylvania, Maryland, Tennessee, Vermont, Rhode Island.

Conclusion
We describe a 3-step procedure for generating keywords related to opioids abuse. Our initial results demonstrated that the new keywords are able to fetch tweets with less noise. Oregon and Washington ranks No. 1 and No. 2 on the number of opioid-related tweets, which could be due to their legal use of Marijuana. Ohio ranks No. 4, and it has second highest number of drug overdose deaths. Our future work includes collecting more tweets and identifying the types of users (e.g. Personal, Organization, etc.) in the Twitter dataset for more insightful analysis.

Acknowledgments
This work is supported in part by NSF ACI 1443054 and by NSF IIS 1350885.

References
Building Better Timeline Interactions for Patient Chart Reviews

Heekyong Park, Ph.D.1,2, Taowei David Wang, Ph.D.1,2,3, Vivian Gainer, M.S.3, Victor M. Castro, M.S.3, Shawn Murphy, M.D., Ph.D.1,2,3

1 Massachusetts General Hospital, Boston, MA, USA; 2 Harvard Medical School, Boston, MA, USA; 3 Partners HealthCare, Somerville, MA, USA

Introduction

Reviewing a patient’s medical history is at the heart of many medical informatics problems. Because the temporality of patient data is important, timeline visualizations are often used to help reviewers. However, naïve implementations of timelines can result in many interaction problems, undoing some of its benefits. We systematically examine our own implementation, generalize common problems, and present our solutions.

Methods

i2b21 has been a platform for developing new informatics workflows. Its Timeline plugin, based on LifeLines2, visualizes a patient’s medical observations. They are displayed as ticks, and details of observations are available with mouse hovers. However, feedback from chart reviewers suggests the following limitations:

1. Using classic zoom and pan to navigate is not ideal. First, they are slow. Secondly, they tend to lose context for the users. For example, after a sequence of zooms and pans, users lose their original place.
2. Ticks can partially or completely occlude others in the visualization. This impedes mapping of the right data, which might result in misinterpretation. In addition, users are forced to use the problematic zoom and pan interactions to get around this issue.
3. There are no visual cues to help keep track of which of the hundreds of observations are viewed or noteworthy.

We developed a visual-semantic layer that sits between mouse clicks and the display of information. The layer receives mouse position and dynamically detect occluded ticks. A list is then presented to display data of these ticks and allows users to traversal to nearby ticks. The occlusion detection and alternative navigational method enable overlapped ticks to be fully interrogated without zoom or pan while keeping users’ temporal context intact. Furthermore, we added a coloring system so that all the already-read observations are colored ‘gray.’ Users can mark interesting observations as ‘yellow.’ The indicators obviate the need to remember which ticks are read, unread, or noteworthy (Figure 1).

Discussion and Conclusion

Unlike many information visualization tasks, chart reviews require constant reading of details while maintaining context. Without specific support, timelines can be cumbersome to use. We present three problems of naïve implementations and propose our solutions. By no means these are the only solutions. Recent advances use a variety of techniques such as a mini timeline for overview to maintain context. We aim to explore these and other solutions and compare them with ours in the future.

References

Identifying Health Interests of the General Public using Reddit
Albert Park, PhD1, Mike Conway, PhD1
1Department of Biomedical Informatics, University of Utah, Salt Lake City, Utah

Introduction
The Health Interests of the General Public (HIGP) are constantly changing and accurately identifying the current HIGP can be a challenge. The general public, however, openly discusses their health interests in health focused communities like r/health, which is a discussion forum from a popular social media platform called Reddit. A popular platform like r/health has the potential to provide opportunities to improve public-health practice. In August 2016, r/health has been active for 8 years with 118,949 members. The members of this community discuss a wide variety of topics ranging from newly emerging topics like Zika to long-standing interests like HIV. In this pilot study, we use r/health as a proxy and aim to identify health topics that are most frequently discussed over the past 8 years.

Methods
1. Data Collection: we used the Reddit Application Programming Interface (API) to download 1,000 all time ‘top’ rated posts determined by members and their associated 48,544 comments made from October 2008 to August 2016.
2. Preprocessing: we preprocessed the dataset to remove common stop words, punctuations, and URLs as well as posts and comments with less than 5 words.
3. Identify Topics: we processed the content of these posts and comments and identified the main topics using a machine learning algorithm called LDA, a topic modelling algorithm. We considered each post and its associated comments as a single document. We extracted and included nouns from each document and then employed the Gensim implementation of LDA1 to identify 25 main topics from this community.
4. Visualization: we created a word cloud visualization to visually represent the importance of topics and their top 25 associated terms, using the weights that are determined by the LDA.
5. Validation: we manually identified main topics from the 100 ‘top’ rated posts and then compared the results against the automatically recognized topics.

The study was exempted from review by the University of Utah’s Institutional Review Board [IRB 00076188].

Results
According to the LDA and the visualization, the members of r/health discussed topics related to nutrition (e.g., food, diet, calorie), medicine (e.g., doctor, healthcare, cancer), money (e.g., insurance, money, claim), and scientific facts or news (e.g., studies, science, research) as shown in Figure 1. Using the manual method, we identified 14 topics (frequency): news (58), personal medical experience (7), expensive medical care (5), vaccine conspiracy (5), addiction (5), medical marijuana (4), public health (3), medical information (3), insurance (2), sex education (2), nutrition (2), advancement in medicine (2), obesity (1), and dental care (1). While the specific topics are similar to automatically identified topics, a majority of the 100 ‘top’ rated posts were started by sharing news concerning medicine or health (e.g., “Drug Goes From $13.50 a Tablet to $750, Overnight - The New York Times”), which shows the potential for using social media for identifying the current HIGP.

Conclusions
We have presented preliminary findings on developing an automated method to identify the current HIGP. Our initial results suggest that, on the basis of our analysis of r/health, we may be able to identify health and medicine related news that is of most interest to the general public. We also provide several interesting directions to future work to accurately track HIGP. Analysis of additional platforms, Facebook and Twitter for example, could bolster our findings.

Acknowledgments: This work was supported by the NIH-NLM T15 LM007124, R00LM011393 & K99LM011393

References

Figure 1. r/health word cloud
Explorative Analyses on Indexing OMOP based Clinical Datasets with DATS

Ethan Park; Diana Guijarro, BA; Imho Jang; Stephen Trac, BS; Nansu Zong, PhD; Elizabeth Bell, MPH; Hyeoneui Kim, RN, MPH, PhD

Department of Biomedical Informatics, University of California, San Diego, La Jolla, CA

Introduction and Background

bioCADDIE (biomedical and healthCare Data Discovery and Indexing Ecosystem, http://biocaddie.org) aims to index biomedical data to help researchers find them for large scale data analyses. Data Tag Suite (DATS) is the standardized metadata specification of bioCADDIE developed by incorporating existing metadata standards in biomedical research and high priority data search use cases (http://biorxiv.org/content/early/2017/01/25/103143). Although some of the use cases reflect the clinical and/or healthcare domains to some extent, they are relatively underrepresented in DATS compared to the content from omics research. Healthcare related content areas appearing in many use cases are mostly limited to the phenotypes represented by disease names, laboratory test results, medication names, and sample demographics. It is unclear whether these limited content areas will sufficiently support dataset search needs for various types of healthcare research. With the launch of the Precision Medicine Initiative (PMI), a need for indexing datasets that represent diverse content areas such as clinical, behavioral, and environmental information became a reality. To ensure the robustness of DATS in representing these content areas, we first evaluated how effectively DATS supports searching clinical datasets structured by the Observational Medical Outcome Partnership (OMOP) Common Data Model (CDM) and 100 dataset search scenarios. We selected the OMOP CDM since it is one of the most widely adopted methods of standardizing healthcare datasets.

Methods

We first conducted mapping between OMOP CDM and DATS at the property level. We also selected 100 non-omics observational and/or experimental studies published in clinical journals for the past 2 years and extracted dataset search use cases by asking ourselves “How would we search for a dataset to replicate this published study?” We then identified the information types required for dataset searches and how these types of key information would be stored in the OMOP CDM based structure. Finally, we evaluated the efficacy of DATS in supporting the 100 dataset search use cases by reviewing to what extent the CDM fields holding key information are mapped to the core DATS items (meaning, this information is always indexed and thus available for search).

Results

The majority of OMOP CDM’s components were mapped to Dimension of DATS. Dimension is defined as “a feature of an entity, i.e., an individual measurable property of the entity being observed,” and takes Entity-Attribute-Value format. For example, patient’s race information “Asian” is indexed to Dimension with its attributes Name and Value that take values of “race” and “Asian” respectively. Pertinent clinical information relevant to the standardized clinical data tables of the OMOP CDM was mostly mapped to Dimension. The topic areas covered by the 100 studies were effectiveness of a treatment/procedure (N=53), comparative effectiveness of treatments/procedures (N=29), diagnostic accuracy (N=5), and others (N=5, e.g., observation on clinical manifestation, workflow analysis, etc.). The information deemed most frequently used for dataset search is medical diagnosis (N=76) followed by the demographic information of patient populations (N=51), which were the clinical content areas already considered for DATS. Twenty-five use cases required rather complex data searches combining more than demographics and diagnosis. The full mapping and the use cases are available at https://tinyurl.com/zuwuz4b.

Discussion

We were able to map the tables and attributes of the OMOP CDM to DATS. However, we found potential challenges to searching healthcare datasets structured with OMOP CDM using the indexing DATS offers because searching for the key information frequently used for healthcare datasets requires searching through the massive information held in Dimension of DATS. More importantly, the dataset search requires evaluating the value of the attributes (e.g., age, gender, diagnosis name, etc.), which is set as optional information in DATS. We are currently discussing these challenges with the DATS developer team. This explorative analysis was done with the use cases that do not fully represent the diverse topic areas of healthcare research; we plan to continue the analysis with the dataset search use cases that represent additional topic areas.

Acknowledgment: This work was supported in part by grant U24AI117966-01 (NIH/BD2K)
Lessons Learned from the Conversion of MIMIC3 to the OHDSI Common Data Model

Joongheum Park MD¹, Liana Tascau MD², and Adler Perotte MD, MA²
¹Englewood Hospital and Medical Center, Englewood, New Jersey; ²Columbia University Medical Center, New York

Introduction

The format and schema of observational health databases can be quite variable, which makes the systematic analysis of disparate databases difficult. The Observational Health Data Science and Informatics (OHDSI) collaborative provides a Common Data Model (CDM) and standard vocabulary. These tools allow for the systematic analysis of disparate observational databases given that they have all been transformed to conform to the OHDSI CDM. MIMIC3 is the newest version of a popular critical care database that is freely available and quite comprehensive in terms of EHR recordings in ICU settings. Although the data model for MIMIC3 is comprehensive, it is not an easily understood analytic database using a well-controlled vocabulary. This work aims to create a standardized version of MIMIC data so that computational methods can be developed and compared based on an easily understood data model and this dataset could be explored with the tools developed by the OHDSI community. More specifically, our goal is to develop a PostgreSQL-based extract, transform, and load (ETL) program that generates a CDM v5-compatible database from MIMIC3.

Materials and Methods

MIMIC3 is the latest iteration of a relational database of critical care patients from Beth Israel Deaconess Medical Center¹. It is comprised of de-identified health data for over 58,000 hospital admissions for 38,645 adults and 7,875 neonates. Entities mapped to the OHDSI CDM were coded using concepts from the standard vocabulary. For example, condition occurrences were mapped to standard concepts from the SNOMED-CT vocabulary and medications were mapped to the RxNorm vocabulary (see http://www.ohdsi.org/web/athena/). PostgreSQL was chosen as the database of choice for this ETL such it could be performed with strictly open-source resources that are compatible with the existing tools for the OHDSI CDM.

Results

We mapped 11 tables from the MIMIC3 database to OHDSI CDM v5 tables. We were able to successfully map demographic, diagnostic, procedural and medication data to the OHDSI CDM. Important issues encountered during the translation include: (i) Representation of temporal granularity. Given that MIMIC3 encodes data from the critical care setting, data elements have precise timestamps and patients evolve on an intraday timescale. However, certain tables in the OHDSI CDM were only capable of representing time at the granularity of days. We modified the data model slightly to accommodate granular timestamps. (ii) Certain data element types in MIMIC3 are not represented with a controlled vocabulary. Therefore, their concept identifiers were assigned by the researchers manually (e.g. SOFA scoring, spontaneous tidal volume). When data elements were represented with a controlled vocabulary, the ETL procedure leveraged the mappings available in the OHDSI standard vocabulary for identifying the appropriate standard concept (e.g. mapping a Read code to a SNOMED-CT code for a condition occurrence). The CDM standardized health system tables and health economics data tables were not populated as relevant data was not available in MIMIC3. Unmapped MIMIC3 tables includes clinical notes except for discharge summaries, culture sensitivity data, procedural data with CPT codes, data regarding caregivers and discharge planning information.

Conclusion

Although there are challenges in converting a database such as MIMIC3 to the OHDSI CDM, it is possible with slight modifications. Our work with the OHDSI collaborative has led to consensus-driven incorporation of timestamps with higher temporal granularity in the OHDSI CDM. Future work will involve further refining this ETL through collaboration with the OHDSI community.

References


2136
Assessment of Laboratory System-Assigned LOINC Codes for Common Tests

Sharidan K. Parr, MD, MSCI\textsuperscript{1,2}, Shuanghui Luo, MS, PhD\textsuperscript{2}, Sina Madani, MD, PhD\textsuperscript{2}, S. Trent Rosenbloom, MD, MPH\textsuperscript{2}

\textsuperscript{1}Veterans Affairs Tennessee Valley Healthcare System, Nashville, TN; \textsuperscript{2}Vanderbilt University Medical Center, Nashville, TN

Introduction
To represent clinical data accurately and consistently across different healthcare institutions, any site-specific information must be represented by and mapped to a standardized terminology, such as The Logical Observation Identifiers Names and Codes (LOINC\textsuperscript{®}). However, mapping can be hindered by idiosyncratic, ambiguous local data representations. We hypothesized that by evaluating the relationships between local laboratory system-generated LOINC mappings and corresponding internal electronic health record data, we would be able to systematically identify examples of LOINC mapping inaccuracy amenable to remediation.

Methods
This project was conducted at Vanderbilt University Medical Center (VUMC). Laboratory results for a subset of 52 common blood chemistry, hematologic, and urine tests were collected from 2012 through 2016. Data was obtained from the Research Derivative, a research-ready data replica derived from VUMC’s clinical enterprise. In our initial assessment of LOINC accuracy, we examined aggregate laboratory results using reference range; median, minimum, and maximum result; and units of measurement. For each locally-defined laboratory test (e.g. hemoglobin), we assessed LOINC code accuracy using the Component, Property, Time Aspect, System, and Scale attributes of the assigned LOINC code. The LOINC Component identifies what is being measured; Property refers to the type of quantity measurement (e.g. mass concentration); Time Aspect distinguishes specimens collected at a moment in time or over a specified time interval; System refers to the sample from which a specimen was obtained (e.g. blood versus urine); and Scale specifies the unit of measurement (e.g. quantitative or ordinal).

We defined LOINC codes of uncertain validity as those having a local test name inconsistent with the LOINC Component; units of measurement inconsistent with the Property, Time Aspect, Units, or Scale of the associated LOINC code; and/or locally-reported specimen type differing from that of the assigned LOINC System. The LOINC codes of uncertain validity were further examined using manual chart review. For each locally-defined test, we also quantified the number of different LOINC codes assigned within each laboratory test, and the number of specimen types assigned within each laboratory test. We identified a subset of laboratory tests associated with multiple LOINC codes and/or multiple specimen types. For these records, we collected patient-level data for manual chart review. We also examined the relationship between the laboratory-generated specimen type and its relationship with the resulting LOINC code.

Results
We examined over 13 million laboratory results from the 52 selected tests. Forty-three tests were associated with a single, correct LOINC code, and one correct specimen type. One test was associated with multiple LOINC codes and multiple specimen types, resulting in one incorrect LOINC code due to an erroneous local specimen type. Two tests each mapped to two LOINC codes and one specimen type, and were associated with accurate LOINC codes. Intra-test LOINC code heterogeneity occurred when laboratory instruments with different LOINC mappings were used to analyze the same local test (e.g. ‘Glucose Urine’). Six tests mapped to a single LOINC code with multiple specimen types, due to mapping schema ambiguity. Within this group, two LOINC codes were inaccurate, resulting from incomplete local concept coverage.

Conclusion
We present a generalizable, systematic approach to evaluating LOINC code accuracy. Using this approach, we detected errors at both the mapping schema level and at the individual test coding level. At the mapping level, errors were either due to ambiguity/heterogeneity in the mapping schema, or incomplete coverage. We found no temporal trend. Optimizing LOINC code accuracy at our institution will require harmonizing the laboratory mapping schema with the locally-defined laboratory tests. This process will require reducing redundancy in the laboratory mapping schema, further examining the role of the technician and the laboratory device in generating mapping features, and improving concept coverage.
Validating Vector Representations of ICD Diagnoses Using Readmission Prediction and Dissimilarity Testing: A Pilot Study

Matthew Parry, DO, Neris M. Nieves-Robbins, MD, Keith Solveson, BS, MS, Eric A. Shry, MD

1Neurology Service and 2Clinical Informatics Division, Madigan US Army Medical Center, Tacoma, WA

Introduction

Advanced analytics including natural language processing (NLP) have been promoted as an effective strategy to reduce hospital readmissions. These tools may not fully exploit the rich categorical data found in patients’ past medical histories (PMH) to forecast risk, due to the format of the data. Readmission risk prediction tools powered by novel NLP techniques that produce exploitable representations of PMH data such as Doc2Vec may be a suitable alternative, but there may be insufficient evidence to support this.

Objective

The purpose of this study is to validate Doc2Vec for analysis of International Classification of Diseases (ICD) codes by showing prediction of all-cause readmission risk and concurrence with physicians’ interpretation of diagnoses.

Methods

An open-source implementation of Doc2Vec was used to analyze 3.5 years of de-identified electronic medical records data from a single tertiary care hospital to establish relationships between 30-60-90 day readmissions and ICD codes. The resultant vectors were used to predict readmission risk using cross-validated logistic regression. An “odd one out of four” dissimilarity test, analogous to the semantic and syntactic reasoning tests used to validate NLP algorithms, compared the model’s ranking of similarity to answers from 33 licensed, blinded physicians.

Results

Readmission prediction accuracy at 30, 60 and 90 days was 91.3% (+/- 0.81%), 95.2% (+/- 0.47%), and 96.6% (+/- 0.33%), respectively. In the dissimilarity test, physicians had significant concordance of 38.3% with the model on the generated questions compared to a random chance of 25% over 660 example (chi-squared p<0.0001).

Conclusions

The results from this small database study may over-estimate the effectiveness of the model, given the limited, homogenous sample used. Further evaluation using larger datasets and data from other facilities may provide a more accurate assessment of the model’s efficacy. To the best of our knowledge, this is the first application of Doc2Vec for analysis of medical diagnoses, as opposed to its standard application of single word analyses outside of healthcare. Results from this study may help promote a greater understanding of the use of diagnoses and vector representations in clinical practice.
The Adolescent Patient Portal: Are we meeting the needs of our most connected patients?

Chase R. Parsons, DO1, Jonathan D. Hron, MD1, Fabienne C. Bourgois, MD MPH1
1Boston Children’s Hospital, Boston, Massachusetts, United States of America

Introduction

With the rise of Electronic Health Record (EHR) adoption in the United States, healthcare organizations have implemented patient portal functionality to fulfill government regulations and promote the “meaningful use” of EHRs1. In pediatrics, research has demonstrated increased patient engagement in both the inpatient and outpatient setting with use of a patient portal2, 3. Despite strong recommendations from The Society for Adolescent Health and Medicine (SAHM) to implement patient portals with confidentiality standards, including differential access for parents and adolescents4, it is suspected that pediatric institutions have heterogeneous policies surrounding adolescent access to the patient portal, with most institutions significantly restricting access. This study aims to evaluate the current state of pediatric patient portals, focusing on differential access policies for the adolescent population.

Methods

Our team, consisting of Clinical Informaticists at an academic children’s hospital, developed a 50-item web-based survey to delineate policies and features of pediatric and adolescent patient portals. The survey covers general user-access policies, patient data and documentation viewing, secure messaging, and medical results access for both the adolescents and their parents/guardians. It also explores whether the adopted policies were guided by institutional decisions, EHR vendor limitations, and/or specific state regulations. We identified our participants using the Children’s Hospital Association (CHA) member directory and electronically delivered our survey to the chief medical informatics officer (CMIO) (or nearest equivalent staff member) of the 226 CHA U.S. children’s hospitals. We compared characteristics of responding hospitals (size, region, freestanding children’s hospital status, number of portal accounts, and EHR implementation year) with survey results. Early data was analyzed using descriptive statistics.

Results

This survey has now been deployed and early results are available. One week after survey deployment, 10 of 226 CHA hospitals responded (4.4%). Responding hospitals were heterogeneous and characteristics were consistent with the overall composition of general acute care hospitals in the U.S. The number of total institutional portal accounts amongst respondents ranged from 3,167 to 1.1 million. Sixty percent of responding institutions enable separate adolescent and parent accounts at the onset of adolescence (12 – 14 years old) and 50% of those require patient assent for full parental access, while 30% of respondents disable both adolescent and parental portal access at this time. The remaining 10% report the capability for only one account which is either used by the parent, patient, or shared. Narrative comments from responding institutions varied, but multiple centers report that the adolescent patient portal is both a “struggle” and a “challenge.” Respondents report that institutional portal policy is limited by both state law and EHR vendor offerings. Larger freestanding children’s hospitals were associated with more granular user access policies aligning with SAHM’s recommendations.

Conclusion

This study, to our knowledge, provides the first evaluation of common practices in adolescent patient portals. Preliminary data demonstrates significant variability in pediatric hospitals’ policies regarding adolescent patient privacy and access in the patient portal. Further data will be collected and analyzed. These results will help inform efforts toward standardizing policies and procedures regarding adolescent patient portal access across pediatric centers and help identify gaps in current vendor-provided portal solutions.

References

America Walks - A Cross-platform Mobile ResearchKit Study of Walking Behavior Based on Passive Step Tracking

Chintan Patel PhD, Nickolas Whiting, Aamir Bhat, Mir Suhail, Priya Menon, Shweta Mishra, Sharib Khan MD MA MPH
Applied Informatics Inc., New York, NY

Background
Nearly two-thirds of Americans own a smartphone and routinely use them to actively seek or manage health and wellness information. The release of ResearchKit, an open source framework and software development kit (SDK), has enabled the creation of study apps with e-consent, surveys, phone sensor recorded tasks and wearables data. However, apps created using ResearchKit are only available to Apple device users, which excludes a large population of potential participants that use other platform smartphones (primarily Android), thereby introducing the likelihood of sampling bias. This paper describes our effort to address this limitation by conducting the America Walks study (https://clinicaltrials.gov/ct2/show/NCT02792439) to track walking behavior using phone step count sensors, using an app for both iOS and Android users.

Methods
To develop a study app that could run on both Apple iOS and Android phones, we first developed ResearchDroid, an Android port of ResearchKit. ResearchDroid deeply integrates with the Android platform and is designed from the ground up to meet the requirements of capturing participant consent, surveys, extracting sensor data and meeting the security considerations for IRB approval. Though ResearchDroid contains similar capabilities to ResearchKit, operating system differences between Android and iOS, such as explicit back button use in Android phones, necessitated some differences in UI implementation. Moreover, whereas iOS provides a standard interface to access sensor and health data (HealthKit), Android phones come with a variety of sensors from different manufacturers, so several connectors were written to obtain sensor data from Android phones.

We next developed the America Walks Study app using ResearchKit for iOS and ResearchDroid for Android users. The entire study was app-enabled; participants pre-screened themselves using a two-question survey and completed the entire informed consent within the app. Upon enrollment, they answered a demographic survey and provided permission for the app to access their step counts data from Apple’s HealthKit or Android step count sensors. Step counts were tracked automatically for 30 days from enrollment. A real-time study dashboard provided participants the ability to view their step counts and compare that to counts of other participants combined. Participants could opt out anytime or extend participation after the 30 day period. The primary objective was to determine walking activity as measured by daily step counts and to compare this across demographic categories.

Results
The study enrolled 210 participants, (114 females, 96 males) of which, 41 used Android phones and 169 used iOS devices. Recruitment was done via social media marketing and listing on ClinicalTrials.gov. Participants spanned across 35 states and 130 cities. A total of 22,219,847 steps were tracked over 300 days (out of the 305 days the study was live) of which 19,619,679 (88.2%) steps were recorded by iOS and 2,600,168 (11.7%) steps were recorded by Android devices. The average daily step count of an iPhone user was 9,868 steps, compared to 3,042 daily steps for an Android user. Maximum step count activity was recorded at noon and 6pm. 63% of participants were active over the 30 day study period, indicating sustained engagement.

Discussion
We successfully demonstrated the possibility of creating a study app that can run on both iOS and Android phones. Though significant programming work was needed to build ResearchDroid, subsequent app development was much faster. Android development using ResearchDroid required equivalent effort as compared to iOS development with ResearchKit. The existence of numerous manufacturers of Android phones causes variability in screen sizes and sensor implementations, which do add debugging and screen rendering complexities. We are currently working on three studies that will be launched as cross-platform apps.

References
For the Common Good: Sharing Data Extracted from Text

Olga V Patterson, PhD\textsuperscript{1,2}, Scott L DuVall, PhD\textsuperscript{1,2}

\textsuperscript{1}VA Salt Lake City Health Care System; \textsuperscript{2}University of Utah, Salt Lake City, UT

Abstract

The Department of Veterans Affairs (VA) Informatics and Computing Infrastructure (VINCI) resource center mission includes the goal of improving the quality of research with services and innovations supporting acquisition and analysis of healthcare data. VINCI is engaged in a number of projects to meet this mission. This poster reports on a new approach to supporting the research community it serves by applying natural language processing (NLP) on all VA notes and reports and making the output available as a new data source.

Methods

VINCI sponsored development of a high-performance scalable rule-based extraction system for identifying quantitative values for left ventricular ejection fraction (LVEF) in clinical notes. LVEF is an important measure that reflects heart health and any study that evaluates heart disease requires this variable, which is why it is commonly requested.

The system is based on the Leo framework\textsuperscript{1} that extends the Apache Unstructured Information Management Architecture Asynchronous Scaleout (UIMA AS)\textsuperscript{2}. The extraction system is light-weight because it utilizes a highly targeted custom lexicon and a specific rule set. For system validation, we evaluated documents from three sources within the medical record – echocardiograms, radiology reports, and clinical notes – which combined have more than 3 billion text documents. Therefore, we created a comprehensive but targeted set by filtering all documents based on the presence of at least one of the following highly sensitive and specific keywords: “ef”, “lvef”, “eje*”, “function”, “fraction”, “dysfunction”. We randomly selected 200 documents from each source. The documents were manually annotated with the values for LVEF and compared against the values found by the NLP tool. Precision of the extracted LVEF values ranged 0.981-0.993 across the three sets with average of 0.986. Recall ranged 0.887-0.935 with average of 0.925. F-score ranged 0.932-0.960 with average of 0.955. A failure analysis revealed that the main source of error is mistaken LVEF values for the normal range that is often included in documents as a reference.

The NLP tool was applied to all records with the first run completed in September 2015 and three additional incremental runs were completed in May 2016, November 2016, and June 2017. After the raw NLP dataset was created, additional data cleaning steps were performed to normalize values and discard values above 100\%. The combined set contains 35 million LVEF values for 3.4 million patients, supplementing the original structured Echo dataset, which contains only 429 thousand values for 277 thousand patients. The current plan is to continue with periodical incremental updates to this dataset.

Conclusion

Since the dataset has been created, over a dozen of studies have already benefitted from having this dataset – without having to rerun the NLP tool each time. The novelty of the approach required development of a new procedure on how to ensure regulatory compliance when a data request is processed. Initially, each data request was reviewed individually, but with the acceptance of a new regulatory procedure, the requests will be processed routinely and access to the new dataset is available to any VA-affiliated researcher.

Providing structured datasets that result from NLP processing enriches available data, which saves time for all other researchers and makes it possible to ask new research questions. VINCI has plans to continue with this approach and provide additional values extracted from text: vital signs, smoking status, ankle brachial index, and others.

Acknowledgements

This work was supported using resources and facilities at the VA Salt Lake City Health Care System with funding from VA Informatics and Computing Infrastructure (VINCI), VA HSR HIR 13-457.

References

Distance learning as a strategy for the education of health informatics professionals: possibilities and challenges
Andréa P. S. Pelogi¹, Alissa P. Penteado¹, Ivan T. Pisa¹, Claudia G. Novoa¹
¹ Department of Health Informatics -Unifesp, São Paulo, Brazil

Introduction
Health informatics is a discipline that involves several areas of knowledge and has been considered of extreme importance(¹). There is a paradigm shift in higher education such as the result of technological advances, the introduction of e-Learning and a greater participation of higher education professors in the area of distance education (DE). It is necessary to explore which is the best way to use scarce resources in order to cause a great impact on the training of such professionals so that they can make the best use of the available technologies to support the provision of health services (²-³). New technologies can help the affective process between professors and learners, who will be mediated by the interaction that technological resources provide, generating relationships which involve respect, trust and cooperation(⁴).

Methods
In total, 682 students were separated into 15 groups. The pedagogical project of the course was designed so that its content was divided into disciplines. Each discipline was developed in 4 axes: 1 - Concept: didactic material; 2 - Collaboration: collaborative activity; 3 - Reflection goal: description of learning; and 4 - Assessment: collaborative tasks. In all moments throughout the course, content-specific tutors clarified doubts, followed-up the attendance and improvement of participants. Professors developed the pedagogical content with the help of a team of instructional designers using the teaching resources available in the LMS that best served the purpose. The course emphasized collaborative tasks from the perspective that the graduate student learns with both professor and peers.

Results
The result obtained showed that 374 (54.83%) of the students passed, 53 (7.77%) failed, and the registered withdrawal was 255 (37.39%). During the course the students participated in self-assessments about their work in collaborative format. The questionnaire consisted of 10 questions. Students' responses demonstrated if they were prepared for the collaborative work: 0.46% never prepared; 2.75% rarely; 24.71%, sometimes; 41.88% many times and 30.21% always.

Discussion
Distance Education uses communication and information technologies to promote both professional and human development. Therefore, this teaching approach makes that student learns without moving him away from his daily activities. Its characteristics enable the elaboration of an online course based on the proposal of intercommunicative teaching, promoting autonomy and the exchange of experiences (²). However, even with the support of machines and virtual learning environments (LMS), the DE requires motivated professionals, reliable technical structure and suitable monitoring. It is necessary to think over on the conceptions of collaborative learning in this teaching modality, and also evaluate the strategies used and the role of the teacher when collaboration occurs in LMSs(³). These are some obstacles to the full use of DE resources: difficulty to access information and communication technologies, difficulty to use tools, lack of time to develop activities, difficulty to communicate with tutors and to manage personal life. Nevertheless, it is well known that DE with collaborative bias represents an interesting possibility for lifelong education considering the challenges faced by professionals in the area of health informatics in the scope of access to continuing education.

Conclusion
The course provided students both individual and collective learning experiences in order to prepare them to work in hospitals and health services where there are information systems and other resources based on the use of information and communication technologies. The obtained results showed that the variety of presentation structures of the conceptual domain favored motivation. This could be verified through the participation in several collaborative learning activities and by the learners' self-assessment questionnaires. Distance Education based on a collaborative method seems to be an efficient, practical pedagogical strategy faced with the need to qualify the health professionals.

References
RxNav 2.0 – A web-based, mobile-responsive RxNorm browser

Lee Peters, M.S. and Olivier Bodenreider, M.D., PhD
U.S. National Library of Medicine, National Institutes of Health, Bethesda, Maryland, USA
Contact information: RXNAVINFO@LIST.NIH.GOV

Motivation
RxNav is a browser for several drug information sources, including RxNorm, RxTerms and NDF-RT. Since its release in 2004, many features have been added over the years, including pill images and drug-drug interaction information. However, the Java Web Start framework originally used was not adaptable to an environment where web resources are increasingly used on mobile devices. In 2016 RxNav was redesigned to conform to modern web application development principles, with the goal of simplifying the user experience and supporting mobile devices.

Features and look of RxNav 2.0
Startup. The redesigned RxNav 2.0 is a web application which can be run by selecting the RxNav button from the RxNav website (https://rxnav.nlm.nih.gov). Based on Java Script, it only requires a browser such as Chrome or Firefox, whereas the original version of RxNav required users to have java installed on their system in order to run the application. RxNav 2.0 has a faster startup time than the original application, as well as faster search times.

Responsive Design. Based responsive design techniques, RxNav 2.0 can be displayed appropriately on workstations, tablets and mobile devices. The original application was limited to devices running java and required a large display area. Figure 1 (left) shows a screenshot of RxNav 2.0 on a mobile phone.

Functionality. Our goal in developing RxNav 2.0 was to keep the functionality of the original application and provide additional features. These new features include:

- **Ubiquitous searching.** Searching can be done while visiting any tab in RxNav 2.0.
- **NDC properties.** The NDC tab displays properties from DailyMed for each NDC.
- **NDC History.** The NDC tab contains a history feature that displays all past and present NDCs for the RxNorm concept.
- **Links to RxClass.** The nodes in the graphs in the Class View tab are linked to RxClass and clicking on a class in RxNav 2.0 opens RxClass for that class.
- **Search history.** A search history allows the user to conveniently revisit previous searches.

Appearance. RxNav 2.0 keeps the basic tab structure of the original application, but the tab views have been modified in some cases to provide better navigation and enhanced look and feel. In particular, RxNav 2.0 offers a simplified graph view limited to ingredients, brand names and drug products (Fig. 1, right).

Acknowledgments: This work was supported by the Intramural Research Program of the NIH, National Library of Medicine.

Figure 1. RxNav 2.0 on a mobile phone (left); new simple graph view (right)
Assessing Inter-Hospital Clinical Processes Variability with Process Mining

Marius Petruc, MD MS¹, Mihail Popescu, PhD²
¹MU Informatics Institute, ²Department of Health Management and Informatics, University of Missouri, Columbia, MO

Introduction

The last decade has brought an increasing awareness that current, unsustainable healthcare cost levels need to be lowered, that the patient experience of care could be much improved and that the health of the population should be advanced. Outcomes and costs depend, among other factors, on the way care and care-related activities are organized and performed so that best practices are followed and variability reduced. As part of a process improvement project at our multi-hospital system, we have started inventorying clinical processes and measuring their variability across hospitals. Time-motion studies have proved successful previously, but they are costly and do not scale easily. Data science methods, especially process mining, are promising to address these limitations. In this poster we describe our experience assessing the feasibility of using process mining techniques to gather process variability and performance data related to the interpretation of ER CT examinations in two hospitals at an academic medical center.

Methods

Process mining leverages data already generated by information systems for other purposes (HIPAA compliance, debugging, etc). For our study, we collected log data generated from January 1st, 2015 to December 31st, 2015 (one year of data) by the radiology information system (RadNet, Cerner Corporation). The initial dataset consisted of 268,361 records representing event traces for all imaging examinations performed across our multi-hospital system. This dataset was imported in RStudio Version 0.99.486 for pre-processing, which consisted of removing duplicate records, transforming the date/time information to a consistent format, and filtering out all traces for events that were not related to Emergency Room CT examinations. In the end we extracted 17,408 ER CT cases, which were further split in two groups based on the location where they were performed: the adult hospital (15,660 cases) and the children’s hospital (1,748 cases). Data analysis was performed using the Disco (Fluxicon Process Laboratories) process mining application using the fuzzy miner algorithm.

Results and conclusions:

Figure 1 shows the two causative workflow nets (Adult-ER-CT and Peds-ER-CT) mined by the fuzzy miner algorithm in the Disco application. The blue boxes represent steps in the process. They are connected to each other by arrows to form paths, the most frequent of which are highlighted by darker arrows. The numbers shown in the boxes and next to the arrows are the actual counts of events. The arrows also show the average number of minutes between each step in the process. Examining the two diagrams shows that there is little variability between the two hospitals in terms of process flow, which is to be expected. However, there are differences in the time it takes to perform specific steps: the peds ER-CT studies seem to be dictated and finalized faster than the adult ones, while the exams themselves appear to be performed faster at the adult hospital.

Figure 1. Adult (left) & pediatric (right) CT mined workflow
Developing a Framework for Evaluating Influencers of Frailty in Aging

Andrew B. Phillips, PhD, RN1,2, Margarita Sordo, MSc, PhD3,4,5, Ruth Palan Lopez PhD, GNP-BC1,3, Jacqueline A. Merrill, PhD, MPH, RN6
1MGH Institute of Health Professions, School of Nursing, Boston, MA; 2Massachusetts General Hospital, Boston, MA; 3Clinical Informatics, Partners eCare, Partners Healthcare, Boston, MA; 4Department of Medicine, Brigham and Women’s Hospital, 5Harvard Medical School, Boston, MA; 6Columbia University, Nursing and Biomedical Informatics, New York, NY

Introduction
Frailty, a decrease in physiological reserve, is a serious threat for older adults that significantly increases their risk of chronic illness, institutionalization, poor quality of life, and death. Early intervention during the pre-frail state may reduce poor outcomes associated with frailty. As there is little consensus on how to measure frailty, there is wide variation in both estimates of frailty prevalence and poor understanding of interactions among known factors that contribute to frailty. There is a need for a dynamic adaptive agent model (AAM) of frailty to gain insight into interactions between individual preconditions and other factors to guide and support decisions on pre-frailty interventions.

Methods
AAMs are a “bottom-up” probabilistic computational strategy for simulating complex systems. In such models, unique agents are assigned micro-level attributes, (i.e. at the individual and environmental level) and a set of computational (if/then) rules to govern agent actions within a system. By running multiple simulations of autonomous agent actions and interactions macro-level outcomes of the system emerge and can be evaluated. To develop parameters for a prototype AAM to test against known frailty prevalence outcomes, we utilized a single instrument, the Survey of Health, Ageing and Retirement in Europe—Frailty Index (SHARE-FI), that incorporates five well accepted criteria for assessing frailty. We used expert opinion to cross map the measures to selected variables in the National Survey of Midlife Development in the US (MIDUS). SHARE-FI measures constitute initial agent attributes and longitudinal distributions from MIDUS are the basis for action rules in our initial model. The results below present example criteria, SHARE-FI measure and operationalization based on MIDUS variable.

Results

<table>
<thead>
<tr>
<th>Frailty Criteria</th>
<th>SHARE-FI Measure, Operationalization of Rule</th>
<th>MIDAS VARIABLE: (ID), Description, Range of Valid Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Shrinking/Weight loss (unintentional)</td>
<td>Self-report (diminution, no change, increase) If weight_a_year_ago - current_weight &gt;= 10 lbs AND &quot;Lose 10 lbs for other reason = 2&quot;</td>
<td>(A1SA30C), Lose10 lbs for other reason, (1 = &quot;Yes&quot;, 2 = &quot;No&quot;) (A1SA28), Weight one year ago, (actual weight) (A1SA27), Current Weight, (actual weight)</td>
</tr>
<tr>
<td>2. Weakness</td>
<td>Grip Strength If Limit_lifting or carrying_groceries &gt; 1</td>
<td>(A1SA17A), Limit lifting or carrying groceries, (1-4) 1 = &quot;A Lot&quot;, 2 = &quot;Some&quot;, 3 = &quot;A little&quot;, 4 = &quot;Not at all&quot;</td>
</tr>
<tr>
<td>3. Poor Endurance/Exhaustion</td>
<td>Self-report (Y/N) If Limit_climbing_stair &gt; 1</td>
<td>(A1SA17C), Limit climbing stairs, (1-4) 1 = &quot;A Lot&quot;, 2 = &quot;Some&quot;, 3 = &quot;A little&quot;, 4 = &quot;Not at all&quot;</td>
</tr>
<tr>
<td>4. Slowness</td>
<td>Self-report (Y/N) If Limit_walking_several_blocks &gt; 1</td>
<td>(A1SA17F), Limit walking several blocks, (1-4) 1 = &quot;A Lot&quot;, 2 = &quot;Some&quot;, 3 = &quot;A little&quot;, 4 = &quot;Not at all&quot;</td>
</tr>
<tr>
<td>5. Low Activity</td>
<td>Self-report: No. activities per week If Physical_fitness_vs_5_years_ago = 3</td>
<td>(A1SA6B), Physical fitness compared to 5 years ago, (1-3) 1 = &quot;Better&quot;, 2 = &quot;No Change&quot;, 3 = &quot;Worse&quot;</td>
</tr>
</tbody>
</table>

Discussion
The results will provide initial data on the relative influence of each of the five criteria on frailty prevalence. The ability to successfully demonstrate predictive capabilities using a small prototype AAM, provides a foundation for increasingly complex models of frailty risk. The inclusion of social, economic, and other environmental factors within the MIDUS dataset will allow such an expansion. Future work will be guided by domain experts, using existing data, and scientific evidence to iteratively incorporate a greater number of data points and interactions.

References
A Simple and Efficient Method for Private Exact Matching

Le Trieu Phong, PhD¹, Xiaqian Jiang, PhD², Shuang Wang, PhD³
¹National Institute of Information and Communications Technology, Tokyo, Japan
²University of California San Diego, La Jolla, CA, USA;

Introduction
Genomic data are playing an ever-important role in medical research and clinical studies but the size of the raw data is very big (e.g., ~500GB for the whole genomes of an individual), which is cost-ineffective to host and analyse them locally in hospitals. A straightforward way is to utilize the cloud computing to store and analyze them on demand. But a major concern is that genomic data are highly sensitive and depositing raw data in a public cloud processes large privacy risks of data owners. In 2016, iDASH privacy workshop called for novel solutions to address this challenge and several teams demonstrated innovative approaches¹. But current solutions still have big overheads, limiting their practicability in applications with time constraints. We propose a simple and efficient solution with a slightly relaxed security assumption in this work.

Methods
As shown in Figure 1, given a genome database, Data Owner first encrypts this database using PRF before sending it to the Cloud Service. At a later time, Data Owner generates a plain query, which she encrypts and sends to Cloud Service. Cloud Service responds with an encryption that Data Owner can use to obtain the boolean result of match or non-match. It is required that Cloud Service should not learn anything about the query and the data. Our protocol is as follows. For a Database = {D₁, ..., Dₙ} of N records, the upload phase at Data Owner does the following steps:
1. Takes a PRF FK and generates a corresponding secret key K.
2. Computes and sends FK(D₁), ..., FK(Dₙ) to Cloud Service.

For the matching phase, Data Owner gets a query Q, computes and sends FK(Q) to Cloud Service who does the following steps:
3. Looks up FK(Q) in the set of FK(D₁), ..., FK(Dₙ).
4. If there is an index i such that FK(Q) = FK(Dᵢ) then return match; else return non-match.

Security can be argued as follows: the function FK is pseudo-random function with a secret key K known only to the data owner, thus, Cloud Service obtains no information from FK(D₁), ..., FK(Dₙ) in the upload phase and FK(Q) in the query phase. Therefore, information on the database and the query is kept secret from Cloud Service. In step (4), secure channel such as TLS/SSL can be used to securely transmit the result of match/non-match. In addition, revealing match/non-match to Cloud Service may be harmless or harmful depending on the structure of the database, the latter case can be addressed via adding dummy records in the database and asking dummy queries.

Results
Our experiment was conducted with a dataset from the iDASH 2016 genome privacy protection competition¹, which contains 100K records at the size of 5.3 MB. We used HMAC-SHA256 as the PRF². Table 1 shows that the proposed method is more efficient than the solution using homomorphic encryption in iDASH 2016 competition.

<table>
<thead>
<tr>
<th>Solutions</th>
<th>Data encryption</th>
<th>Encrypted data size</th>
<th>Secure matching</th>
<th>Query size</th>
</tr>
</thead>
<tbody>
<tr>
<td>iDASH competition¹</td>
<td>1.86(s)</td>
<td>24(MB)</td>
<td>3.09(s)</td>
<td>0.64(MB)</td>
</tr>
<tr>
<td>Proposed method</td>
<td>1.94(s)</td>
<td>4.3(MB)</td>
<td>0.0016(s)</td>
<td>65(bytes)</td>
</tr>
</tbody>
</table>

Limitation
To allow fair performance comparison, the proposed algorithm was only evaluated on a small scale dataset with 100K records in VCF files provided by the iDASH 2016 genome privacy protection competition. In our future work, we will evaluate the performance over a large dataset. The proposed method mainly focuses on the secure outsourcing protocol for a single site, which means the proposed algorithm does not support data and queries that need to be encrypted from different sites. The proposed method only support exact comparison in a secure manner, which justified the further investigation to support secure fuzzy matching efficiently.

References
A Usability Study to Evaluate the Impact of a Novel Automated Brain Tumor Assessment Application

Edgar A. Rios Piedra MSc, Iren Orosz MD, Mary Zide PhD, Suzie El-Saden MD, Ricky K. Taira PhD, Alex A.T. Bui PhD, William Hsu PhD

Medical Imaging Informatics Group, Radiological Sciences, David Geffen School of Medicine, University of California Los Angeles, USA.
Bioengineering, University of California Los Angeles, USA.

Introduction. Efficient and accurate evaluation of brain tumor imaging studies is an important facet of managing neuro-oncology patients but is a particularly challenging task due to heterogeneity in tumor appearance and the time required to perform a comprehensive evaluation. We have developed a novel image analysis application that automatically delineates tumor components and visualizes how tumor volume changes over time. In this project, we conducted a usability study among neuroradiologists to evaluate whether our computer aided system for automated image analysis can alleviate the challenges of image interpretation in this patient population.

Methods. On this exploratory study, eight experienced neuroradiologists evaluated the automated tumor segmentation application [1] on a set of six different patients with multiple follow ups (Figure 1), and afterwards answered a usability survey developed by the authors and consisting of ten questions. The questionnaire includes an evaluation of the user interface (intuitiveness of single and multiple data points), display (data display and arrangement), understanding of data (understanding probability maps and graphs, results have informative value), and program usability (performance, work facilitation). Results were gauged using a Likert-type scale ranging from 0 (no agreement) to 5 (complete agreement) (Figure 2) and found high agreement (4.62 ± 0.51) regarding the user interface being helpful to better evaluate data while a smaller proportion indicated it would improve their diagnostic workflow (4.12 ± .99). Additionally, visual aids that can display error-range in tumor measurements (e.g., timeline) were identified by the users (six out of eight) to be significantly helpful during the decision-making process.

Conclusion. With this exploratory usability study, we demonstrated that computer-aided approaches have a positive effect on the data understanding and effective work of neuroradiologists. Future work includes improving the application based on the feedback received on the surveys and expanding the evaluation sample size.

References
Presentation of Clinical Pathways for Improved Provider-Patient Shared Decision Making

Jose Pineda, PhD(s)1, Lina Bouayad, PhD1,2, Min Chen, PhD1
‘Florida International University, Miami, FL, ‘James A. Haley Veterans’ Hospital, Tampa, FL

Introduction

The study of clinical pathways to guide the diagnosis and treatment of patients has been shown to improve their prognosis. However, prior researchers have limited their studies to analyzing the usefulness of these pathways for the physicians performing the diagnoses, while paying little attention to how the patients could learn from this data themselves. Yet, these pathways offer major potential for informing patients about their conditions and treatments.

In the past couple of decades, a major move towards patients and physicians making decisions together has taken place1. Despite intentions to expand on this practice, researchers have identified barriers for patients to participate with the most significant being inadequate provisions of information2. By collecting the knowledge found in clinical pathways, matching it to the patients based on their characteristics and prior diagnosis as found on their personal health records (PHR), and presenting the results to patients in the form of decision aids, we argue that it will be possible to improve the patients’ understanding about their conditions as diagnosed by their physicians, so that the physician and the patient can have more meaningful discussion about the possible treatments options available.

To our knowledge, this would be the first study to use analytics to analyze clinical pathways and match them to the patient’s data in order to improve the patient’s knowledge. If successful, this approach could be used to improve interactions between physicians and their patients leading to more effective shared decision making.

Methods

In this study, we aim at extracting and presenting sets of possible treatment options to the patients. The data used in this study is extracted from 135,962 electronic discharge records of 285 hospitals in the state of Florida. For every patient admitted, 192 different variables are used to describe different diagnoses, procedures performed, and relative dates of admission. A unique patient identifier allows us to track patient admissions across hospitals and over time. We start by focusing on Congestive Heart Failure (CHF) patients, who have some of the highest readmission rates (23%-26%) of any type of patient.

First, patients’ matching is made by using patient characteristics and set of diagnoses. Next, we create clinical pathways using process mining techniques on 3 data elements: 1) a case identifier, 2) an activity identifier, and 3) a calculated time stamp based on the relative dates of admission and diagnoses. In order to make our dataset more human readable, we then convert all extracted clinical pathways from ICD9 and CPT code to matching description using data dictionaries published by the Agency for Healthcare Research and Quality (AHRQ). Sets of treatment options are then evaluated based on different criteria, namely 1) outcome – combination of efficacy and side effects represented by the status at discharge, and 2) procedure cost.

Results

Results including the sets of diagnoses and matching treatment options to be presented to patients will be presented at the conference.

Conclusion

In the present study we extract and present a set of treatment options to patients. Results from this study can help empower patients and drive their engagement.

References

Andrew Placona, MS, Fengjuan Wang, MS, MBA, Shantanu Phatakawala, MS, Nick Toure, BS, Evolent Health, Arlington, VA

Introduction
Medicaid finances 45% of U.S. births totaling $27.6 billion dollars annually\(^1\). While attempts to identify high risk maternity patients, in advance to improve outcomes and lower costs, the models are not used in daily practice because there are questions about availability of timely data, model validity, and accuracy\(^2\). Medicaid enrollment changes and churn exacerbates these issues because many predictive models rely on robust data sets, which may not be readily available. In addition, members may have survey data, pharmacy data, lab, and claims data all pointing to the same condition or severity within a condition. The objective was to develop an algorithm capable of integrating disparate data to assess the risk of pre-term, respiratory distress, or neonatal abstinence by the 20th week of pregnancy.

Methods
We identified 3539 births, between 2015 and 2016 in a large Medicaid plan in KY, that had at least 9 months of continuous eligibility for the mother, and could be linked to a live birth through ID or address match algorithm. We coded any birth pre-term, respiratory distress, or neonatal abstinence as 1 and the rest as 0 to create a binary outcome. Next we pulled medical, pharmacy, labs, and health assessment data, when available, for the members for the past two years prior to 20th week of gestation by normalizing for gestational age at birth. Because of data sparsity and reoccurring signals, we developed data munging process to combine similar information from different data sources. Our example for Periodontal Disease highlights our feature engineering process (See Figure 1):

Features included Diabetes, smoking, Substance Abuse (SA), Periodontal Disease (PD) in addition to demographic features such as Age. We employed a LASSO Logistic Regression classification framework using R’s Glmnet package with 5-fold cross validation.

Table 1: Selected Sample Dataset Descriptive Statistics

<table>
<thead>
<tr>
<th>Avg Age</th>
<th>% Diabetes</th>
<th>% smoking</th>
<th>% SA</th>
<th>% PD</th>
</tr>
</thead>
<tbody>
<tr>
<td>26.01</td>
<td>14.91%</td>
<td>37.37%</td>
<td>11.24%</td>
<td>6.47%</td>
</tr>
</tbody>
</table>

Results
Our model had a validation c-statistic of .7. Significant features included: age, history of preterm, Diabetes, Substance Abuse, number of opioids.

Discussion
The model has a 10 percent higher c-statistic performance than other models built using comparable data sources\(^3\). It is possible model is does not generalize outside the dataset. Exploration is needed to test the validity of the model. We need to assess whether these predictions are useful to practitioners in the field. Three measures of interest include whether the predictions are valid, timely, and can effect change in patients.

References
A comprehensive Variant Analysis System for Molecular Pathology

A. Pozhikov PhD, S. Bolisetty MTech, MS, F. Martinez BS MBA, K. Honnavali MCA, T. Kirkpatrick MS, M. Guirgus, R. Rout, A. Shah, PhD MBA, R. Pillai MD
Beckman Research Institute and City of Hope National Medical Center, Duarte, CA

Introduction

Molecular analysis of tumor biopsies (molecular pathology) is critical for determining the course of patient’s treatment. Genomic sequencing of the biopsy specimen reveals the SNPs, fusions, CNVs and other genetic rearrangements occurring in oncogenes and oncosuppressors. Depending on the gene involved, either a targeted therapy or chemotherapy/radiation/surgery may be prescribed. Variant information handling, interpretation, and reporting of the results are the biggest bottlenecks in the tail-end workflow of the NGS-based testing. Several software applications, open source, commercially licensed, and custom designed such as ANNOVAR, AnnTools, VARIANT, SnPEff, SeattleSeq, Annotate-It, Appistry, IonReporter, NextGENE etc are available for variant annotation and visualization. The common features shared by these applications are SNP and insertion and deletion identification, gene and variant annotation, in silico prediction score annotations, and variant filtering features. For a comprehensive overview of the existing tools see “Standards and Guidelines for the Interpretation and Reporting of Sequence Variants in Cancer” by Li, Marilyn M. et al. The Journal of Molecular Diagnostics, 19:4-23.

Our group (Research Informatics and Systems) developed a comprehensive information system for molecular pathology (HOPESEQ). Unlike the above-mentioned existing packages, the HOPESEQ offers the following features: (i) archiving variants determined by sequencing; (ii) maintaining a knowledge base (KB) of previously classified variants as pathogenic or benign; (iii) matching patients’ variants to public databases, i.e., COSMIC and ExAC; (iv) providing natural language based searching to interpret published literature and clinical trials and (v) visualizing the variants of the patients archived within HOPESEQ.

Methodology

Raw sequencing files are processed with an experimentally-verified pipeline, which maps the sequencing reads to the human genome (e.g., HG19) and detects variants. The detected variants are filtered to satisfy minimum sequencing depth criteria followed by calculation of allele frequency. The allele frequency is important for estimating the pathogenic potential of the variant, i.e., there is a different impact of zygocity on oncogenes vs. oncosuppressors. The filtered variants are submitted to HOPESEQ (see fig. above) by a technician.

A pathologist and a geneticist (or more than one geneticist) review the submitted variants using the above 5 key features of the HOPESEQ. The purpose of the review is to make a determination of whether a variant is pathogenic or not. Specifically, if the variant was previously saved into the KB, the determination of the variant’s pathogenicity is based on that record in the KB. If the variant has never been seen before, the pathologist and the geneticists can use HOPESEQ and consult against the external databases (COSMIC, ExAC).

HOPESEQ automatically matches the variant in question to the records in COSMIC and ExAC. Natural language processing (NLP) aids the pathologist to interpret the literature referred by the external databases. If no records are available in the external databases, HOPESEQ visually shows which part of the gene product (e.g., enzyme, transcription factor) is affected enabling the pathologist to at least provide a mechanistic estimate of the variant’s pathogenic potential. Additionally, HOPESEQ facilitates a discovery of appropriate clinical trials based on the inclusion criteria addressing the variants of the patient. Finally, a report is generated that includes selected pathogenic variants, appropriate drugs and potentially suitable clinical trials.

Current status of HOPESEQ

The HOPESEQ platform is currently being used by the Pathology Department at the City of Hope. To our knowledge, no other existing software offers a cohesive interplay between the variant pathogenicity prediction and expert reviewing, whereby the prediction is aided by NLP and protein structure consideration.
Mobile system to gather patient-generated data in Parkinson’s Disease treatment.

Marco Prenassi, MSc1, Pierluigi D’Antrassi, MSc2, Sergio Barbieri, MD2, Lorenzo Rossi, PhD3, Simona Mrakic-Sposta2, Sara Marceglia, PhD1,2
1University of Trieste, Trieste, Italy; 2Fondazione IRCCS Ca’ Granda Ospedale Maggiore Policlinico, Milano, Italy; 3Newronika S.r.l., Milano, Italy;

Introduction

Deep brain stimulation (DBS) implants require a stabilization period to benefit patients affected by Parkinson’s disease (PD). A correct and constant monitoring of this period is essential to regulate their therapies. In a homecare environment is necessary an easy-to-use, comfortable, affordable, and reliable system. The scope of this work is to test if a commercially available and affordable accelerometric sensor bracelet paired with a mobile device is capable of assessing motor symptoms without causing discomfort to the patient.

Methods

A dedicated Android App was developed to provide a clinical diary to be filled in by the patients at predefined times. It is paired with a Pebble Time smartwatch that acquires tri-axial accelerometric data from the wrist. An offline Matlab algorithm was implemented to evaluate PD bradykinesia. For this study we enrolled 4 patients undergoing DBS implant surgery. During an 8-hours session the patients were asked to fill in the diary and wear the smartwatch. A clinician assessed their condition compiling a Unified Parkinson’s Disease Rating Scale part III (UPDRSIII) form several times during these sessions. After the 8-hr trial, a clinician asked the patients if the smartwatch was uncomfortable during the day (options: “No” or “Yes”, if the answer was “Yes”, please explain why).

Results

At the end of the session, the data was collected and elaborated remotely with an optimized algorithm that finds the Bradykinesia Acceleration Scores (BAS) in data bins of 4 minutes (Figure 1A). The BAS algorithm was modified from the patent by Griffiths & Horne; BAS is lower when the patient is bradykinetic. The correlation between UPDRS III scores and the BAS for 5 different recording sessions, shown in Figure 1B, is -0.582 (P<0.009, Pearson). The bracelet was well tolerated (everyone answered “No” without further comments). Three of 4 diaries were correctly filled in by the patients. The correlation between UPDRS III scores and the diary scores were not significant.

Conclusions

The smartwatch was well tolerated and the mobile app was used by the majority of patients to fill in the diary, a cumbersome activity if done on paper. The mobile app was able to support data collection and the wearable device was reliable to monitor the motor symptoms of the patients without causing discomfort.

References

Promoting adherence to antiretroviral therapy in HIV-infected patients using mobile phone text messaging technology

Yuri QUINTANA PhD1,2, Eduardo Gonzalez MARTORELL MD4, Darren FAHY1

1Division of Clinical Informatics, Beth Israel Deaconess Medical Center, Boston, MA
2Harvard Medical School, Boston, MA

Abstract
We conducted a systematic review to determine whether mobile phone text-messaging or mobile apps are efficacious in enhancing adherence to ART in patients with HIV infection.

Introduction
An estimated1 36.7 million people were living with Human Immunodeficiency Virus infection worldwide in 2015. The impact of Human Immunodeficiency Virus/Acquired Immunodeficiency Syndrome (HIV/AIDS) has been particularly hard on countries with limited resources. Antiretroviral therapy (ART) can help people with HIV infection to live longer, healthier lives, but adherence to ART can be challenging. It is important to find better ways to help patients stay adherent to their ART regimens. Mobile phone text-messaging has the potential to help promote adherence in these patients.

Methods
We conducted a computerized literature searches with no date or language delimiter using the databases PubMed/MEDLINE (NCBI), Embase (Elsevier), Web of Science (Thomson Reuters), Cochrane Central (EBSCO), CINAHL (EBSCO). The literature search was completed on Date January 19, 2017. We included any intervention evaluated with a comparison group where the intervention was delivered digitally (not just voice alone) via mobile phone or smart phone. We excluded interventions that were not evaluated (no data collected for evaluation of the mhealth intervention), those interventions focused only on the health provider and not involving patients using the intervention, and papers that just reported on the feasibility of intervention. Adherence was measured with the following outcome metrics: adherence over time, EAMD cumulative adherence, missed dose past week, pharmacy adherence over time, self-reported adherence, self-reported adherence over time, viral load adherence.

Observations
We retrieved 1188 articles, and after removing duplicates, we had 555 articles. We found 489 were irrelevant and did not match our inclusion criteria. We reviewed 66 papers with full text and excluded 46, leaving 20 papers that matched our inclusion criteria. We reviewed these for bias using the Cochrane review methodology. Table 1 shows the types of bias we found in the study papers. We found that 7 of these described interventions with no improvement in medication adherence, and 13 had positive outcomes when compared to the control group.

<table>
<thead>
<tr>
<th>Type of Bias</th>
<th>Low</th>
<th>High</th>
<th>Unclear</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequence Generation</td>
<td>10</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>Allocation Concealment</td>
<td>11</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>Blinding of Participants and Personnel</td>
<td>4</td>
<td>15</td>
<td>1</td>
</tr>
<tr>
<td>Blinding of Outcome of Assessors</td>
<td>7</td>
<td>7</td>
<td>6</td>
</tr>
<tr>
<td>Incomplete Outcome Data</td>
<td>17</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Selective Outcome Reporting</td>
<td>17</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Other Sources of Bias</td>
<td>13</td>
<td>3</td>
<td>4</td>
</tr>
</tbody>
</table>

Conclusions
Some mobile applications are showing examples of successful implementations, but more rigorous methodology and documentation are needed to reduce the potential risks of bias.

Reference
MeSHgram: An Open Source Tool to Visually Browse Co-occurrence of MeSH Terms in PubMed

Satyajeet Raje, PhD¹, Ravi Teja Bhupatiraju, MBBS, PhD¹, Abdelrahman Hosny, MS², Ben Busby, PhD¹

¹U.S. National Library of Medicine, National Institutes of Health, Bethesda, MD, USA
²Center for Quantitative Medicine, UConn Health, Farmington, CT, USA

Introduction
We present MeSHgram as a tool for convenient, visual and interactive exploration of the co-occurrence of MeSH terms over the entire PubMed corpus. The tool can assist in the quantification of known research patterns as well as potentially aiding novel hypotheses generation.

Methods
We used the NLM PubMed corpus¹ as our data source. As of Jan 2017, the corpus contained approximately 24.5 million publications from 1809 to 2016². Publications indexed in PubMed have human curated Medical Subject Heading (MeSH) terms associated with them. We leveraged these MeSH terms in MeSHgram.

We parsed the entire PubMed corpus and extracted the ID, year of publication and MeSH terms associated with each document. We excluded duplicate items such as revision entries and those with no MeSH terms, resulting in approximately 23 million publications in our database (MongoDB). We use a Python web controller (CherryPy) and JavaScript widgets (nvd3, jqcloud) in the front-end for the implementation.

Features
The tool allows searching with multiple MeSH terms via an auto-complete search box. It displays article counts by year for the searched terms along with the number of co-occurrences. It also displays a word cloud for other MeSH terms that appear alongside the searched MeSH terms to provide contextual visualization of other associated MeSH terms. Visually selecting specific year ranges in the graph updates the graph and the word cloud.

Acknowledgements
This work is supported in part by an appointment to the NLM Research Participation Program administered by ORISE through an interagency agreement between the U.S Dept. of Energy and the NLM. MeSHgram was developed as part of a hackathon organized by NCBI in Jan 2017. MeSHgram is currently under development and available at www.meshgram.org. The source code is open and available on GitHub under MIT license.

Figure 1. Screenshot of MeSHgram displaying co-occurrences of MeSH terms "Rubella" and "Rubella Vaccine"

References
Reducing Uncoded Allergy Information in the Electronic Health Record

Harley Z. Ramelson, MD, MPH
1Brigham and Women’s Hospital, Boston, MA; 2Partners HealthCare, Wellesley, MA; 2Harvard Medical School, Boston, MA

Abstract
Through a series of user interface design modifications, the percentage of all uncoded allergies being entered by users was reduced from 18-23% to around 8%. The increased coded information allows drug-allergy checking to be performed thereby improving patient safety.

Introduction
The LMR is an in-house developed EHR that used across the Partners HealthCare enterprise. In order to address the high rate of uncoded allergies in the LMR, a series of user interface changes were initiated.

Methods
The following user interface changes were implemented across two releases.

1) Requiring users to go through the search process and review a listing of search results of allergens
2) On the search results page, the functionality allows a user to enter an allergen in the record that may not be present in the dictionary. This enhancement presented a message next to that allergen indicating that it is uncoded and will not be used for allergy alerts.
3) If the uncoded allergen was selected, a pop-up alert is presented to the user again emphasizing that the allergen is uncoded and will not be used in clinical decision support, reporting and performance measurements.
4) The allergy views were enhanced to include the label “uncoded” next to the allergen. This would prompt the user to be able to codify the allergen and remove it from the uncoded category.

Results
Prior to the changes being made, the percentage of allergens that were uncoded in the LMR ranged from 18% to 23%. Following the first set of changes, this was reduced to around 15%. Following the second set of changes, the percentage was further reduced to around 8%.

Conclusion
Modifications in the user interface in an electronic health record can reduce the percentage of uncoded allergies in an Electronic Health Record, thereby improving quality of the clinical data and increasing the effectiveness of clinical decision support allergy alerting.
Setting the Stage for Implementing an EHR-Integrated Social Knowledge Networking (SKN) System on Medication Reconciliation: Phase 1 Results
Pavani Rangachari, PhD
Augusta University, Augusta, GA

Background: Despite the policy impetus towards meaningful use of EHR Medication Reconciliation (MedRec) Technology (to prevent medication errors during transitions of care), hospital adherence has lagged for one chief reason—low physician engagement, stemming from lack of consensus about which physician is responsible for managing a patient’s medication list. In October 2016, Augusta University received a two-year grant from AHRQ, to implement a Social Knowledge Networking (SKN) system for enabling its health system, AU Health, to progress from “limited use” of EHR MedRec technology, to “meaningful use.” The hypothesis is that an SKN system would bring together a diverse group of practitioners across settings of care, to facilitate the exchange of tacit (practice-based) knowledge on issues related to EHR MedRec, which in turn, is expected to increase practitioners’ engagement in addressing those issues, to enable learning and meaningful use of EHR. The goal of PHASE 1 of the project (completed in April 2017), was to identify a consensus-set of issues related to EHR MedRec across multiple subgroups of practitioners involved in the MedRec process, i.e., physicians, nurses, and pharmacists. A two-round Delphi process (interviews and survey) was used to achieve this goal. This process in turn, facilitated the development of a Reporting Tool on Issues Related to EHR MedRec, which, along with an existing SKN tool (Microsoft Yammer); was integrated into the EHR at AU Health, at the end of PHASE 1. PHASE 2 (launched in May 2017), involves conducting a 52-week pilot test of the EHR-integrated SKN system in outpatient & inpatient units at AU Health. The purpose of this paper is to describe the results of PHASE 1.

Methods: Phase 1 included 15 individual semi-structured interviews with physicians, nurses, and pharmacists at AU Health, to identify issues related to EHR MedRec. Thematic Analysis helped identify 56 issue-items grouped under 9 issue categories. These results in turn, were used to conduct a survey a larger group of physician, nurses, and pharmacists at AU Health. The survey sought practitioners’ importance-rating of all issue-items identified from interviews, on a 7-point Likert Scale. A total of 140 responses were received (out of 225 recipients), for a 62% response rate. Data were subjected to 1) Descriptive Analysis of importance-ratings of issue-items; 2) Factor Analysis to validate issue categories and generate indices, and 3) Index T-test Analysis to assess differences in rating of issue-items by professional affiliation and by unit affiliation.

Results: Factor analysis served to validate the first 8 of the following 9 issue categories identified from Thematic Analysis of the interviews. On average, the issue items under each of the 9 issue categories were rated as “Important” or “Extremely Important” by 70%-90% of all survey respondents: 1) Care-Coordination Issues (89%); 2) Patient Education Issues (84%); 3) Workflow Issues (83%); 4) Resource Issues (82%); 5) Ownership & Accountability Issues (81%); 6) Process of Care Issues (81%); 7) Workforce Training Issues (79%); 8) IT-Related Issues (78%); and 9) Documentation Issues (74%). The Index T-test Analysis revealed no statistically significant differences (at the 5% significance level) in the importance-rating of issues by professional affiliation or unit affiliation, indicating a consensus on issues related to EHR MedRec. A central theme that emerged from findings was an absence of a shared understanding of 1) the responsibilities of the various practitioner groups involved in the EHR MedRec process; and of 2) the importance of EHR MedRec as a tool for preventing errors during care transitions. This central issue in turn contributed to an overall sense of skepticism regarding the value of EHR MedRec, leading to system workarounds, and “limited use” of the EHR MedRec functionality.

Conclusions: Results suggest that successful implementation of EHR MedRec would first and foremost, require a shared understanding of the process for MedRec among multiple stakeholder groups. Importantly, results indicate considerable potential for an EHR-integrated SKN system to enable knowledge exchange among practitioners across care settings, to address aforementioned issues, and enable engagement, learning, and EHR meaningful use. If the hypothesis holds, federal EHR vendors could be encouraged to incorporate SKN features into EHR systems.

References

This project was supported by grant number R21HS024335 from the Agency for Healthcare Research and Quality.
Use of the popHealth Open-Source Quality Measure Engine to Support Cardiovascular Care at Small- and Medium-Sized Practices

Luke V. Rasmussen¹, Andy Schriever², Pauline Kenly¹, Dejan Jovanov³, Stephen D. Persell¹, Theresa L. Walunas¹, Abel N. Kho¹

¹Northwestern University Feinberg School of Medicine, Chicago, IL; ²Park Street Solutions, Naperville, IL; ³Illinois Department of Public Health, Chicago, IL

Overview: The prevalence of electronic health records (EHRs) has grown since the advent of the Health Information Technology for Economic and Clinical Health (HITECH) Act of 2009¹, which incentivized adoption of EHRs through the Meaningful Use (MU) program. One aspect of MU has been recording and reporting of electronic clinical quality measures (eCQMs). As part of the AHRQ-funded EvidenceNOW network, the Healthy Hearts in the Heartland (H3) study has pursued the centralized calculation of cardiovascular health eCQMs from small- and medium-sized practices in Illinois, Wisconsin and Indiana. The use of a centralized calculation engine facilitated transparency and consistency in the calculation process across practices and vendors. Here we describe the use of the open-source popHealth software (https://github.com/OSEHRA/popHealth) as a means to that goal.

Methods: Four eCQM measures were implemented for the study: Aspirin, Blood Pressure, Cholesterol and Smoking (ABCS), using the HL7 Health Quality Measure Format (HQMF) definitions. This included the use of a draft Cholesterol measure accessible from the Measure Authoring Tool (MAT). Data were extracted from commercial EHR systems using a variety of methods (CCDA exports, integrated reporting workbench, or direct database access) and sent to a central study server for processing using software licensed from Park Street Solutions. The resulting data extracts were formatted as JSON documents in the format used internally by popHealth. Additional automation was built on top of the existing popHealth code base to streamline the management of quarterly data refreshes. An overview of the data flow and system components involved is shown in Figure 1.

Results: We established data feeds from 77 practices (representing seven different commercial EHR vendors) between June 2015 and February 2017, representing over 250,000 patient records. Currently 11 practice facilitators involved with the study and two practice office managers access popHealth to assess clinical quality scores.

Discussion: We experienced challenges obtaining sufficiently detailed patient data from some of the EHRs, including 1) a lack of batch export capabilities, 2) insufficient detail in CCDA and reporting exports (e.g., missing dates), and 3) a lack of historical data (e.g., exporting only the most recent blood pressure reading). Despite these limitations, we were able to demonstrate the popHealth platform to be a useful centralized eCQM calculation engine. The platform has been shown to be scalable, and has allowed aggregate eCQM calculations across health care networks and geographical regions. Ongoing work includes a formal validation of popHealth calculations against manual chart reviews. Since popHealth is capable of ingesting HQMF documents for measure logic, this solution may be easily expanded beyond the initial ABCS four measures with minimal effort. If the identified limitations can be overcome, popHealth may become a useful way for small and medium-sized practice groups to apply and report eCQMs and might eventually be used to support their participation in value-based payment or public reporting programs.

Acknowledgements: This research was supported by AHRQ grant number 5R18HS023921-02.

References

Exploring Barriers to the use of Electronic Health Records in the Trauma Room Using a Cognitive Work Analysis

Ryan J. Rasmussen MS, FNP-C, Doctoral Candidate1, Jane M. Carrington PhD, RN2
1 University of Arizona, Tucson, Arizona; 2 University of Arizona, Tucson, Arizona

Background. Trauma is one of the leading causes of death for individuals under 45 years old1 accounting for 192,000 deaths/year in the US in 20142. When both healthcare costs and lost productivity are calculated the total cost of trauma is more than 489 billion dollars annually in the United States alone3. When working in a multidisciplinary team of this size effective communication of critical information is often difficult to accomplish and information is delayed in reaching the intended target. The purpose of this presentation is to describe a planned research project where Cognitive Work Analysis (CWA) will be used to explore inter-trauma team communication to inform EHR design and adoption and increase patient outcomes.

Description of Problem. The trauma room is unique area of the hospital that employs complex communication patterns within and among trauma team members. Despite its promise to increase effective communication, the current electronic health record (EHR) has yet to meet the needs of the trauma team to support inter-team communication. Because of this most trauma rooms have not fully implemented or adopted the “real-time” use of an EHR during the care of trauma patients. The Effective Nurse-to-Nurse Communication framework describes nurse-to-nurse communication of a clinical event4 and provides a better understanding of how nurses communicate with each other and other parts of the medical team. This and Information Theory guide this research5. The objective of this study will be to understand the barriers to the use of electronic health records in the trauma room by increasing our understanding of intra-trauma communication.

Method. A CWA is a method of research designed to analyze a workplace and guide the use of technology within that workplace5. CWA has the unique ability to look at a current system, analyze the system, and suggest recommendations for future implementations and designs. CWA has great potential for contributing to the improvement of inter-team communication and facilitate the effective use of an EHR within the trauma room. Consistent with CWA described by Jenkins et al5, interviews and observations will be used to inform current inter-team communication. Interviews with trauma team members will be digitally recorded and transcribed for analysis. Observations will be documented using video recordings. The trauma room will be equipped with both video and audio recording devices that will be activated when the trauma patient arrives. These recordings will be analyzed for actual patterns of communication and documentation. Interviews will be analyzed for perceived patterns and barriers to communication and documentation. Data will then be analyzed according to the five domains of CWA: work domain analysis, control task analysis, strategies analysis, social-organizational and cooperation analysis, and worker competencies analysis6.

Impact. The findings of this study will increase our understanding of inter-team communication in the trauma room and result in suggested improvements to increase patient outcomes. This study will also inform design of the EHR that supports inter-team communication during trauma.

Conclusion. Increasing our understanding of intra-team communication during trauma can lead towards successful implementation and adoption of the EHR in the emergency department trauma room and contribute to increased patient safety.

References
Introduction
The United States Critical Illness and Injury Trials Group Program for Emergency Preparedness (USCIITG-PREP) has been coordinating multi-site exploratory studies to identify and overcome challenges associated with prospective data collection during public health emergencies. Efficient data collection is essential in this context, particularly if new medical countermeasures are being evaluated. An approach being considered is the population of clinical case report forms (CRFs) directly from Electronic Health Record systems (EHRs) in order to expedite real-time data collection while minimizing additional burden upon clinical providers. Eighty-three percent of sites from our 2016 Influenza Study reported transcription of up to 98.5% of CRF fields directly from an EHR. Integrating research data collection workflows within existing EHR clinical workflows has the potential to significantly decrease time and effort requirements for all providers involved with a given study. The overall aim of this project is to integrate research CRFs with EHRs, including reuse of data previously collected for clinical care.

Methods
USCIITG-PREP has initiated a new feasibility study in November 2016, focused on the treatment of severe cases of Influenza requiring Intensive Care Unit (ICU) admission. A new CRF has been created and implemented using an open-source and stand-alone electronic data capture tool. The new study provides the opportunity to evaluate the integration of the same CRF with a commercially available EHR system. This evaluation includes four steps: 1) selection of relevant EHR data and EHR data collection tool(s), taking into account typical ICU clinical workflows; 2) identification of CRF fields that map to structured clinical data elements (SCDEs) within the EHR; 3) validation of identified fields with clinical domain experts; and 4) implementation of prototype CRF integrated with the EHR, aligned with existing clinical workflows and including opportunities for data reuse. Once the prototype is implemented, we will explore options to de-identify and aggregate data from different sites, and also describe the process to disseminate the CRF to sites using the same EHR system. Throughout the evaluation, detailed documentation about available alternatives and challenges will be created, as well as records outlining resources and effort to implement the prototype.

Results
The Influenza Study CRF has 82 fields total, including details about laboratory confirmation, disease severity, treatment, and outcomes. The CRF includes 29 boolean, 18 single-select, 5 multi-select, and 30 free text fields. When identifying relevant EHR fields, special consideration must be taken to confirm data is being retrieved from the correct encounter and with the appropriate format. For example, dates might need to be converted to time periods relative to the first ICU admission. In the Influenza Study eCRF eight single-select fields represent relative dates. We have started steps 1 and 2 and plan to complete the prototype implementation by May 2017. The current data entry tools being considered include structured forms and flowsheets within the EHR. Preliminary results reveal challenges in translating dates to relative date fields, as well as calculating number of treatment days and obtaining details assessed during admission.

Conclusion
Data entry burden on clinical providers is a prominent barrier to multi-site and large-scale clinical research studies. This barrier is even more pertinent to prospective studies during public health emergencies, when real-time data collection is essential. USCIITG-PREP’s effort to evaluate EHR data collection tools is key to decreasing data entry time and improving the ability for rapid dissemination during emergent events. A CRF integrated with an EHR can support reuse of existing EHR data and decrease the need for manual data collection and transcription to a stand-alone data capture tool. If integration with a single EHR is successful, future evaluation will be necessary to understand how easily the EHR-based CRF can be disseminated to sites using disparate EHR systems.

Acknowledgements: This project is funded by FDA and BARDA in collaboration with USCIITG-PREP.

References
Examining Nurses’ Use of Electronic Health Records to Document Symptoms in Acute Care Settings

Blaine Reeder¹, Mustafa Ozkaynak¹, Mary Beth Flynn Makic¹, Karen H. Sousa¹
¹University of Colorado College of Nursing, Aurora, CO

Introduction
A person’s perception of their own symptoms is a prominent factor in the decision to seek acute-care admission. Nurses perform symptom assessment and management as a core part of their practice and these activities longitudinally influence patient symptoms. Knowledge of how nurses’ use electronic health record (EHR) systems to document patient symptoms is incomplete, though clinical experiences of our research team members indicate that nurses often document symptom data in narrative notes. Our recent systematic review identified the need to improve symptom assessment documentation and design of systems that support this activity in acute care settings [1]. Accepted wisdom is that nurses document symptoms using the EHR but the extent to which this happens is unknown. Therefore, the motivation and aim of this qualitative descriptive study is to examine how nurses use the EHR system to document patient symptom assessment and management in acute care settings.

Methods
We conducted semi-structured interviews with nurses (n=14) who work at inpatient units of an academic hospital in the Mountain West region of the United States. Interview questions related to patient care responsibilities, patient assessment, symptom management, work experience, education and training, and use of information technologies to support patient care. The first and second authors conducted all interviews. A codebook for data analysis was iteratively developed with input from all authors. A data-driven, inductive coding approach was employed with first and second authors independently coding transcripts and then meeting to reconcile differences. After three transcripts, the first author coded the remaining transcripts with periodic checks by the second author for code drift.

Results
Participants generally had a positive view of the EHR as tool to support the patient care they provide, noting that symptom documentation and overall patient care was important. A common response was: “If it wasn’t charted, it wasn’t done” (Participant 8). Beyond patient care, a primary motivation for documenting patient care was for legal reasons.

With regard to general patterns of use, nurses reviewed other nurses’ documentation going back only 24 hours. Not surprisingly, symptom documentation varied by unit due to patient conditions. Symptoms reported as important are: pain, numbness, nausea, vomiting, numbness, tingling, alteration in cognition, dizziness, shortness of breath, dry mouth, cough, weight loss, constipation, mouth sores (oncology). As noted by one participant: “Pain is a huge symptom that is a problem in all patient populations” (Participant 7).

One identified challenge is that nurses are instructed to “chart by exception” though some participants disagreed with this practice due to loss of individual information about each patient. Another identified challenge is that increased patient care needs (workload) decreases quality of documentation (e.g.: timeliness and completeness). In these circumstances, nurses prioritize patient care over documentation. Other common barriers to system use are: system downtime, freezing or hanging; checklists that do not portray the patient; having to search where to document; charting the same thing in multiple places; and finding where other nurses have documented patient information. Some nurses identified the practice of copy/paste as contributing to poor documentation.

Conclusion
Our findings indicate opportunities to better support nursing practice by redesigning EHR interfaces and providing longitudinal reports of patient symptom status from nursing documentation. Full results and design recommendations from this study will be presented at the AMIA Annual Symposium 2017 in Washington, DC.

Acknowledgements
This study was supported by funding from the Colorado Collaborative for Nursing Research (CCNR).

References
Natural Language Data Sampling: Principles & Strategies in Corpus Selection

Ruth M. Reeves, PhD1,2, Nancy H. Gentry, RN, MSN1,2, Glenn T. Gobbel, DVM, PhD, MS1,2, Fern FitzHenry, RN, PhD1,2, Michael E. Matheny, MD, MS, MPH1,2

1Tennessee Valley Healthcare System VA, Nashville, TN; 2Vanderbilt University, Nashville, TN

INTRODUCTION: Designing a model of study phenomena allows hypothesis generation, testing and evaluation of the model’s validity against real-world data. But drawing inferences from the model are unlikely to prove valid outside the model unless its design incorporates methods of representing variation within the population under consideration. To achieve adequate representativeness of the data available, researchers sample some portion of it. Sampling from a patient cohort will involve considerations of the characteristics of the clinical model to determine suitable data types to represent variations across the population and how to stratify, for example, demographic data or diagnostic categories versus how to weight study variables such as lab, medication or behavioral values. While clinical models are often built with multiple data streams and require complex analytical methods to achieve integration into the model, it is important to distinguish a stream of raw data from a data stream which is itself the output of a model. Sampling patients by age group uses raw data (number of years since birth), while sampling patients whose clinical notes contain references to dehydration makes use of a language model. Building a natural language (NL) model for use in providing information from NL data sources differs in significant ways from building a model of clinical outcomes or clinical risks. Clinical models making use of information extracted from NL text need to intersect with the NL model associated with the clinical phenomena, but a successful NL model requires independent sampling methodology. METHODS: We reviewed the document sampling methods in 4 Natural Language Processing (NLP) projects aimed at extracting specific concepts for downstream analytics, each with differing corpora and targeted clinical concepts, all using narratives documents within the VA medical record system. The cohort selection for each project was performed prior to the document sampling, using structured data pertinent to each project (e.g., ICD codes, drug orders, lab values). Our corpora were derived from these cohorts and included the following four data sets: CCC: Cardiac Catheterization Clinical Note Corpus (starting corpus of 9,007,164 narrative notes); CRR: Cirrhosis Risk Radiology Report Corpus (starting corpus of 91,881,399 radiology reports); OIC: Opioid Induced Constipation Clinical Note Corpus (starting corpus of 1,158,681,416 narrative notes); PTSD: PTSD Treatment Mental Health Corpus (starting corpus of 4,090,310 narrative notes). Rather than risking possible over-fitting by pre-mining documents for instances of targeted concepts, we used available meta-data to construct each project corpus. The possible document features used in filter and categorize the documents were (1) document title, (2) provider type of document author, (3) clinic type producing the document, (4) document distribution across clinical workflow of targeted cohort (for example, ER, admission, procedure, inpatient progress, discharge, outpatient progress), (5) document distribution across cohort-specific determinates (e.g., over-sampling by age, gender, etc.). We were interested in finding strategies producing corpora with a high yield of targeted clinical concepts. We determined the proportion of the corpora that yielded documents with no targeted clinical concepts, as a (negative) measure of sampling success by assessing the annotated (tagged) sub-samples of each corpus for documents yielding zero annotations. RESULTS: Table 1 illustrates the feature sets used to generate each corpus, and the rate at which each produced zero concept documents. The feature set used to form the CCC performed the best (2% rate), using all 5 features; the worst being the OIC (40%) rate, using feature sets 1,2 and 5.

Table 1. Document Meta-Data Feature Sets Forming NLP Project Corpora with 0-Concept Document Rate

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Unfiltered Corpus</th>
<th>Feature Sets</th>
<th>Filtered Corpus</th>
<th>Tagged % 0-hits</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCC</td>
<td>9,007,164 narrative notes</td>
<td>(1); (2); (3); (4); (5)</td>
<td>1,256,685 clinical notes</td>
<td>1,568</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2%</td>
</tr>
<tr>
<td>CRR</td>
<td>91,881,399 radiology reports</td>
<td>(1); (2); (3); (5)</td>
<td>6,024,284 thorax image reports</td>
<td>3,600</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5%</td>
</tr>
<tr>
<td>OIC</td>
<td>1,158,681,416 narrative notes</td>
<td>(1); (2); (5)</td>
<td>11,707,693 clinical notes</td>
<td>3,840</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>40%</td>
</tr>
<tr>
<td>PTSD</td>
<td>4,090,310 narrative notes</td>
<td>(1)</td>
<td>116,162 mental health notes</td>
<td>911</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>23%</td>
</tr>
</tbody>
</table>

DISCUSSION: Successful document sampling resulted from filtering out irrelevant note types and balancing distribution across document features. The ability to filter and categorize by document feature depends to a large extent on the quality of the information in the feature set. For example, document titles among nursing notes tend to be highly underspecified. The sampling strategist can often rely on combining feature sets to produce better results. The corpus with the worst performance, the OIC had a very high rate of zero annotation documents, due to the inclusion of nursing notes authored from mental health services, most of which were not germane to the concepts (opioid use, constipation, constipation treatment, constipation procedures). Filtering nursing notes by clinic type (in essence, applying feature 3) would have reduced the zero-annotation documents to a rate of 14%. Acknowledgements: This research was reviewed & approved by Tennessee Valley Healthcare System (TVHS) IRB, Nashville VA facility. The material is based on work supported by Veterans Health Administration, Office of Health Services Research & Development, with resources and use of facilities at the VA TVHS. Funding sources: VA HSRD IIR 12-362; VA HSRD IIR 13-052; VA HSRD HIR 09-003; IIR 292-1
Creating a Corpus Resource for Text Simplification R & D

Debra Revere, MLIS, MA¹, Partha Mukherjee, PhD², David Kauchak, PhD³, Gondy Leroy, PhD²

¹Univ of WA, Seattle, WA; ²Univ of AZ, Tucson, AZ; ³Pomona College, Claremont, CA

Introduction

Text simplification (TS) involves rewriting complex text into simpler language that is easier to understand while retaining meaning. An important TS resource is text in which the difficulty of the text is known. Previous corpora have either relied on the text source (e.g., blogs vs. medical articles) or human perception to judge difficulty. These assessments have some value, but are subjective. To better quantify text difficulty, we have created a collection of texts and questions (Qs) which test knowledge (As) within those texts in order to systematically quantify text difficulty. Our Q/A process combines auto-generated and human-annotated methods to create a training corpus from texts of varying difficulty within the medical domain. To minimize factors that might influence performance other than text difficulty, we standardized the process using Heilman's (2011) framework which parses each sentence and then uses a rule-based approach to generate Qs based on syntactic structure¹.

Text Simplification Corpus Creation

Following auto Q/A generation, annotators received one Q/A file for 60 medical conditions that contained: 1) auto-generated Qs in Wh- (i.e., What, When, etc.) and True/False formats; 2) the A embedded in its complete original text; 3) the A; and 4) a score which indicated viability of the text to answer the Q. Files with scores lower than 2.5; illegible characters (e.g., HajduGČôCheney syndrome); or non-medical topics were eliminated—leaving a total of 8,637 Q/A item sets across 56 articles. Human annotators randomly selected 36 articles on common (e.g., dyslexia) and uncommon (e.g., Spondylocostal dysostosis) topics to ensure a wide variety of language and complexity. Annotators corrected grammatical errors and conducted up to 3 manual simplifications per Q/A item set to create a Q/A set that was simpler but retained lexical structure—syntactically reducing complexity by splitting sentences, eliminating non-standard abbreviations, or rephrasing strings containing double negatives. Table 1 provides annotation examples.

Table 1. Auto-generated Q/A item set input and human annotated Q/A item set output examples.

<table>
<thead>
<tr>
<th>Auto Q/A</th>
<th>Annotated Q/A</th>
<th>True/False-Q</th>
<th>Rule-Based-Q</th>
</tr>
</thead>
<tbody>
<tr>
<td>Is the disease related to the SRRT gene?</td>
<td>The disease is related to the SRRT</td>
<td>Spondylocostal dysplasia</td>
<td></td>
</tr>
<tr>
<td>What disease is related to the SRRT gene?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>What is acute rheumatic fever rare in children?</td>
<td>In children, acute rheumatic fever is rare in most of the developed world.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>In most of the developed world is prevalence of acute rheumatic fever rare in children?</td>
<td>YES</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Results

The final training corpus contains 1264 questions: 670 Wh-Qs with 4 annotated multiple-choice As (1 correct, 3 incorrect) and 594 True/False-Qs with YES, NO, Sometimes and Not enough information As.

Conclusion

This corpus is the foundation for future TS work in which we will create smaller text fragments with matching Qs at different difficulty levels and generate comprehension data for readers with varying levels of literacy.

Acknowledgments: Research reported in this poster was supported by the National Library of Medicine of the National Institutes of Health under award number R01LM011975. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

Family Data Linkage Methods Using EHR Data

Pedro Rivera, MS¹, Megan Hoopes, MPH¹, Stuart Cowburn, MS, MPH¹, Jon E. Puro, MPA¹
¹OCHIN, Inc., Portland, OR

Abstract

Although many EHRs have functionality for documenting family relationships among patients, explicit linkage between the health records of family members is still lacking in many systems. Such linkage is important for research on pregnancy, birth, health policy, childhood outcomes¹, and the study of conditions with genetic links. We present a generic methodology to impute reliable family linkages using EHR data commonly shared among family members, focusing on maternal-child relationships. A gold-standard dataset of explicit family relationships was utilized for validation.

Introduction

This study describes methods used to link family member data in electronic health records. To accomplish this linkage, we identified a variety of ‘target’ data elements, commonly captured in EHR systems that could potentially be used to establish family linkages. We then conducted a series of experiments to identify potential family linkages. Each experiment utilized a different combination of target data elements to predict family linkages. The results of each experiment were compared to a gold-standard dataset of explicit family linkages to determine the sensitivity and precision of each experimental approach. We focused on mother-child links for two reasons: (1) we felt we could more reliably identify these from available EHR fields than paternal relationships, and (2) areas of research interest requiring these linked data focus on reproductive health and early childhood outcomes, in which maternal factors are of greatest direct interest.

Evaluation Results

We identified 382K patient-mother links in a dataset of 2.1 million patients who had one or more face-to-face visits in the last 5 years. Only 44.5% of these linkages were explicitly documented in the EHR. The number of links was considerably higher for younger patients (link rate for patients ≤ age 18=55%).

Our gold-standard consisted of 25.5K patient-mother links explicitly documented in the Epic EHR both as an emergency contact and as a family relationship in the billing guarantor account(s). The use of home phone numbers as the common data element to group patients into potential household units resulted in the highest number of imputed links (sensitivity 0.8, precision 0.99), followed by Oregon Medicaid case number (sensitivity 0.34, precision 0.98) and geocoded last known address (sensitivity 0.18, precision 0.94).

Conclusion

These data lay the groundwork for future studies requiring data of family members be linked, particularly mothers and their children. This study shows that it is possible to impute such family linkages using standardized methods and available EHR data. A limitation of the current work is the focus on linking maternal-child relationships; future work would benefit from extending this method to paternal relationships and broader family units. Such methods would be important to the study of neighborhood-level factors (i.e., the social determinants of health) and conditions with a genetic component.

References

Sub-Phenotyping of Crohn’s Disease Using a Large Electronic Record Cohort

Jamie R. Robinson, M.D., Lisa Bastarache, B.S., Robert J. Carroll, Ph.D., Elizabeth A. Scoville, M.D., David A. Schwartz, M.D., Joshua C. Denny, M.D., M.S.
Vanderbilt University Medical Center, Nashville, TN

Introduction
Crohn’s disease (CD) is a chronic inflammatory disease of the gastrointestinal tract with increasing prevalence worldwide. The phenotype of CD is marked by variable clinical features and disease course, with the etiology thought be multi-factorial. Previous studies have suggested a genetic basis for the clinical heterogeneity of this disease. We sought to study the influence of known CD genetic risk loci on clinical heterogeneity via case-only phenome-wide association studies (PheWAS).

Methods
A retrospective, cross-sectional study was performed using the Vanderbilt University Medical Center (VUMC) Synthetic Derivative (SD), a de-identified version of the electronic health record linked to the VUMC biorepository, BioVU. A validated phenotyping algorithm was used to identify individuals with CD. Individuals genotyped on the Illumina HumanExome array platform with principle-component analysis indicating European ancestry were analyzed using a case-only PheWAS for each genetic risk loci, with pairwise comparisons conducted using logistic regression, adjusted for age and sex. Essentially, for each PheWAS, cases and controls were drawn only from among CD cases.

Results
360 individuals with CD meeting selection criteria were identified with extant genetic data. 38 single nucleotide polymorphisms (SNPs) associated with CD were reproduced in this cohort. In a case-only PheWAS, rs6596075 within IBD5 was associated with coagulation defects (OR 3.35, p=1.2x10^-5), acute post-hemorrhagic anemia (OR 2.94, p=7.2x10^-5), and pulmonary insufficiency after trauma (OR 4.38, p=2.7x10^-5) (Figure 1A-B).

Conclusion
This study shows the ability for case-only PheWAS to identify disease comorbidities. The polymorphism with greatest association with Crohn’s disease, nucleotide-binding oligomerization domain 2 (NOD2), showed no association with Crohn’s disease subtypes, suggesting less prevalent polymorphisms, as seen within IBD5, may contribute more significantly to disease heterogeneity. Larger studies are needed to fully characterize these effects.

References
Critical Factors for Developing a Personalized Warfarin Dosing Protocol: A Scoping Review Analysis

Jacee Robison1, MS, RN & Jia-Wen Guo2, PhD, RN
13M Healthcare Information Systems, Murray, Utah, U.S.A.
2University of Utah College of Nursing, Salt Lake City, Utah, U.S.A.

Abstract
Currently there is a lack of dosing protocols for Warfarin which encompass all the essential information for personalizing dosing. A scoping review was performed in this study to extract factors relevant to the use of Warfarin from existing literature and clinical documents. We identified factors vital in personalizing Warfarin dosing and grouped them into four components as genomic, demographic, health behaviors and clinical. This result can be used to develop a protocol for personalizing warfarin dosing.

Introduction
Warfarin is associated with a high incidence of serious adverse drug events due to its narrow therapeutic range. Patients metabolize warfarin differently, therefore making it important for individualized warfarin dosing to be more commonplace. Using a personalized approach would allow patients more accurate treatment and prevention strategies. However, there are some patient populations understudied and important patient factors left out in current dosing protocols. The research question for this study was: what essential factors are needed in order to develop a protocol for personalizing warfarin dosing?

Purpose
The purpose was to identify the critical factors that need to be considered when developing a protocol for personalizing warfarin dosing from existing literature and clinical guidelines.

Methods
A scoping review study was performed to identify factors which are important in determining the Warfarin dosage from existing literature and clinical guidelines. Following Arksey and O’Malley’s framework for scoping reviews, we started from conducting literature search based on the proposed research question, screening and selecting the relevant publications, extracting data by using narrative review, to summarize and report the results. We searched PubMed with the terms including Warfarin, protocol, prescription, dose, pharmacogenomics, personalized medicine, and precision medicine from January 2007 to February 2017. Google Scholar was also used to retrieve grey literature. Thirty-two articles and documents were relevant and included in the review process. Two reviewers independently extracted data from the retrieved articles.

Results
Inter-reviewer agreement regarding the four critical factors was assessed and the results were 99.91% in agreement, with a Cohen’s Kappa score of 0.93. The critical factors needed to develop a protocol for personalizing warfarin dosing were divided into four domains: genomic factors, demographic factors, health behaviors and clinical factors. Genetic variations play a significant role in warfarin metabolism. For example, CYP2C9, one common polymorphism, results in slowing down the body metabolism of Warfarin; VKORC1 causes the dose-anticoagulant effect of Warfarin. Demographic factors such as ethnicity, sex and age impacts Warfarin metabolism; age affects the ability to metabolize medications. Health behaviors including smoking status, diet and alcohol use also affect warfarin metabolism; alcohol use can lead to an increased risk for bleeding. Additionally, clinical factors such as drug interactions, co-morbidities and those receiving cancer treatment are important to consider when dosing; those taking amiodarone can experience a decrease in Warfarin metabolism requiring dosing adjustments.

Discussion
Further research needs to be conducted for those currently undergoing cancer treatment and those with kidney and liver diseases in order to better understand how these individuals process Warfarin. There is also a need for Clinical Decision Support development in order to make the complexity of the genomic data easier for clinicians. Comparative effectiveness research (CER) would be needed in order to accurately compare benefits and harms of alternate methods and protocols. This would assist patients, clinicians, policy-makers and purchasers to make informed decisions thereby improving healthcare in the individual and in the community.
Towards the Effective Management of EHR Configuration Integrity

Roberto Rocha MD, PhD,1,2,3 Priyaranjan Tokachichu MD, MS,1 Margarita Sordo PhD,1,2,3 Christopher Vitale, PharmD,1 David Dubois, MBA,1 Saverio Maviglia MD, MS1,2,3
1Clinical Informatics, Partners eCare, Partners Healthcare System, 2Brigham and Women’s Hospital, 3Harvard Medical School, Boston, MA

Introduction
An important maintenance challenge associated with electronic healthcare records (EHRs) is the need to detect and resolve broken configuration dependencies. For example, we recently updated diagnosis terms associated with ICD-10-CM to accommodate changes introduced with the FY 2017 release. Within our EHR system, terms were automatically remapped, but other system configuration items, such as ICD-based value sets, were not updated. As a result, problem list entries had the correct new codes, but other configuration items had to be manually updated to remove deprecated codes and/or include new ones. Likewise, patient data records that refer to deprecated codes continue to exist within the EHR, but subsequent use of these records requires mechanisms to recognize that new codes might supersede deprecated ones, while other codes are simply no longer valid. Similar configuration integrity challenges exist for many dictionaries and configuration artifacts used by EHR systems, but mechanisms to effectively detect and manage dependencies are not always available. At Partners, we are using a commercially available knowledge management system (KMS)[1] to represent and manage high priority configuration dependencies. Our focus is on configuration items typically associated with clinical decision support rules, and our goal is to use the KMS to identify and remediate broken dependencies proactively.

Methods
The KMS used at Partners allows the representation of any data or knowledge structure using configurable semantic models. A few dictionaries from our EHR system are represented in the KMS. For each dictionary, only high utility fields are considered. Each EHR dictionary becomes a semantic entity in the KMS, and each field is translated into either an attribute (e.g. numbers, dates, strings) or a relation (e.g. semantic-enabled links). Once created, relations provide the optimal mechanism to represent and track structural and semantic dependencies. Each dictionary record from the EHR becomes a valid instance of the corresponding semantic model in the KMS. The KMS includes robust mechanisms for versioning instances, and also enforces the integrity of each dependency (link) using a well-defined lifecycle. Lifecycle validation considers the state of each entity instance, with more restrictive rules as instances move from work-in-progress state to in-active-use state.

Results
We have implemented semantic models for two EHR dictionaries. The first dictionary represents diagnoses, and its model includes 33 EHR-specific properties (e.g. attributes and relations). The second dictionary represents value sets, and its model includes 45 properties. We are importing dictionary records extracted from the EHR, resulting in over 2 million diagnosis instances (multiple revisions), and over 3,500 value set instances. Many value set instances include links to diagnosis instances, enabling our team to track and manage dependencies as these dictionaries are updated and used within the EHR. Modeling each dictionary required 40-60 hours of an experienced informatician.

Conclusion
Effectively managing large number of dependencies (links) is an important challenge that requires well-defined processes and advanced tools. We are addressing this problem using a sophisticated KMS and expect to proactively identify and remediate broken configuration dependencies within our EHR, ideally before resulting in system malfunctions.[2] The KMS allows us to manage structural dependencies and also enables the use of semantic expressions to validate the meaning of the links. Efforts to manage link integrity within linked data initiatives offer promising methods that we plan to evaluate and compare against our current approach.

References
Automated Extraction of Pediatric Obese and Allergic Related
Asthma Phenotypes from the Electronic Health Record

Mindy K. Ross, MD¹, Javier Sanz¹, Douglas Bell MD PhD¹
¹University of California Los Angeles, Los Angeles, Ca USA

Introduction

Asthma is one of the most common chronic childhood illnesses, affecting over 6 million children in the United States. One challenge with asthma is its variability in underlying inflammation and treatment response. In order to understand this variation, different categories of asthma patients have been described that reflect underlying pathophysiology (i.e. allergic-asthma and obese-asthma).

These specific asthma phenotypes have been identified with data from protocolled clinical trials; however, they have not been extracted from the electronic health record (EHR) in an automated fashion. While some literature has reported on methods to identify asthma cases from the EHR, there are no widely used computable phenotype algorithms for general asthma, let alone specific categories of asthma. We feel that it is important to automatically extract patient cohorts from the EHR for applications such as clinical decision support and identifying patient cohorts for research.

Typically, reliance on ICD codes alone is not sufficient to accurately extract disease cases. We are in the early phase of the development of algorithms to extract pediatric asthma patients who meet strict criteria for allergic-not-obese (A⁺/O⁻), obese-not-allergic (A⁻/O⁺), allergic-and-obese (A⁺/O⁺), and neither-obese-nor-allergic (A⁻/O⁻). The aim is to define four separate groups for future application to decision support and predictive analytics.

Methods

For this IRB-approved study, our EHR (Epic Systems) Clarity data warehouse was queried from March 1, 2013 to present. There are approximately 66,000 adults and children with the diagnosis of asthma documented in their record. Algorithm rules were based on criteria described from previous studies of identifying cases of asthma, obesity, and allergic-related conditions. ¹,² Asthma: Our definition for an asthma patient included a) ICD-9 or ICD-10 codes of 493.*, J45* and b) ≥4 prescriptions for short-acting β-agonist rescue inhalers or ≥1 controller medication AND ≥1 short-acting β-agonist rescue inhaler based on the high-specificity algorithm by Paecheco, et al. The exclusion criteria were ICD-9 or ICD-10 diagnoses of any diagnoses that could lead to an inaccurate diagnosis of asthma or have pulmonary involvement as a complication of the underlying condition such as: chronic lung disease, pulmonary fibrosis, neuromuscular, cardiovascular, transplant, immunodeficiency, oncologic, metabolic, autoimmune disease, etc. Obesity: Based upon the PheKB algorithm for childhood obesity, we defined obese as a body mass index (BMI) >95 percentile. Allergy: We evaluated diagnosis based upon ICD-9 or ICD-10 codes (i.e. allergic rhinitis), laboratories (i.e. specific allergen IgE), and medications (i.e. ≥1 prescription over time of any combination of allergy medication).

Results

The number of pediatric patients with an ICD code of asthma was 7,746. After exclusion criteria were applied, 6,450 patients remained. When asthma medication was included, there were 3,606 patients that were labeled as asthma. We next evaluated allergic asthma patients based off of this cohort. Using allergy-related ICD codes, there were 1,924 patients identified. After allergy labs and medication criteria, 364 patients qualified as allergic-asthmatic. In regard to obese patients, we are currently working to develop a method to extract the pediatric body mass percentile from the data warehouse in an accurate manner, as it is not directly calculated in a manner similar to the BMI and is not stored in easily accessible locations. Limited sampling of our data appears to be accurate, but once we complete the identification of the obese-related cohorts, our next step is to more extensively manually review each category of phenotype for sensitivity (recall) and positive predictive value (precision) of the definition algorithm.

Discussion/Conclusion

Accurate identification of complex medical conditions from the EHR can be achieved with cautious definition of inclusion criteria. This is the groundwork to cohort identification for applications such as clinical decision support. In addition, this will be valuable for clinical trial cohort identification and application toward precision medicine.

² https://phekb.org/phenotype/severe-early-childhood-obesity
Can Automated Retrieval of Data from Emergency Department Physician Notes Enhance Radiology Order Entry Process?

Justin F. Rousseau, MD, MMSc1,2, Ivan K. Ip, MD, MPH1, Ali S. Raja, MD, MBA, MPH1, Vladimir I. Valtchinov, PhD1, Jeremiah D. Schuur, MD, MHS1, Ramin Khorasani, MD, MPH1

1Brigham and Women’s Hospital, Harvard Medical School, Boston, MA; 2Dell Medical School at The University of Texas at Austin, Austin, TX

Abstract

Physician notes completed prior to image ordering were reviewed from 666 Emergency Department (ED) encounters with a chief complaint of headache and a completed head CT during the visit. We compared natural language processing-extracted concepts relevant to headaches from notes to image order requisitions. History of present illness was signed in 23% of encounters prior to image ordering. Significantly more concepts relevant to headache were extracted from notes than orders (median 3 vs. 1; p<0.0001).

Introduction

Clinicians are often required to enter redundant data in the electronic health record (EHR). For instance, in most commercial EHR implementations, clinical information is required both in the history of present illness (e.g., symptoms) in notes and in imaging order requisitions, often in structured form for quality and billing purposes, in the same EHR. Even when requisition forms are customized by imaging study, workflow limitations often lead clinicians to select the minimum number of indications required to order the imaging study. When a paucity of clinical information is communicated from ordering physicians to radiologists at the time of order entry, suboptimal imaging interpretations and patient care may result. We sought to evaluate clinical data documented in EHR notes prior to image ordering for ED patients presenting with headache. We hypothesized that data are present in the notes that could potentially be used to augment or prepopulate codified indications provided in the image order requisition.

Methods

In this Institutional Review Board-approved, retrospective observational study performed between 4/1/2013 and 9/30/2014 at an adult quaternary academic hospital, we evaluated data documented in EHR notes prior to image ordering for Emergency Department (ED) patients presenting with headache. We reviewed data from 666 consecutive ED encounters for patients with chief complaint of headaches with head CT performed during the visit. We compared the count of relevant Concept Unique Identifiers (CUIs) specific to headache in the ED notes, extracted via ontology-based natural language processing, to the count of indications in imaging order requisitions. We evaluated the percentage of head CTs graded by physician graders as appropriate based on extracted concepts compared to requisition indications.

Results

The history of present illness note section was completed prior to image order entry in 23% (154/666) of encounters. The number of CUIs specific to headache per note (median=3) was significantly greater than the number of indications per encounter identified in the imaging order requisition (median=1, the minimum required; Wilcoxon signed rank p<0.0001). Extracted concepts were unique from order requisition indications in 93% of cases. Extracted concepts added value in 23% of encounters when available by providing information to support the grading of the head CT as appropriate when the indications alone were not enough to grade the head CT as appropriate.

Conclusion

We identified unique, relevant clinical information present in unstructured physician notes in the EHR that were completed at the time of image ordering, with a significantly greater number of CUIs specific to headache extracted from these notes than the number of indications per encounter identified in the image order requisition. EHR clinical documentation provides a source of valuable information that, when available, could be used in a semi-automated fashion to augment communication between ordering physicians and radiologists. Future work is needed to evaluate whether automated retrieval of data from physician notes can inform clinical decision support and prepopulate imaging order requisitions to optimize efficient communication of clinically useful information in the ordering process.
Physician negation of nursing concepts in the electronic health record

Khawllah Roussi¹, Karen Dunn Lopez¹, Barbara Di Eugenio¹, Andrew D. Boyd¹
¹University of Illinois at Chicago, Chicago, IL;

Introduction

Electronic health records provide a shared understanding among health professions. Recently, nursing data has been shown to predict hospital readmissions.¹ Prior work has focused on the lack of common vocabulary between physicians and nurses (physicians only used 21% of nursing concepts).² Negation of a term means denying the existence of the concept and is patient dependent, (i.e. patient denies chest pain). In our prior study² negation was included in the use of concepts, but what the professional is trying to express are contrasting ideas. The study objective was to determine the frequency of term negation related to professions’ utilization of each others concepts.

Methods

Seven years of physician discharge summaries were processed by MedLEE (a medical information extraction program) to generate 53,423 physician² concepts: Unified Medical Language System (UMLS) concept unique identifiers (CUIs) and negation frequency. The discharge summaries were both free text and smart form summaries. A list of negated terms were generated and were then compared to the list of 1,011 NANDA (Nursing Diagnoses), NIC (Nursing Interventions Classification) and NOC (Nursing Outcomes Classification) terms which were gathered from the Hands-On Automated Nursing Data System (HANDS).³ For the overlapping terms (terms having identical CUI numbers), a comparison study of frequency of physician use and negation was performed.

Results

We discovered 19,737 physician concepts were negated at least once occurring 2,694,421 times. Out of the negated physician concepts, a total of 122 nursing terms were negated out of the 222 nursing terms physicians use. Of the 122 negated physician terms, 30 of the terms were also nursing concepts that were negated more than 10 times a year. The 33% of the most commonly negated nursing terms yielded an average frequency negation of 19.5%.

In 122 negated physician usage of nursing terms (55%), 62 NIC terms, 33 NOC terms, and 33 of the NANDA-I were included. The most frequently negated terms were Swallowing Problem, Bowel Incontinence, Neurological Status at 91.6%, 71.4% and 62.7% (respectively).

Discussion

Of the clinical concepts common to physicians and nurses, 55% of the terms were negated at least once. Of the entire list of physician terms, only 36.9% were ever negated. As reported before, very few nursing terms are used in physician language. The negation of Swallowing problems is likely due to review of systems, but is an important finding for both. The challenge of over half of nursing terms negated in physician utilization raises questions about meaning.² This highlights the little use of nursing data in physician language and how the utilization between the professions of the same concepts can differ, subsequently underscoring language differences between nurses and physicians.

Conclusion

The findings demonstrate even common terms between physician and nurses are used differently between the professions. When reading reports focused on concept it is easy to miss “not”, in standardized tests questions will capitalize NOT to ensure the correct meaning is portrayed. In collaboration, focusing on the context and how the concepts are written is just as important as the concept. A challenge is some nursing documentation practices does not allow negation of the concepts, creating a potential irreconcilable documentation difference between the professions.

References

Oncology Precision Network

Kerry Rowe, PhD, Lincoln Nadauld, MD
Intermountain Healthcare, Salt Lake City, Utah

Abstract
Cancer care is entering a new era of personalized treatment based on molecular profiles. Gathering evidence on molecularly-defined cohorts can inform treatment decisions for precision medicines. A multi-institutional genomics data consortium has been established to pool real-world data that provides large-scale, statistically powerful treatment insights. Clinical, molecular, treatment and outcome data is shared using methods that standardize and normalize data for comparisons across multiple health systems.

Description
Combining real-world evidence on molecularly-defined cohorts can inform treatment decisions for precision medicines. The use of molecular data to stratify populations leads to small sample sizes, limiting our ability to improve care. It is critical to pool real-world data across multiple institutions to draw large-scale, statistically powerful treatment insights.

A consortium led by founding members Stanford University, Intermountain Healthcare and Providence Health Systems has created a data-sharing network called the Oncology Precision Network (OPeN). OPeN combines data from electronic medical record systems, data warehouses, registries and other clinical systems at each institution. Using standard vocabularies such as RXNorm and LOINC, data is aggregated and linked with molecular profiles for patient. Using the services of the genomics software company, Syapse, a secure and trusted network is created where clinicians can review large sample sizes of patients grouped by cancer type and molecular profiles. Data is de-identified and stored in database hosted by Syapse. It is currently used for research but goal is to provide resource to assist in treatment decisions. A sample view of the data is shown here:

Work at Intermountain has focused on gathering data from all clinical sources where data exists for oncology patients. A number of queries and procedures have been created to extract data from our data warehouse and electronic medical record system. The clinical data is inserted into CDA documents that are securely transmitted to Syapse where it is combined with tumor sequencing data. This enables our data to be integrated with similar data from other institutions and removes the barriers associated with the use of different electronic medical record systems.
Development of an Automated Annual Cumulative Susceptibility Reporting Process

W. Cliff Rutter, PharmD, MS\(^1\), David S. Burgess, PharmD\(^1\)
University of Kentucky College of Pharmacy, Lexington, KY

**Background:**
The Clinical Laboratory Standards Institute (CLSI) recommends the use of cumulative susceptibility reports\(^1\), or antibiograms, to track the incidence of antimicrobial resistance. In addition to surveillance of resistance, clinicians consult the antibiogram to inform empiric antimicrobial therapy selection. Typically, antibiograms are developed manually by the clinical microbiology laboratory and require a significant time and personnel commitment. The purpose of this study was to develop a quick and reliable process to generate institutional antibiograms automatically.

**Methods:**
Antimicrobial susceptibility testing data from 1/1/2014 through 12/31/14 were obtained from the clinical microbiology laboratory for antibiogram creation. An algorithm was created to handle susceptibility data using open-source programs (RStudio v0.98 and R v3.1.2)\(^2,3\). Original raw data contain observations for multiple facilities, surveillance cultures, and multiple observations for individual isolates. Data were initially cleaned by selecting the locations of interest, eliminating surveillance cultures (i.e., axilla, groin, nares/perirectal swabs), and eliminating exact duplicate observations. Data were then transposed from narrow format into wide format with only one row for each observation. This “semi-clean” data still contains multiple observations per isolate, duplicate isolates, and many missing values generated by the transposition process. The final clean dataset is generated by combining multiple observations into one primary observation. This is accomplished by combining testing methods (Etest\(^\circledR\) and MIC), selecting the first isolate of the year, and removing duplicated isolates by selecting the more resistant biotype. CLSI breakpoints are applied to the current data and percent susceptible is calculated for each drug and organism to generate the antibiogram, which may undergo post-processing for optimal visual interpretation of antibiogram results. Between-method susceptibility results were compared with a chi-squared or Fisher’s exact test, as appropriate.

**Results:**
Antibiogram generation was significantly faster after development of the algorithm. Prior to algorithm development, the antibiogram was developed manually using Microsoft Excel\(^\circledR\) taking approximately 1 month of active work. The algorithm produces a completed antibiogram with CLSI interpretations of the MIC data in less than 60 minutes from raw data. The automated process produced similar results to the manually generated antibiogram with the same number of isolates. Results for non-beta-lactam antibiotics were nearly identical. The automated process produced higher percent susceptible for beta-lactams compared to the manually generated antibiogram, for example, *Klebsiella pneumoniae* was significantly less resistant to cefepime in the automatically generated antibiogram (98% vs. 92%, \(p=0.009\)). This suggests a practitioner bias towards beta-lactam resistant isolates instead of overall resistance profile.

**Conclusion:**
Antibiogram development can be significantly streamlined though the use of open-source data science technology. Automated antibiogram-generation functionality significantly reduces the time and resource commitment of previous antibiogram creation methods. This allows for more advanced analyses of the data available. Additionally, the automated process may be less susceptible to clinician bias in the isolate selection process.

**References:**
MyHealthKeeper: Clinical Trial of Personal Health Record-Based Healthcare Management

Borim Ryu1,2, Nari Kim1, Eun-Young Heo1, Sooyoung Yoo1, Se Young Jung1, Yoojung Kim2, Joongsik Lee2, and Jeong-Whun Kim1

1Seoul National University Bundang Hospital, Seongnam-si, Gyeonggi-do, KOREA; 2Seoul National University, Seoul, KOREA

Abstract

Personal health record (PHR)-based care management systems can be expected to improve patient engagement and data-driven medical diagnosis in clinical settings. In this study, we conducted a clinical trial in patients with lifestyle-related diseases involving the collection of data on physical activity and diet using wearable devices and the delivery of care interventions over 4 weeks. The results of our study showed how patients and clinicians apply personal health data and mobile applications. We have successfully demonstrated a practical approach to implementing and monitoring a mobile-based health intervention for self-management in clinical settings.

Introduction

Widespread interest in using PHRs has been reported and their uptake has been increasing. Interactive self-management technology is reported to be suitable and useful in a variety of chronic health conditions. The use of PHR applications in improving health care is promising, but integration into routine clinical care and population health management has proven challenging. The purpose of this study was to demonstrate the design of a clinical trial to assess the effectiveness of a PHR-based mobile health application given the name MyHealthKeeper, which can help patients with obesity to manage their individual daily routine and improve lifestyle behaviors.

Methods and Results

We utilized an activity tracker (Misfit) to collect individual physical activity data and developed a smartphone application to record every meal and daily logs. We also monitored data collection in a web-based administration environment. For the trial, 80 persons were randomly assigned to a PHR feedback care group (intervention group, n=51) or a usual-care group (control group, n=29). All participants underwent a paper-based survey, laboratory test, physical examination, and interview on their opinions. Clinical workflow observations were also processed to monitor and evaluate the study design. During the 4-week intervention period, we collected health-related mobile data, and the PHR feedback group visited the clinic twice and received PHR-based clinical diagnoses and recommendations using their EMR. The mobile application developed for self-reporting was successful in collecting personalized daily health data. Additionally, clinicians were able to promote physical activity to their patients. A total of 68 participants (44 in the intervention group and 24 controls) completed the study. In the results, the PHR intervention group showed significant weight loss compared to control group participants (average –1.4 kg, 95% CI 0.9–1.9, p<.001) in the final week (week 4). In addition, other factors such as BMI (average –0.4 kg/m^2, 95% CI 0.3–0.6, p<.001) and TG (average –46.7 mg/dL, 95% CI 17.6–75.8, p<.05) also changed during the study period.

Conclusion

We have developed an innovative testing approach to integrate PHR data into the clinical workflow in order to support lifestyle interventions. The results of this study suggest that PHR-based intervention is successful in preventing weight gain with modest weight loss and improvement in lifestyle behaviors.

Acknowledgement

This research was supported by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant no.: HI14C3213)

References

Innovative Method for Measuring Nursing Workload

Virginia K. Saba, EdD, RN, FACMI, FAAN\(^1\); Luann Whittenburg, PhD, RN, FNP, FHIMSS, PMP, FAAN\(^2\)

\(^1\)President and CEO, SabaCare, Inc. Arlington, VA; \(^2\)JP Systems, Alexandria, VA

**Keywords:** Clinical Informatics, Implementation, Care Plan, Clinical Care Classification, CCC, Nursing Workload, RVU

1. **Overview**
An innovative method for measuring nursing workload was developed by integrating a patient’s individualized plan of care (POC) in an electronic clinical system implemented with a standardized nursing terminology: the Clinical Care Classification (CCC) System, and CCC Information Model. The methodology uses POC nursing concepts with predetermined, valid Relative Value Units (RVUs) and CCC Base Value to determine workload.

2. **Methodology**
The CCC Terminology concepts provide the structure for the documentation of a coded patient POC following the six steps of the CCC Information Model. Once the POC has been implemented and tracked for a specific inpatient episode of illness, the four CCC Nurse Intervention Action Type Qualifiers in the POC’s for a given clinical/nursing condition are summarized and used to calculate nurse care with predetermined RVUs and which determine nursing workload (time). The CCC Information Model’s structure (six steps) is followed to track the patient’s POC; the four CCC Action Type Qualifiers (Monitor, Perform, Teach, or Manage). These are used to clarify and expand the focus of the CCC Nursing Interventions (core) and summarized; and their predetermined RVUs used to determine the nursing workload for the patient’s POC. The workload measurement is calculated using the time span of each Nursing Service Encounter Action and RVUs developed by RVS Inc.

3. **Example:** A CCC System Core Nursing Intervention Terminology Concept has been selected, such as CCC ‘Pulmonary Care’ to implement and track provider or nurse orders for inpatient care via the POC which is used to document the nursing care provided. The nurse during the patient’s hospitalization: a) ‘Monitors Pulmonary Care’ during evenings and nights, b) ‘Performs Pulmonary Care’ during the day shift, c) ‘Teaches Pulmonary Care’ to the patient and family, and d) ‘Manages Pulmonary Care’ by calling for new orders. In each nursing situation the Nursing Intervention Action Type Qualifier is documented and coded and on discharge aggregated.

This advanced analytic method was developed to integrate a standardized nursing terminology, the CCC System, into a patient’s clinical information system POC to provide reliable data on nursing workload based on predetermined RVUs, which reflect the value of nursing time (Table 1). This study extends a 2013 workload measurement study by coupling a Base Value of a disease condition with valid RVUs to determine the intensity of care required for the patient’s medical condition. The ability to accurately measure workload provides a care workload algorithm for use in determining future nursing workload predictions for the care of Pneumonia.

<table>
<thead>
<tr>
<th>Intervention Type</th>
<th>Frequency</th>
<th>Time</th>
<th>RVU Factor</th>
<th>Workload Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assess Pulmonary Care</td>
<td>5</td>
<td>50</td>
<td>Calculated</td>
<td>38%</td>
</tr>
<tr>
<td>Perform Pulmonary Care</td>
<td>20</td>
<td>200</td>
<td>Calculated</td>
<td>55%</td>
</tr>
<tr>
<td>Teach Pulmonary Care</td>
<td>5</td>
<td>50</td>
<td>Calculated</td>
<td>5%</td>
</tr>
<tr>
<td>Manage Pulmonary Care</td>
<td>3</td>
<td>30</td>
<td>Calculated</td>
<td>2%</td>
</tr>
<tr>
<td>TOTAL:</td>
<td>33</td>
<td>330</td>
<td></td>
<td>100%</td>
</tr>
</tbody>
</table>

4. **Evaluation Results**
This innovative method of contextual workload analytics supports the design of a clinical information system for a nursing POC benefiting patient care coordination. The coding of nursing data advances the calculations of nursing workload (time). This method can correlate the Nursing Intervention Actions with the patient’s Nursing Diagnoses, and when combined with Nursing Interventions provide a protocol for the ‘Pulmonary Care’ POC.

5. **Conclusions**
This method integrates the CCC System’s standardized nursing terminology and CCC Information Model to demonstrate that a POC does provide reliable nurse workload calculations based on pre-determined valid RVUs, which reflect the workload value of nursing.
Remote Usability Testing Method for an EHR Platform

Inkeri Saiku, MSc1, Yalini Senathirajah, PhD2, Elizabeth M. Borycki, RN, PhD3, Andre W. Kushniruk, PhD3, Upasana Malik, MSc4, Jinglu Wang, MSc1
1Weill Cornell Medicine, New York, NY; 2Hofstra University, Long Island, NY; 3School of Health Information Science, University of Victoria, Victoria, Canada; 4SUNY Downstate Medical Center, Brooklyn, NY

Introduction
Scheduling suitable testing times with busy users is always a hurdle with any kind of usability testing, but when conducting a usability test for a clinical system with the participants being medical residents, being able to offer flexible test times and locations becomes imperative in order to recruit a fulsome sample of study participants. Remote usability testing can be a solution. In addition to being flexible in terms of testing times, remote usability testing can also allow for recruiting of participants from a larger geographic area. For these reasons a remote usability testing method was constructed for the needs of usability testing of a novel EHR platform, MedWISER. In this poster we describe our approach.

Method
To gain user insights and to test the usability of a novel EHR platform, MedWISER, remote usability testing in conjunction with think-aloud was employed. The prototype being tested can be accessed online and therefore it is especially suitable for remote testing. For a think-aloud usability test to be conducted remotely, both users’ actions on the screen as well as the audio of their thinking aloud needs to be recorded. Eight software applications for screen and audio recording were tested: Go-to meeting, TeamViewer, Validately, Loop11, TryMyUI, Morae, Inspectlet and Join.me. Criteria for selection were: 1) Ease of use for the participants. For example, downloading and installing the needed software should be as easy possible for the participants. In addition, participants might not have admin rights on the computer they are using, therefore downloading the software should not require admin rights. 2) Quality of audio and video recording. 3) Availability for different operating systems. Of the tested software, TeamViewer was selected.

The TeamViewer software was used as follows for the usability tests: the participant is asked to download the TeamViewer software before the scheduled test session. At the time of the test session and before establishing the TeamViewer connection, the moderator calls the participant by phone and asks the participant to remove all identifying or sensitive information anywhere on the participant’s screen e.g. participants name or patient data. Before connecting the moderator confirms in the call with the participant that there is no identifying or sensitive information anywhere on the participant’s screen e.g. participants name or patient data. To achieve this, the participant is asked to open a new (incognito) browser window and close all messaging programs, such as iMessage or What’sUp, that might bring up private messages on the screen during the recording. After the TeamViewer connection is established, the phone call is ended and verbal communication between participant and moderator continues via TeamViewer so that the audio is recorded along with the screen recording. Through the TeamViewer text chat option, the participant first receives links to demo videos of the think-aloud method and of the features of the EHR prototype followed by a link to the prototype. Within the MedWISE interface the participant is given tasks to solve clinical cases while thinking aloud. In a one hour session, the participant goes through 2-3 patient cases. After reviewing the cases the participant is asked to rate the ease of use and usefulness of the system, as well as comparing it to the systems they are currently using.

Results and Conclusion
The described method was considered successful for conducting remote think-aloud usability testing. The quality of screen capture and audio were suitable for recording and analyzing the use of the system being tested. Options for participating in the test either in the project premises or at participant’s own offices were also offered, but most participants (70%) chose the online method. Reasons participants mentioned for preferring the remote test were more flexible test hours and the convenience of participating from home. Ten test sessions have already been carried out and we will continue conducting more usability tests with this method.

Acknowledgements: This work was funded by AHRQ #5R01HS023708-02
Anticoagulation Safety Risk Assessment: Evaluating compliance with an anticoagulation protocol through secondary use of EHR data

Hojjat Salmasian, MD, MPH, PhD, Eric Venker, PharmD, MD, Emilia Hermann, MD, MPH, Iheanacho Obinnaya Emeruwa, MD, MBA, Daniel Farrell, MD, Jason Adelman, MD, MS

1The Value Institute, NewYork-Presbyterian Hospital, New York, NY; 2Columbia University, New York, NY

Introduction

The American Medical Informatics Association (AMIA) and the Office of the National Coordinator (ONC) are proponents of “secondary use of EHR data”, where electronic healthcare data are aggregated for purposes other than direct patient care. One important secondary use of EHR data is to support quality improvement and risk assessment initiatives. We aimed to use EHR data to evaluate compliance with the hospital protocol for intravenous heparin, an anticoagulant used to treat serious cardiovascular diseases such as atrial fibrillation, deep vein thrombosis and acute coronary syndrome. Intravenous heparin is considered a high-alert medication, as both under-treatment and over-treatment can lead to serious adverse outcomes; its dosing is adjusted based on the aPTT lab levels.

Methods

In this cross-sectional study, we analyzed all heparin orders placed in a one year period at two major hospitals in New York City. We excluded orders for ‘prophylactic’ use of heparin, or those for patients younger than 18 years of age, as the indications and dosing recommendations differ in these populations. In addition, we restricted the data to orders placed to treat deep venous thrombosis (DVT), pulmonary embolism (PE), Atrial Fibrillation (AFib), or Acute Coronary Syndrome (ACS). Our guidelines are based on best available evidence, and recommend a target value of 76 to 112 for aPTT in patients with these indications. Our guidelines recommend initiating heparin with a bolus dose of 2,000 to 10,000 units, followed by an infusion drip at the rate of 400 to 2,500 units/hr (based on weight and indication).

Results

Of 28,838 heparin orders included in the analysis, 14,224 were placed to treat DVT/PE, 10,012 to treat AFib, and 4,602 to treat ACS. Together, these orders were placed for 2,275 unique patients and in 2,450 unique admissions. An initial bolus dose was used only in 7.5%, 16.8% and 22.6% of patients with AFib, DVT/PE, and ACS, respectively. The first heparin drip order was also mostly at a lower than recommended dose (mean log expected-to-observed dose = -0.29, SD = 0.70). Overall, 55% of patients achieved target aPTT within 24 hours of heparin initiation. Target aPTT was more likely to be achieved whenever heparin therapy started with a bolus dose (p=0.039) and when the heparin drip was ordered at the recommended dose (OR = 1.18, p=0.013).

Conclusion

Secondary use of EHR data can play a significant role in designing quality improvement (QI) initiatives that are informed by the common practice behavior and target the most salient issues. Our analysis allowed us to identify two potential causes for suboptimal anticoagulation therapy at this institution: providers underused bolus heparin orders, and the initial infusion rate was often a lower rate than what is recommended for the patient’s weight and indication. These issues can be mitigated through the use of clinical decision support, whereby the guidelines would be incorporated into the order set and appropriate drip dose would be recommended at the point of care.

References

Detecting Adverse Drug Event Safety Signals from MEDLINE Reports: Challenges in Employing Cross-terminology Mapping of MeSH to MedDRA

Abhivyakti Sawarkar, MD, MMSc; Alfred Sorbello, DO, MPH; Anna Ripple, MLS; Olivier Bodenreider MD, PhD

1US Food and Drug Administration, Center for Drug Evaluation and Research, Silver Spring, MD, USA; 2Lister Hill National Center for Biomedical Communications, U.S. National Library of Medicine, Bethesda, MD, USA

Background: The Center for Drug Evaluation and Research/Office of Translational Sciences Data Mining Team at the US Food and Drug Administration (FDA) has recently developed a web-based prototype analytical software tool to automate detection of adverse drug event (ADE) safety signals from MEDLINE reports through quantitative data mining of Medical Subject Heading (MeSH) indexing terms. The prototype tool greatly enhances the FDA’s capacity to integrate outputs from MEDLINE with other data mining streams, such as the FDA Adverse Event Reporting System (FAERS), to support multimodal pharmacovigilance. The tool incorporates a system in which MeSH terms are leveraged to extract ADE associations from MEDLINE citations, and the ADE content of a MEDLINE report is conveyed through a single MeSH descriptor. A key element to facilitating interoperability of the data mining outputs from MEDLINE and FAERS is to standardize the representation of ADE content using the Medical Dictionary for Regulatory Activities (MedDRA). In this study, we assess the feasibility of rendering MeSH descriptors for ADEs to MedDRA leveraging the Unified Medical Language System (UMLS) to map MeSH descriptors to equivalent or finer-grained MedDRA preferred terms (PT).

Methods: We manually reviewed the cross-terminology mapping by assessing the completeness and quality of the one-to-one and one-to-many MeSH descriptor-to-MedDRA PT relationships in the context of an arbitrarily selected use case, ciprofloxacin. More specifically, we assessed how well the MedDRA PTs represented the clinical ADE content expressed by the MeSH descriptors in the source MEDLINE citations.

Results: We identified 219 unique drug-AE associations for ciprofloxacin in which the ADE content of the MEDLINE citations was conveyed by unique MeSH descriptors. There were 84 one-to-one MeSH-to-MedDRA PT mappings and 135 one-to-many MeSH-to-MedDRA PT mappings. Manual assessment of the one-to-one mappings confirmed that the linked terms conveyed similar meaning in translating the clinical ADE content of the source citations between the two terminologies. However, the one-to-many mappings included some MeSH descriptors mapped to a multitude of MedDRA PTs (e.g., ‘musculoskeletal diseases’ mapped to 276 unique MedDRA PTs) that required manual review to select the PT that most reasonably translated the clinical content between the terminologies. Overall, the major categories of cross-terminology challenges involved MEDLINE citations annotated with (1) general MeSH headings (e.g., ‘heart diseases’) that appeared broad and less distinct compared to the granular mapped MedDRA PTs and that may conflate multiple related major topics, and (2) pre-coordinated descriptors (e.g., ‘Drug Hypersensitivity Syndrome’) that are conceptually broad in scope and subsume distinct clinically important conditions (e.g., drug reaction with eosinophilia and systemic symptoms (DRESS syndrome)).

Conclusions: This study illustrates some of the challenges in cross-terminology mapping from MeSH to MedDRA and illustrates the limitations inherent in leveraging single MeSH descriptors to convey highly specific ADE content. While one-to-one MeSH descriptors mapped to single MedDRA PTs provide the best links to translate clinical ADE content between the terminologies, the accuracy of translating clinical content between MeSH and MedDRA in one-to-many mappings may be severely impacted by the granularity mismatch between MeSH and MedDRA. Future work will explore text mining of abstracts and full text citations with machine learning to resolve the translation of ADE content between MeSH and MedDRA and thereby support automation.

References

Acknowledgements: This project was supported in part by appointment to the research participation program at CDER administered by Oak Ridge Institute for Science and Education (ORISE) for the FDA. Funding support also received from FDA/CDER/Office of Translational Sciences and the Intramural Research Program, NIH, National Library of Medicine. Disclaimer: The views expressed are those of the authors and do not necessarily represent the views of the US FDA, the NIH, or the US Government.
Development of a Conceptual Model for Clinical Deployment of e-Patient-Reported Outcomes

Courtney Segal, BA¹, Cynthia LeRouge, PhD¹, Elizabeth Kellogg, MPH¹, Elizabeth Austin, MPH¹, Danielle Lavallee, PharmD, PhD,¹
¹University of Washington School, Seattle, WA, USA

Introduction

Patient-reported outcomes (PRO) provide insight on patients’ experience and perspectives regarding treatment and outcomes. The growing availability of electronic PRO (ePRO) in clinical settings has the potential to improve the efficiency of data collection and analysis and to provide new opportunities to bring meaningful evidence to decision makers and patients in innovative ways. Despite the increasing pressure for using PRO measures (due to shifts towards patient-centered healthcare and payment reform initiatives), comprehensive models of clinical use do not exist. Literature reviews of PRO systems do not focus on the development and clinical use of e-PROs. The purpose of this research is to develop a conceptual framework that reflects a socio-technical systems perspective of ePROs in clinical use. The TURF (Task, User, Representation, and Function) EHR Usability Framework provides a system-level view that reflects both the social and technical aspects of an EHR system.¹ We use the TURF framework to systematically explore the constructs of clinical use of PRO systems.

Methods

The search parameters in EBSCO Host and PubMed were ((patient reported outcomes)) AND (implementation OR workflow OR usability OR adoption) NOT ((clinical trial)) among articles in English and academic journals published over the last five years, resulting in 268 articles. Investigators reviewed the 268 abstracts for inclusion and exclusion criteria (n=121). An additional 23 articles were excluded after reviewing the full text, resulting in 98 articles that were included for the detailed review. The TURF model served as a high-level coding schema that guided open coding within each TURF construct using Dedoose software. Resulting detail codes and supporting article text was reviewed by a team of appropriate experts (e.g., visualization and human-centered design experts for codes under Representations) at regular junctures to further refine the coding schema, validate propriety of the codes assigned to text, and, in the final stages, to identify cross-cutting themes.

Results

The literature review identified several emerging themes and common challenges for deploying ePROs in the clinical setting (summarized into design principles). FUNCTION - a PRO system can function to advance efficient care, care coordination, decision support, patient engagement, quality assurance and population health, . Patient-facing functions should support self-management of diseases and symptom tracking, and interactive functions (e.g., annotation of symptom charts). TASK - the length of the PRO measures selected should be optimized to reduce patient burden and improve data completeness; automated tasks support clinician workflow and improve data access and analysis; PRO/EMR integration is conducive to care coordination for patients and clinician/ staff communication and. REPRESENTATION – ePRO design should consider the use of color, reporting of MCID scores, raw scores versus incremental changes, and personalization of content.. USERS – PRO system design should consider non-English speaking, medically frail, and aged patients. Active motivation efforts (e.g., champion, case review, system-wide adoption, identified clinical value, peer support) are needed to promote clinician adoption.

Conclusion

We provide the first literature review of ePROs in clinical use and a resulting contextualized framework that reveals current approaches to ePRO design and practice and challenges to successful use. This framework can be used by developers, practices, and researchers to design and deploy PRO systems in clinical settings. We also demonstrate that the TURF EHR Usability Framework can be applied to the PRO context. The information gleaned from this effort will be synthesized with a larger ongoing project involving direct data collection resulting in ePRO design guidelines that reflect a systems, visualization, and workflow perspective.

References

Evaluating the Appropriateness of Medication Related Clinical Decision Support Alert Overrides in the Inpatient and Outpatient Settings

Diane L Seger RPh1,2, Adrian Wong, PharmD1,5, Mary G Amato PharmD MPH1,5, Sarah P. Slight MPPhD PGDip1,3, Patrick E Beeler MD1, Olivia Dalleur MPharm PhD1,4, Tewodros Eguale MD PhD1,5, Christine A. Rehr BA1,2, Julie M Fiskio BS1,2, Karen C Nanji MD MPH6,7, David W. Bates MD MSc1,2,7

1 The Center for Patient Safety Research and Practice, Division of General Internal Medicine, Brigham and Women’s Hospital, Boston, MA, USA; 2Partners Healthcare, Somerville, MA, USA; 3Newcastle University, Newcastle-upon-Tyne, UK; 4Louvain Drug Research Institute, Universite catholique de Louvain (UCL), Belgium; 5MCPHS University, Boston, MA; 6Department of Anesthesia, Critical Care and Pain Medicine, Massachusetts General Hospital, Boston, MA, USA; 7Harvard Medical School, Boston, MA, USA

Abstract: Clinical decision support (CDS) systems can provide clinicians with knowledge and patient-specific information to help guide their decision making. Balancing CDS with the risk for overexposure (e.g., alert fatigue) is needed. The objective of this study was to characterize and compare CDS override rates and appropriateness of these overrides in the inpatient and outpatient setting within our healthcare system.

Introduction: Medication-related clinical decision support (CDS) systems are a valuable tool for improving safety and quality of care by preventing medication errors. However, despite extensively modifying the CDS in our legacy electronic health record to improve user acceptance, we continue to observe a high level of overrides for many medication alert domains. In this study, we evaluated the overrides for drug-allergy interactions (DAI), drug-drug interactions (DDIs), duplicate therapy, geriatric, and renal medication substitution alerts in the inpatient and outpatient settings to determine if the overrides were appropriate or not.

Methods: Our primary outcome measure was whether a CDS medication alert was overridden appropriately. Data on these overrides were obtained over a period of four years (Jan 2009 to Dec 2012) for inpatients at Brigham and Women’s Hospital and outpatients within Partners HealthCare. The alert override reasons were collected and a sampling of each alert type override was evaluated for appropriateness by a trained multidisciplinary group.

Results: The most common reason given for overriding the DAI alerts in both settings was “patient had taken the drug previously without allergic reaction.” The most common reason for DDI overrides in both settings was “will monitor as recommended.” A chi-square test was used to compare inpatient vs. outpatient override appropriateness for each alert category. Details of the override rate and the appropriateness rate of the overrides for the inpatient and outpatient settings are detailed in Table 1.

<table>
<thead>
<tr>
<th>Table 1. Override rate and appropriateness of overrides for inpatient and outpatient settings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type of Alert</td>
</tr>
<tr>
<td>----------------</td>
</tr>
<tr>
<td>Drug-allergy</td>
</tr>
<tr>
<td>Drug-drug</td>
</tr>
<tr>
<td>Duplicate</td>
</tr>
<tr>
<td>Renal</td>
</tr>
<tr>
<td>Geriatric</td>
</tr>
</tbody>
</table>

Appro = appropriate; Inpt = inpatient setting; Outpt = outpatient setting; OR = override; UTO = unable to obtain

Conclusion: The rate of medication-related CDS overrides continues to be high in both settings. DAI and duplicate therapy alerts were often appropriately overridden and should be refined to reduce the number of displayed alerts, in turn minimizing the risk for alert fatigue. Future research should focus on 1) the improvement of the clinical relevance of alerts to increase acceptance of DDI, geriatric, and renal alerts and thereby reduce inappropriate overrides and 2) seeing if adverse drug events are caused by these overrides in both settings.


This study was funded by grant #U19HS021094 from the Agency for Healthcare Research and Quality (AHRQ).
DeepSuggest: Query Expansion for Clinical Notes by Deep Learning

Emre Sezgin, PhD, Koushik Jinna, MS, Tran Bourgeois, MPH, Soheil Moosavinasab, MS, Yungui Huang, PhD, Simon Lin, MD, MBA
The Research Institute at Nationwide Children’s Hospital, Columbus, OH

Introduction

Today, more than 70% of the health information are stored in unstructured clinical notes, and healthcare providers (HCP) increasingly rely on the text search systems. However, these clinical notes may have spelling variations, typos, personal acronyms, and informal words. Although EPIC’s new search feature aims to help HCPs finding information quickly, it has limited capability in recognizing acronyms or suggesting spelling variations. To fill this gap, we propose an alternative search system (DeepSuggest) expanding queries by suggesting spelling variations, acronyms and other semantically relevant words through an unsupervised deep learning algorithm.

Methods and Data Retrieval

Our training corpus consists of over 50 million clinical notes (~69GB) documenting encounters at Nationwide Children’s Hospital between 2006 and 2016. After removing words with less than 30 occurrences, we created a vocabulary having 6.3 million unique 1-3 grams, representing 5.5 billion total words.

Word2Vec is an unsupervised learning approach that uses neural networks and word embedding to map the words into a vector space. In this process, the Continuous Bag-of-Words (CBOW) algorithm of Word2Vec (2 iterations and 400 dimensions) was used to create a statistical model of 1-3 gram word similarity across the corpus using word proximity in the notes (2 days 9 hrs of runtime). The nearest neighbor model was utilized to find the closest and semantically relevant N-Grams from the trained model. The model created by Word2Vec is indexed by Annoy (a library for nearest neighbors) and retrieved quickly and efficiently (<1sec). The model can be trained incrementally with more notes over time and run in a distributed environment to accommodate larger corpora.

Results and Evaluation

As an example result, DeepSuggest was able to suggest “tonsillectomy” (correction of “tonsilectomy”), “t&a” and “t/a” (alternative synonym for “tonsillectomy and adenoidectomy” procedure) or “tonsillectomy” or “adenoidectomy” (narrower concepts). In another test, a query for “worried about” (e.g. to identify patient concerns) led suggesting semantically related words, “afraid” and “fear”, and morphologically similar words, “worry about” and “worrying about”.

We will use Preferred Names in SNOMED as queries (count in corpus > 50), and use their synonyms to evaluate recall rate of DeepSuggest. We will also invite domain experts to evaluate suggestions that do not exist in SNOMED to calculate precision and NDCG. To understand the effectiveness and usefulness of DeepSuggest, a usability test was designed. The pilot study helped to build the study protocol and questions. 10 residents at NCH will be invited to join usability study and to test DeepSuggest in comparison to Epic search. We hypothesized that DeepSuggest is more effective and useful in assisting for clinical decision making than Epic search in clinical queries.

Conclusion

In the proof of concept stage, DeepSuggest proved its functionality after its implementation using our corpus, suggesting spelling variations, acronyms and semantically relevant words. The system is promising for a higher recall rate in EHR searches, and thus, increasing productivity in health care delivery.

References

Mapping CT Codes from a Health Information Exchange to LOINC

Jason S. Shapiro, MD¹, Anton O. Beitia, MD¹, Tina Lowry MS¹, Daniel J. Vreeman, PT, DPT, MSc², Bradley N. Delman, MD¹, George Loo, DrPH¹, Frederick Thum III, MD¹
¹Icahn School of Medicine at Mount Sinai, New York, NY; ²Regenstrief Institute, Indiana University School of Medicine, Indianapolis, IN

Introduction
As a prerequisite to building an alerting system to notify providers at the point of order entry of prior CT exams performed anywhere across a health information exchange (HIE), we are mapping proprietary CT names and codes from the Healthix HIE in the New York metropolitan area to LOINC®. We present preliminary results of LOINC’s coverage of the most frequently performed diagnostic CT exams at 26 sites within the Healthix HIE.

Methods
Healthix currently has 41 sites that transmit radiology data. In order to expedite mapping for a pilot prior CT alerting system, and using a data set of CTs performed between January 2013 and October 2015, we elected to first map up to the 200 most frequently performed diagnostic CT exams from each site to LOINC. If a site had fewer than 200 diagnostic codes, we mapped all diagnostic codes from that site. The rationale was that these CTs (as opposed to procedural/interventional or less frequently used diagnostic CTs) were the most likely to be repeated. At the time of this study, we had completed mapping 26 of the 41 sites. Sites were prioritized for mapping based on convenience and data availability. Mapping was performed manually by one of the authors (AOB), a radiologist and informatician with domain expertise, with the assistance of LOINC’s desktop mapping program RELMA®, to match local exam descriptions with LOINC names. We calculated the total number of CT codes for all sites and the absolute number and percentage of codes (concept-type coverage) for which we had an initial LOINC match, and the number and percent of codes for which there was no initial LOINC match. We also counted the number of codes with ambiguous descriptions which required additional investigation in order to disambiguate (e.g., a local site’s “CT Hematuria Protocol” required report review to ascertain that it was a “CT abdomen and pelvis +/- IV contrast”), because these codes are particularly time consuming to map. For exams without an initial LOINC match, we submitted requests to LOINC for the creation of new codes, assigned these exams temporary codes, and then tracked new LOINC codes that were created based on our requests. Because the requests were submitted to LOINC in batches over time, some are still pending a final decision. We recalculate concept type coverage for the 200 most frequently performed exams with the updated LOINC CT list.

Results

Table 1. Breakdown of CT codes at 26 Healthix sites

<table>
<thead>
<tr>
<th>Total CT codes</th>
<th>Total diagnostic CT codes</th>
<th>Ambiguous CT codes (see text)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5715</td>
<td>4306</td>
<td>268</td>
</tr>
</tbody>
</table>

Table 2. Concept coverage of LOINC for up to the 200 most frequent Diagnostic CT codes at 26 Healthix sites

<table>
<thead>
<tr>
<th>CT codes with initial LOINC match</th>
<th>CT codes w/o initial LOINC match</th>
<th>Initial Concept Type Coverage (see above)</th>
<th>CT codes with new LOINC created</th>
<th>CT codes with match after new LOINC creation</th>
<th>Concept type coverage with new LOINCs</th>
</tr>
</thead>
<tbody>
<tr>
<td>4006</td>
<td>300</td>
<td>93.03%</td>
<td>93</td>
<td>4099</td>
<td>95.19%</td>
</tr>
</tbody>
</table>

Discussion/Conclusion
LOINC provided excellent coverage for the most frequently performed CT exams with concept coverage initially at 93% and improving to 95% with the addition of new codes. We are continually submitting requests for new codes to LOINC and expect concept type coverage to continually improve. Coverage will never be 100% because there are certain exam code types for which LOINC does not plan to create new codes, including nonspecific exam codes, such as “CT orbits, sella, ear, and/or posterior fossa”, in which the exam can cover any one or any combination of the areas specified. The benefit for LOINC to maintain specificity of anatomic location in the exam codes outweighs the loss of not having a code for such non-specific exams. As secondary use of HIE data becomes more commonplace, mapping to standard terminologies such as LOINC will be important. This study may help others gain a sense of effort involved and problems encountered in similar large mapping efforts for radiology codes.
Robust Lung Nodule Classification using 2.5D Convolutional Neural Network

Shiwen Shen¹,², Alex A.T. Bui² PhD, William Hsu² PhD
¹Department of Bioengineering, ²Medical Imaging Informatics (MII) Group, Department of Radiological Sciences; University of California, Los Angeles, CA, USA

Introduction

Discovering lung nodules in low-dose computed tomography (CT) is a difficult task. Moreover, it is challenging for conventional model-based lung nodule classifiers trained on one dataset to achieve comparable performance on a completely independent dataset. In this work, we develop a robust novel lung nodule classification model using a 2.5D deep convolutional neural network (CNN). The contributions of this work are: 1) a deep CNN architecture trained on axial, sagittal, and coronal views; and 2) improved model robustness with stacked small convolution (conv) filters (3×3) design enabling more layers, combined with dropout techniques and online training data augmentation scheme. We also present external evaluation results on a complete independent dataset.

Methods

The proposed deep CNN was trained and validated using 893 CT scans from the Lung Image Database Consortium (LIDC). Nodule annotations from four radiologists were used as positive subjects (lung nodules). The non-nodule objects were generated using adaptive multi-level thresholding and morphological operations. Figure 1 illustrates the overall 11 layer CNN architecture. For each candidate, a 3D cube of 40×40×40 mm is extracted, and then resized to 64×64×64 pixels (resolution of 0.625 mm/pixel). The center slices of this cube in the axial, sagittal, and coronal views are then aggregated as a 2.5D input (64×64×3). We used a conv module design comprising two 3×3 conv layers and one 2×2 max pooling layer (with stride of 2). Stacking two conv layers achieved an effective receptive field of 5×5, significantly reducing the number of model parameters while increasing the discriminative power. Three conv modules were used in total, consisting of 32, 64, and 128 kernels for each of their two conv layers, respectively. The output of the third conv module was linked to a fully-connected (FC) layer with 128 rectified linear units (ReLUs). Dropout and online data augmentation were implemented on to reduce training overfitting. We then performed an external validation using data from UCLA, consisting of 158 CT scans with annotated lung nodules.

Results

The scans in the LIDC dataset were randomly divided into five subsets of similar size. Three subsets were used for training, one for validation, and one for testing. This dataset contained a total of 6,776 nodules. The test subset contained 207 scans, with 1,262 nodules and 8,281 non-nodules. For the test set, our method obtained better results in all metrics compared to [1]: an area under the curve (AUC) of 0.991, sensitivity (SEN) of 95.6%, specificity (SPE) of 97.0%, and accuracy (ACC) of 96.8% (vs. AUC of 0.922, SEN of 91.9%, SPE of 94.7%, and ACC of 94.3%). The UCLA dataset, which included 158 nodules and 3,938 non-nodules, were used as an independent dataset to validate the model performance. The model achieved an AUC of 0.948, SEN of 88.6%, SPE of 88.7%, and ACC of 88.7%.

Conclusion

We present a novel robust lung nodule classification approach using a 2.5D deep convolutional neural network. We achieved significantly higher performance and robustness compared to previous work.

Figure 1. Illustration of 2.5D deep convolutional neural network architecture.

Reference

Accelerating Rare Disease Diagnosis with Collaborative Filtering

Feichen Shen, Ph.D., Sijia Liu, M.S., Yanshan Wang, Ph.D., Liwei Wang, Ph.D., Naveed, Afzal, Ph.D., Hongfang Liu, Ph.D.
Department of Health Sciences Research
Mayo Clinic, Rochester, MN, USA

Abstract

Patients with rare diseases are frequently misdiagnosed or undiagnosed due to the lack of knowledge and experience of care providers. We hypothesize that patients’ phenotypic information in electronic medical records can be leveraged to accelerate disease diagnosis. In this study, we proposed a collaborative filtering system enriched with natural language processing and semantic techniques to assist rare disease diagnosis based on phenotypic characterization. Results showed that our approach could stratify patients with similar rare diseases.

Introduction

The problem of diagnosing a patient with a rare disease based on patients’ phenotypic information is similar to recommending a product to a customer, and therefore it is natural to propose the use of collaborating filtering for rare disease diagnosis. In this study, we proposed a patient based collaborative filtering system, enriched with natural language processing (NLP) and semantic techniques, to assist rare disease diagnosis based on phenotypes extracted from free-text clinical notes.

Method

The system consists of two modules: i) a preprocessing module leveraging natural language processing and semantic processing techniques to identify patients with rare diseases and collect their phenotypes by building an annotation pipeline with Human Phenotype Ontology (Robinson PN et al., 2008) and MetaMap (Aronson AR, 2001); and ii) a user-based collaborative filtering model to recommend similar patients and possible disease recommendations. Four similarity measurements (Tanimoto (TANI: Jaccard, 1901), Overlap (OL: Szymkiewicz, 1934), Fager & McGowan (FMG: Fager et al., 1963), and Log likelihood ratio (LLRS: Dunning, 1993)) and two neighborhood algorithms (K Nearest Neighbor (KNN: Altman, 1992) and Threshold Patient Neighbor (TPN: Bellogin et al., 2014)) are applied.

Results

We used a subset of Mayo Clinic 2010-2015 clinical notes, including 24,000 patients with 437 rare diseases and 2,400 phenotypes. We applied a 10-fold cross validation to determine the optimal number of neighbor k and threshold t, and used three-level matching criteria to measure similarity between any two diseases: String matching, SNOMED-CT semantic matching (similar if shared the same ancestor within 3 generations), and GARD (Genetic and rare diseases information center) category matching (similar if in the same rare disease category). For patient recommendation, Table 1 gives the mean average precision (MAP) and precision-recall area under curve (PRAUC) for each experiment applied on all patients. For rare disease prediction, we applied TANI with KNN as an optimal algorithm. String, SNOMED-CT and GARD matching found 49.4%, 56.6%, and 92.1% correct predicted rare diseases, and weighted micro-average F measure for them are 0.4, 0.44, and 0.71.

Table 1. MAP/PRAUC for patient recommendation with 8 experiments and 3 matching criteria (highest value in bold).

<table>
<thead>
<tr>
<th></th>
<th>TANI+KNN</th>
<th>TANI+TPN</th>
<th>LLRS+KNN</th>
<th>LLRS+TPN</th>
<th>OL+KNN</th>
<th>OL+TPN</th>
<th>FMG+KNN</th>
<th>FMG+TPN</th>
</tr>
</thead>
<tbody>
<tr>
<td>String</td>
<td>0.4229/0.4119</td>
<td>0.4228/0.4113</td>
<td>0.3736/0.3750</td>
<td>0.3672/0.3558</td>
<td>0.2671/0.2665</td>
<td>0.2672/0.2667</td>
<td>0.2124/0.2187</td>
<td>0.3007/0.2968</td>
</tr>
<tr>
<td>SNOMED</td>
<td>0.4727/0.4614</td>
<td>0.4729/0.4612</td>
<td>0.4297/0.4292</td>
<td>0.4215/0.4106</td>
<td>0.2989/0.3005</td>
<td>0.2988/0.3005</td>
<td>0.2387/0.2487</td>
<td>0.3303/0.3296</td>
</tr>
<tr>
<td>GARD</td>
<td>0.7525/0.7457</td>
<td>0.7530/0.7464</td>
<td>0.7116/0.7147</td>
<td>0.7084/0.7023</td>
<td>0.6114/0.6271</td>
<td>0.6094/0.6262</td>
<td>0.5964/0.6129</td>
<td>0.68410.6856</td>
</tr>
</tbody>
</table>

Conclusion

We have investigated patient based collaborative filtering with NLP and semantic techniques on cohort to assist rare disease diagnosis. We demonstrated its potential in facilitating rare disease prediction.

Acknowledgements

The authors gratefully acknowledge the support from the National Institute of Health (NIH) grants R01GM102282, R01LM011934, R01EB19403, and OT3TR002019.

2181
An approach for preventive healthcare for elderly people in Japan using an information and communication technology-based system

Akiko Shibuya, RN, PHN, PhD\textsuperscript{1}, Tsuneo Yamada, MSc\textsuperscript{2}, Jimpei Misawa, PhD\textsuperscript{1}, Yukihiro Maeda\textsuperscript{1},
Rie Ichikawa, MD\textsuperscript{1,3}, Teruyoshi Hishiki, MD, PhD\textsuperscript{4}, Yoshiaki Kondo, MD, PhD\textsuperscript{1}
\textsuperscript{1}Department of Health Care Services Management, \textsuperscript{3}Department of Pediatrics and Child Health, Nihon University School of Medicine, Tokyo, Japan; \textsuperscript{2}The Medical Information System Development Center, Tokyo, Japan; \textsuperscript{4}Department of Information Sciences, Toho University, Chiba, Japan

Abstract
We developed a preventive health care system for elderly people using information and communication technologies (ICT). Users can record their health information (e.g., weight, physical condition, and mental health status) in the system via a password-protected portal at Amami city of Island, Japan, and can receive individually tailored advice, healthcare assessments, and messages from healthcare professionals via their computers and mobile phones. This system may have a significant impact on preventive healthcare for elderly people.

Introduction
With the dramatic aging of the population, a need exists to support preventive healthcare services for the elderly\textsuperscript{1}. Information and communication technologies (ICT), which include services such as communication, consultations, and monitoring, have a significant potential to be useful in enhancing quality of life and providing preventive healthcare for elderly people who live far away from or have limited access to healthcare systems\textsuperscript{2}. However, concerns still exist regarding barriers to ICT use by elderly people\textsuperscript{3}. In a previous study, lack of convenient access to the Internet and use of modern technologies, such as a personal computer, as well as difficulty in obtaining technical support, were considered important barriers to ICT for elderly people\textsuperscript{3}. We used ICT to develop a preventive healthcare system for elderly people.

Methods and results
Figure 1 shows the architecture of our ICT-based preventive healthcare system, which is accessible via a password-protected portal. With this system, elderly users can record their health information (such as body weight, exercise, physical condition, and mental health status) via a password-protected portal and can view their data in graphical forms or charts on their computers and mobile phones (Figure 2). Furthermore, both elderly users and healthcare professionals can access the information to obtain health data. Healthcare professionals can provide individually tailored advice and healthcare assessments, as well as send messages to patients regarding lifestyle factors such as physical exercise, mental health recommendations, and other important medical information regarding preventive healthcare. Family members who live far away (i.e., outside of the island) can be given access to sign-in names and passwords to monitor elderly people's information and communicate about and support the recommended preventive healthcare measures.

To verify the efficacy of the system, 18 elderly people (aged $\geq$65) in Amami City of Island, located 1,300 km southwest of Tokyo, participated in the trial. Participants first received lectures regarding the health site, after which they recorded their health status information and viewed the results in graphical form or charts on their computers. We then assessed their satisfaction and system usability using a questionnaire survey. Overall, 76% of subjects rated the usefulness of the system as excellent. And, 86% of subjects rated the operability of the system as excellent or good. This finding indicates that a preventive healthcare system using ICT was positively received, and it is believed that the system may be useful for providing preventive healthcare to elderly people.

Conclusion
The present study suggests that a preventive healthcare system using ICT may help meet various support needs of elderly people in terms of recommending preventive healthcare measures and enhancing quality of life.

References
Anesthesia Instrument Recognition System

Scott Siegel, Masters, Agyeiwaa Agyei, Anis Davoudi, Bachelors, Patrick Tighe, MD, Parisa Rashidi, PhD
University of Florida, Gainesville, Florida, USA

Abstract

Our goal is to develop a novel intraoperative anesthesia instrument detection and characterization system using deep region-based convolutional neural networks. This system will be designed to operate in real-time during surgical procedures and will provide surgical teams with the assurance that all anesthesia equipment required for a specific procedure is present without having to check manually.

Introduction

Surgical and anesthesia safety checklists are commonly used during surgical procedures and have reduced the rate of complications in the operating room (OR)\(^1\). Despite their success, there are still concerns about their time consumption and efficiency\(^2\). Preoperatively checking for the presence of anesthesia equipment is one example that can take a nurse anesthetist (NA) almost two minutes to complete if no items are missing\(^3\). An autonomous anesthesia instrument recognition system could perform a preoperative check for anesthesia equipment in less than a second. Our hypothesis is that a using a deep region-based convolutional neural network (R-CNN) for this task would reduce the time required to complete surgical safety checklists and improve efficiency in the OR.

We are utilizing the Faster R-CNN network to detect the presence of anesthesia equipment. This network utilizes Tensorflow as a backend and was trained end-to-end via backpropagation and stochastic gradient descent\(^4\). Images were extracted from videos of various anesthesia equipment were recorded as they were rotated and placed in different positions. A preliminary test was run on a training set that consisted of two objects (oxygen mask and syringe). The pretrained ImageNet VGG-16 network was used to initialize the fully convolutional portion of the network. A Mean Average Precision (MAP) of 90.6% for both objects was achieved. We are planning on testing other models in place of the VGG-16 network in the future and increasing the number of instruments to be detected.

![Oxygen Mask](image)

**Figure 1.** Example output of Faster R-CNN network.

Conclusion

This project utilizes recent advances in machine learning and object detection to eliminate the need for NAs to manually check for the presence of anesthesia equipment. Implementing this system preoperatively would reduce the time required to complete surgical safety checklists and give NAs more time to complete other tasks. Combining this system with other devices in the future could improve efficiency in the OR and reduce the rate of complications.

References

An Observational Study of Clinical Application Downtimes in an EHR

Rizwan Siwani, MBBS, Mohamed O. Seisa, MD, Vitaly Herasevich MD, PhD
Department of Anesthesiology and Perioperative Medicine, Multidisciplinary Epidemiology and Translational Research in Intensive Care (METRIC)

Mayo Clinic, Rochester, MN

Abstract

Health care delivery at our institution depends on our comprehensive electronic health record that is comprised of several applications. Any system is susceptible to downtime, which can negatively impact patient health outcomes. In this study we evaluated the incidence and causes of downtimes of applications critical to care delivery in a large academic tertiary care center.

Introduction

There has been an increase in uptake of EHR systems by hospitals in the United States. A data brief from the Office of National Coordinator of Health Information Technology reported that 96% of hospitals had a certified EHR in 2015. This increased use has exposed hospitals to the unintended consequence of vulnerability to system downtimes which can have a negative impact on care delivery and therefore, patient outcome.

Methods

We conducted a retrospective observational study of clinical application downtimes over a period of six years from January 2011 to March 2016. Six clinical applications used in the operating room setting were evaluated: the anesthesia information management system (AIMS), the picture archiving and communication system (PACS), an integrated clinical viewer that pulls in information from several sources into an easily navigable interface (ICV), computerized provider order entry system (CPOE), and the applications used to schedule surgeries (Surgical List Viewer) and to record information about all surgeries (Surgical Information Recording Systems). An institutional database was used to identify downtime incidents and data on the following characteristics was collected: application(s) affected, time and date, duration, cause, nature of disruption- impairment or outage and scheduled or unscheduled.

Results

There were 484 downtime incidents over 63 months; 25% scheduled and 75% unscheduled. 93% were characterized as impairments and 7% as outages. The median duration was 0.4 hours with an interquartile range of 0.02 to 1.39. The most commonly identified root cause was related to applications (58.7%), then systems software (33.5%). Human errors were minimal and there were no downtimes due to security violations.

Conclusion

Unplanned downtimes of applications in an electronic health record system occur and vary in severity. This study showed that individual applications within a comprehensive EHR are susceptible to application-specific failures. This information can be used to inform response strategies tailored to the application affected. Also, the clinical impact of downtimes is unclear. Accordingly, we are conducting a study to evaluate the effect of downtimes on patient outcomes.

References

Do They Talk About Risk? An exploratory study for developing tools to prevent delirium in older hospitalized patients

Stacey Slager MS,1, 2 Teresa Taft, BS1, 3 Daniel Nystrom, MS1, 3, Bryan Gibson, DPT, PhD1, 3, and Charlene Weir RN, PhD1, 3

1IDEAS 2.0, George E Whalen VA Medical Center, Salt Lake City, Utah; 2Division of Epidemiology, University Of Utah, Salt Lake City, Utah, 3Department of Biomedical Informatics, University Of Utah, Salt Lake City, Utah

Background: Delirium is sudden, cognitive dysfunction, which presents as confusion, a change in level of consciousness, and reduced ability to regulate attention. Delirium occurs in 11-42 % of hospitalized patients and is associated with longer hospital stays, increased admission to long-term care, increased mortality.1 Prior work has demonstrated that 30-40% of cases of delirium can be prevented, if risk factors can be identified.2 In this study we sought to determine the degree to which risk factors for delirium are documented in the electronic health record (EHR) (to address the feasibility of clinical decision support (CDS)) and/or discussed in handovers and rounds.

Methods: We conducted prospective observations of verbal communication between providers of patients aged 65-89 admitted to three VA medical Centers in the Western US between October 2015 and April 2016. These observations were recorded and transcribed. This poster reports on an initial analysis of 11 patients. To determine whether risk factors for delirium were documented in the EHR, we conducted chart reviews for each patient from the day of admit until either they were discharged or Day 4 of their stay, which we compared with observational data of nurse handoffs which were recorded across 2 shifts. We counted the prevalence of documented risk factors3 across observations and notes. To provide clinical context we also collected the terms used in reference to a patient’s mental status as well as risky medications mentioned.

Results: In preliminary analysis of the first 48 hours of 11 inpatients from the Salt Lake City VA, the data showed there were a median of 3 risk factors (0-8 instances of risk per patient) documented in the chart. In nurse handoffs and physician rounds there were a median of 6 risk factors (0-16 instances of risk per patient) mentioned. The agreement between these two methods for the same patient is very low. Figure 1 shows the difference.

Nurses and physician demonstrated clear differences in verbal and written communication regarding mental status. Nurses spend more time with patients and notice behavioral and mental status changes, whereas physicians pay closer attention to medications, infections, and labs. The poster will include the terms used in reference to a patient’s mental status as well as risky medications mentioned.

Conclusion: In this pilot work we have found that risk factors for delirium are commonly documented in the EHR but are much less frequently discussed by Nurses and Physicians in their handoffs and rounds. In addition, attribution of risk factors to delirium risk itself is rare. This finding suggests that computerized decision support to help clinicians prevent delirium based on data entered in the chart is feasible and may be necessary. Current work is addressing the design and implementation of such a decision support system that would aggregate risk-related data to alert providers to pay attention to acute mental status changes. The overarching goal is to create a decision support intervention that will facilitate early recognition of risk and appropriate intervention.

References
Health Care Providers’ Experiences of Moving from a Home-grown EHR System to a Commercial System

Sarah P Slight, MPH, PhD1,2, Diane L Seger, RPh1,3, Christine A. Rehr BA1,3, Sabrina A. Fowler3,4, Elizabeth R Silvers, BA1,3, Adrian Wong, PharmD1,5, Mary G Amato, PharmD, MPH1,5, Nivethietha Maniam1,3, Michael Swerdloff BS1,3, David W Bates, MD, MSc1

1The Center for Patient Safety Research and Practice, Division of General Internal Medicine, Brigham and Women’s Hospital, Boston, MA; 2Newcastle University, Newcastle upon-Tyne, UK; 3Partners Healthcare, Wellesley, MA; 4Michigan State University, East Lansing, MI; 5MCPHS University, Boston, MA.

Introduction: Electronic health record (EHR) systems hold the promise of improving the quality and safety of patient care in hospitals. There has been widespread adoption of EHRs over the last few years, encouraged by the financial incentives offered through the US Meaningful Use Program. Different forms of EHR systems exist, and some hospitals have chosen to replace their existing systems with a fully-integrated clinical and administrative information systems. This study aimed to explore users’ experiences and views on using a home-grown EHR system, and after switching to a commercial system to explore their views on their current system.

Methods: After obtaining the necessary ethical and institutional approvals, a diverse range of outpatient and inpatient providers (including attending physicians, pharmacists, nurse practitioners, physician assistants and residents) were selected. These providers worked at a well-established, integrated delivery system in the northeastern U.S., which was undergoing a move from a home-grown EHR system to a commercial system. All interviews were digitally audio-recorded with each participant’s verbal consent and transcribed verbatim. We adopted an iterative approach to analysis, which enabled us to refine questions, and pursue emerging themes and concepts during subsequent data gathering. A workable list of main- and sub-themes was developed inductively and applied systematically to these data with the aid of the computerized qualitative data analysis software QSR N-Vivo V.11.

Results: Seventy semi-structured interviews were conducted between July 2013 and Jan 2016, each lasting from 10 to 74 minutes. Twenty-seven of these interviews occurred after the EHR transition had taken place, so providers were able to directly compare the move from one system to the other. Five key themes emerged that described important differences between the two systems: (1) how medications were displayed when ordering, (2) default drug doses and directions, (3) the location of pertinent patient information, (4) the timing of medication related alerts, and (5) the quantity of medication-related alerts and their navigation. Medications were displayed in a single drop-down list in the home-grown system, compared to three separate lists in the vendor system, causing some confusion around what the purpose and value of each of these lists were in the new system. Providers found some of the default medication doses suggested by the vendor system to be inappropriate. The vendor system replaced both discrete inpatient and outpatient versions of the home-grown system with a single integrated system, where all the patient’s notes were documented in one place. Although there were clear advantages to doing this, users believed it was now harder to find key information. All clinical decision support alerts were triggered at the time of ordering in the home-grown system; however, in the vendor system, they were presented at the time of signing and shown on one screen, which had advantages and disadvantages. Compared to the homegrown system, providers reported more alert fatigue using the vendor systems; they perceived this to be due to more complex system navigation and an increased quantity of alerts.

Conclusion: Improving the quality and safety of patient care through adoption of EHR systems is a priority area for governments and policy makers worldwide. The five key themes identified in this study provide some useful and potentially transferable insights into how potentially disruptive this change can be from a user’s perspective. This qualitative study was part of a larger project evaluating health care providers’ alert override behaviors, which may explain why the results focused mainly on medication-related elements. This study was funded by grant #U19HS021094 from the Agency for Healthcare Research and Quality (AHRQ).
HIPAA and eCR: Are You Sharing the ‘Minimum Necessary’?

Crystal L. Snare, DVM, MPH1, Julie Simon, MSPH1, Rita Altamore, MD, MPH2, Kim Peifer, MPH3

1Infectious Disease Assessment Unit, Washington State Department of Health, Olympia Washington; 2Office of the State Health Officer, Washington State Department of Health, Shoreline, Washington; 3Communicable Disease Epidemiology, Washington State Department of Health, Shoreline, Washington

Introduction

The Washington State Department of Health (DOH) is engaged in a CDC-funded project to promote electronic case reporting (eCR) of sexually transmitted infections (STIs) with a goal of improving completeness of reporting, when compared to paper reports. This pilot involves submission of reports for chlamydia, gonorrhea or syphilis-positive patients to DOH from a clinical partner using existing certified Electronic Health Record (EHR) and standards-based solutions. According to CFR 164.512(b)(1)(i), the Health Insurance Portability and Accountability Act (HIPAA), providers must share the minimum data needed to satisfy public health investigations, keeping in mind that other laws can be more restrictive about data relating to minors, addiction, genetics, domestic violence, mental health and HIV/AIDS. We summarize findings from the assessment of standard EHR generated records for the presence of contents that satisfy public health reporting requirements and contents which exceed the minimum data necessary.

Methods

DOH partnered with Planned Parenthood of the Great Northwest and Hawaiian Islands (PPGNHI), who uses a certified EHR (NextGen). CDA documents complying with the Continuity of Care Document (CCD) standard were generated by the EHR and securely transmitted to DOH. The data elements and values in these files were compared to those requested by the DOH STI public health reporting form. Data elements present in the CCDs, but not requested on the reporting form, were assessed in light of the HIPAA ‘minimum necessary’ rules.

Results

Sixty-two (62) CCDs were received, representing 17 patients; and included documents in both the EHR’s longitudinal and encounter-based styles. Extraneous or sensitive information was present for 59% (10/17) of patients when compared with information requested by paper case report forms. Examples include 18% (3/17 patients) whose CCD included mental health medications, 6% (1/17) with sensitive procedures, and 12% (2/17) with unneeded diagnoses included. HIV status is included if tested at any encounter reflected in the CCD.

Discussion

We found that unwanted sensitive information is present in unmodified CCD documents submitted as public health case reports, which would not be consistent with HIPAA’s ‘minimum necessary’ requirement. Additionally, some information needed by public health, such as partner treatment plan, is not present; previous work describes this mismatch. Information such as HIV status is present if a relevant test is performed in connection with one of the encounters included in the CCD. This information would be “unnecessary for public health” for reporting non-STI diseases (for example, pertussis or campylobacteriosis) -- but when present in an STI report the information is insufficient, since the date and results of any prior test were absent. Overall, public health case reporting needs were not fully met by data in unmodified CCD and in some cases data present exceeded the HIPAA mandate to share only the minimum data needed to accomplish appropriate public health action. Limitations include using files from a single instance of a single EHR and reliance on manual selection of positive cases. Next steps include exploration of triggering options, further gap analysis, and analysis of larger and more representative sets of records. While public health is committed to reducing burdens on reporters, case reporting implemented using unmodified CCDs produced by certified EHRs will present challenges, both in the necessary information that is absent, and in the unnecessary and unwanted information that is present.

References

Introduction and Background:
The internet has enabled the sharing and viewing of vast amounts of information, including biomedical datasets. However, unlike journal articles, there is no common site where datasets can all be found. Furthermore, due to the myriad of ways data can be developed and stored, finding one dataset does not mean the searcher will have access to datasets with further or related information on the subject.

The biomedical and healthcare data discovery index ecosystem, also known as bioCADDIE, is an effort between informaticians, biomedical researchers, administrators, and publishers to develop a Data Discovery Index (DDI) in order to enable the discovery of available biomedical and biological data in an efficient platform. The initiative is funded as part of the Big Data to Knowledge (BD2K) program by the National Institutes of Health. As part of this initiative, the DatA Tag Suite (DATS) developed in order to describe the metadata and structure a dataset requires to populate the DDI prototype known as DataMed. One of our teams at UCSD is focused on taking multiple data repositories that have data available to the public, ensuring that the necessary metadata to populate DataMed is available. So far there have been over 64 repositories that have been indexed in DataMed. In doing this, we have found there to be a large variety of ways that data repositories organize and represent their metadata, highlighting the need for standardization of datasets and metadata. In this abstract, we illustrate these inconsistencies with several example repositories to emphasize the multiple needs for a standardized platform such as DataMed.

Methods:
We first do an initial inspection to check if the repository is a good fit for DataMed, including checking to see if the datasets represented on the site, and their associated documentation, are available to the public. If a repository passes this initial inspection, it proceeds to the mapping stage. In this stage, we focus on linking the metadata available for the datasets in the repository to the relevant equivalent in DATS. Much of this work is done by downloading the datasets and associated documentation. We then manually find the metadata information of the dataset to the standardized metadata items of DATS. Once this is done the mapping is used to ingest the datasets of the repository into the DataMed pipeline.

Results:
The problem that we are currently encountering while doing the mapping is the heterogeneity in how metadata of datasets is defined and represented. For example, in DATS the entity-property used to identify the name of a dataset is “Dataset: Title”. In comparison, different repositories we have mapped represent this information in many different ways. The National Database for Autism Research (https://ndar.nih.gov) links this value to “General: Title”, while Nature Scientific (http://www.nature.com/srep) refers to this as “Assays: Study Assay File Name”. Many repositories don’t give titles for their datasets, but instead have what we call “Dataset: Identifier”, referring to a unique and primary identifier for a dataset, typically a combination of numbers and letters. In the Human Microbiome project (HMP, http://hmpdacc.org) alone, they host three different types of datasets, all of which refer to the primary identifier in different ways. In the HMP NCBI dataset, the value is referred to as “Submission: Primary_ID”, while in the HMP Metagenomic Samples dataset it is referred to as “SequencingProject: Sequence Read Archive ID”, and in the HMP Reference Genomes dataset it is referred to simply as “Project: HMP ID”. This continues through multiple metadata properties, ranging from information about grants and funding, to software used in data analysis. More examples can be found at https://drive.google.com/file/d/0BxvVDngaYh6vc0NTZ3pjBVkxZVE/view.

Conclusion:
The creation of a platform to allow access to datasets requires the standardization of the data presented, much like what PubMed has done for journal articles. Though at the moment much of this standardization must be manually curated before indexing, we hope that the introduction DataMed and DATs will eventually bring about a level of conformity to data production and organization so as to improve biomedical data ease of access.

Acknowledgements:
This work is supported in part through the NIH Big Data to Knowledge, Grant 1U24AI117966-01
Bayesian Prediction of Asthma Exacerbation in Children

Sunghwan Sohn, PhD1, Young J. Juhn, MD2, Sungrim Moon, PhD1, Chung-Il Wi, MD2, Katherine S. King1, Euijung Ryu, PhD1, Hongfang Liu, PhD1

1Division of Biomedical Statistics and Informatics, Mayo Clinic, Rochester, MN; 2Department of Pediatric and Adolescent Medicine, Mayo Clinic, Rochester, MN

Introduction

Asthma exacerbation is an acute or subacute episode of progressive worsening of symptoms of asthma, including shortness of breath, wheezing, cough, and chest tightness which is not responded to rescue or controller medications, so requiring systemic corticosteroids or visit to urgent care. It is a key cause of major morbidity related to asthma and a long-term decline of lung function. Thus, it is crucially important to have a tool for predicting asthma exacerbation enabling a potential preemptive intervention. In this study, we examined various predictors of asthma exacerbation and employed a Bayesian network (BN) to investigate the collective predictive capability of the predictors for future asthma exacerbation within a year. The predictive performance of the BN was also compared with logistic regression (LR), which is commonly used in the clinical domain.

Materials and Methods

The study cohort consists of 590 children who participated in an ongoing asthma intervention study at Mayo Clinic, Rochester MN. We used patient demographics and medical histories over the past three years (June-1-2012 to May-31-2015) to predict asthma exacerbation within post one year (June-1-2015 to May-31-2016). Of these 590 subjects, 300 subjects (35 exacerbation cases) were set aside as the training set, and 290 subjects (48 exacerbation cases) were set aside as the test set. We investigated various predictors of asthma exacerbation (selected through discussion with clinicians) in electronic medical records (28 total variables, Table 1) that were retrieved from structured data and were also extracted from unstructured data (i.e., clinical notes) using natural language processing techniques. Information gain was applied to rank variables. The prediction performance of a BN and multivariate LR has been investigated using all variables versus top five variables (bold in Table 1). We also examined a resampling technique to balance class distribution (i.e., oversampling).

Results

Table 2 contains the receiver operating characteristic (ROC) area under the curve (AUC) of predictive models with different settings. A BN performed better than LR when using all 28 variables, but they were similar under the five selected variables. Figure 1 demonstrates the ROC curves of a BN and LR using all variables. A resampling technique increased predictive performance with the exception of LR using all variables.

Discussion

Patient past histories relevant to previous asthma conditions and management (top five variables) appeared to be important predictors of future asthma exacerbation. The predictive models on five selected variables demonstrated predictive capabilities comparable to the model using all variables without significant loss of performance (BN) or with even better performance (LR). This suggests that we may use as few predictive variables as possible with adequate performance with less burden of variable extraction, realizing a parsimonious model to be implemented in clinical settings. Our predictive models of asthma exacerbation demonstrate the potential to identify high-risk asthma exacerbation patients and thus assist clinicians in determining appropriate intervention to improve asthma management and outcomes.

Table 1. Predictors of asthma exacerbation (descending order of information gain)

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Information Gain</th>
</tr>
</thead>
<tbody>
<tr>
<td>previous exacerbation</td>
<td>0.765</td>
</tr>
<tr>
<td>symptom</td>
<td>0.691</td>
</tr>
<tr>
<td>hospital visit</td>
<td>0.691</td>
</tr>
<tr>
<td>rescuer medication</td>
<td>0.691</td>
</tr>
<tr>
<td>controller medication</td>
<td>0.691</td>
</tr>
<tr>
<td>skin test</td>
<td>0.634</td>
</tr>
<tr>
<td>pulmonary test</td>
<td>0.634</td>
</tr>
<tr>
<td>NAEPP recommended treatment</td>
<td>0.634</td>
</tr>
<tr>
<td>IgE test</td>
<td>0.634</td>
</tr>
<tr>
<td>race</td>
<td>0.525</td>
</tr>
<tr>
<td>sex</td>
<td>0.525</td>
</tr>
<tr>
<td>missed school</td>
<td>0.525</td>
</tr>
<tr>
<td>asthma control test score</td>
<td>0.525</td>
</tr>
<tr>
<td>viral infection</td>
<td>0.525</td>
</tr>
<tr>
<td>hay fever</td>
<td>0.525</td>
</tr>
<tr>
<td>vitamin D deficiency</td>
<td>0.525</td>
</tr>
<tr>
<td>smoker</td>
<td>0.525</td>
</tr>
<tr>
<td>eczema</td>
<td>0.525</td>
</tr>
<tr>
<td>socioeconomic status</td>
<td>0.525</td>
</tr>
<tr>
<td>season</td>
<td>0.525</td>
</tr>
<tr>
<td>age</td>
<td>0.525</td>
</tr>
<tr>
<td>asthma type</td>
<td>0.525</td>
</tr>
<tr>
<td>comorbidities</td>
<td>0.525</td>
</tr>
<tr>
<td>frequent rescuer user</td>
<td>0.525</td>
</tr>
<tr>
<td>overweight</td>
<td>0.525</td>
</tr>
<tr>
<td>asthma severity</td>
<td>0.525</td>
</tr>
<tr>
<td>asthma compliance</td>
<td>0.525</td>
</tr>
<tr>
<td>food allergy</td>
<td>0.525</td>
</tr>
</tbody>
</table>

Table 2 Prediction of asthma exacerbation (AUC)

<table>
<thead>
<tr>
<th>Variable</th>
<th>BN AUC (SE)</th>
<th>LR AUC (SE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>0.775 (0.765)</td>
<td>0.643 (0.683)</td>
</tr>
<tr>
<td>Top 5</td>
<td>0.736 (0.691)</td>
<td>0.734 (0.680)</td>
</tr>
</tbody>
</table>

( ) without resampling
Mining Associations between Drug Abuse, Age, and Suicide Attempts

Brinda Somasundaram, MS
George Mason University, Fairfax, Virginia

Introduction

According to the Substance Abuse and Mental Health Services Administration (SAMHSA), drug-related suicide attempts in the U.S. has dramatically increased by 41% between 2005 and 2011. This rate is increasing given easy access to both prescription and over-the-counter medications, combined with different drugs including synthetics, making combinations highly lethal. According to the National Comorbidity Survey in 1999, individuals with a substance use disorder are almost 6 times more likely to report a lifetime suicide attempt than those without a substance use disorder. In our study, we use Association-Mining Algorithm (Market-Basket Analysis) to find the most common drug-cocktail combinations related to suicide when admitted in the Emergency Department (ED).

Methods

Data: Drug Abuse Warning Network (DAWN) data collected between 2005 and 2011, were analyzed for Emergency Department (ED) visits involving drug combination-related suicide attempts among patients for the various age categories. ED visits in DAWN include only suicide attempts that involve drugs, but are not limited to drug overdoses.

Tools Used: We used the Data Mining tool Weka, to apply the association mining algorithm to find best rules (most common combinations) of our values for every patient record. The data have been pre-processed using SQL, only records indicating suicide as a reason for hospital admission and records with confirmation of drug abuse were used. Each record values were binarized for the presence or absence of drugs. The data were divided into subsets based on ten age categories such as age 5 or younger, 6 to 11, 12 to 17, 18 to 20, 21 to 24, 25 to 29, 30 to 34, 35 to 44, 45 to 54, 55 to 64, age 65 or older, -8 (age undocumented). The binary data subsets related to suicide were imported to Weka and association mining algorithms (Apriori and FP-Growth) were applied to find the most common combination of drugs used for every age category.

Results

Among patients with suicide attempts, alcohol was almost always present when using drugs. Patients in the age category 45-54 had the highest suicide attempts. These patients most frequently combined alcohol, cocaine, and heroin. Patients in the age category 35-44 had the second highest suicide attempt. Alcohol, heroin and cocaine were the most frequently used. The patients in the age category 12-17 had the sixth highest suicide attempts. Interestingly, these patients frequently were on alcohol, Ibuprofen and Acetaminophen, in that order.

Conclusion

Suicide is a significant public health problem. Health providers should be aware that individuals with substance use disorders are at elevated risk for suicide, take note and educate the patients about the lethality of both drug usage and combining drugs for recreational and non-recreational use. The presented study indicated that adults and teenagers who attempt suicide differ in drugs they are intoxicated with. While alcohol is the main drug in all categories, cocaine and heroin is present in adults, while Ibuprofen and Acetaminophen is present in teenagers.

Acknowledgments

The author is thankful to Prof. Janusz Wojtusiak, Chief, Health Informatics Program, George Mason University, for his guidance in preparation of this poster.

References

Exploration of Semantic Expressions to Define Classification Rules

Margarita Sordo, MSc, PhD,1,2,3 Benjamin J. Gross, MD, MMSc,1
Priyaranjan Tokachichu, MD, MS,1 Christopher J. Vitale, PharmD,1
Saverio M. Maviglia, MD, MSc,1,2,3 Roberto A. Rocha, MD, PhD1,2,3
1Clinical Informatics, Partners eCare, Partners Healthcare System,
2Brigham and Women’s Hospital, 3Harvard Medical School, Boston, MA

Introduction

Creating classification rules for clinical decision support, data reporting, and clinical registries requires manual selection of concepts from reference terminologies into value sets. Manual selection of concepts based on terms and synonyms easily becomes a difficult and time-consuming task that requires specific terminology expertise, and frequently leads to inconsistent and inaccurate value sets. An alternative approach is to rely on semantic relations among concepts in a reference terminology. This alternative approach requires similar expertise and is only possible when the reference terminology associates and decomposes concepts using semantic relations. These semantic properties enable a more sustainable approach that relies on the meaning of the concepts to be gathered.

Methods

We propose the use of queries defined using semantic expressions to automatically select concepts required for value sets. Our clinical knowledge management system (CKMS)1 enables the implementation of queries using semantic expressions. These expressions can reference any semantic relation defined within the system, including those obtained from reference terminologies such as SNOMED CT.2 Given the large number of classification rules that we need to create and maintain, we have been exploring the use of queries created using SNOMED CT semantic relations to generate disease-specific value sets. We implemented a small number of queries and have been evaluating the resulting value sets to confirm the utility of the new method. By using semantic relation:value pairs to identify desirable concepts, we can traverse ontologies to identify relevant concepts based on their clinical meaning. For example, we used the following to identify all relevant concepts for a) Bacterial Pneumonia: causative agent: bacterium (organism) and finding site: lung structure (body structure) and pathological process: infectious disease (disorder); b) Autism: is-a: Pervasive developmental disorder and semantic group: mental or behavioral dysfunction.

Results

Semantic relations in SNOMED CT reflect a hierarchical structure that allows us to select a starting point at the appropriate level, retrieve related concepts in different parts of the terminology, and narrow down our query to specific concepts. This is the case with Autism, where our starting point was the is-a relation Pervasive developmental disorder (which turned out to be too general for our purposes, returning diseases not related to Autism, e.g. Rett’s disorder). Adding semantic group: mental or behavioral dysfunction limited results to concepts specific to Autism. As for Bacterial pneumonia, we selected the semantic relations based on the etiology of the disease: bacterial infection in the lungs. This allowed us to retrieve concepts representing this etiology, even though neither “bacterial” nor “pneumonia” were present in the concept description, “Abscess of lung and mediastinum”.

Conclusions

Queries using semantic expressions seem to provide a powerful alternative to manually searching and retrieving concepts using terms and synonyms. Selecting concepts based on their meaning offers a reliable method to define value sets for classification rules. By taking advantage of semantic relations present within reference terminologies, we can tailor the scope of our search (Autism example) and build queries based on the etiology of a disease (Bacterial Pneumonia). Moreover, such queries can automatically update resulting value sets when reference terminologies are updated. We will continue to evaluate this method, taking into account limitations imposed by existing reference terminologies, including comparisons with manually defined value sets.

References

Multi-Agent System Architecture to Measure Patient Engagement Based on Automatic Patient Authored Text Analysis

Fábio Rezende de Souza BSc1,2, Márcia Ito MD, PhD2, Maíra Gatti De Bayser, PhD3
1Universidade Federal do ABC, Santo André, SP, Brazil; 2IBM Research, São Paulo, SP, Brazil, 3IBM Research, Rio de Janeiro, RJ, Brazil

Introduction

The impact of chronic health conditions in the population requires the adoption of methods to improve the effectiveness of clinical treatments. Patient Engagement can be defined as a set of actions that a patient takes to obtain the greatest benefit from his/her clinical treatment [1]: improving a patient engagement produces a positive impact on his/her treatment outcomes. A multi-agent software architecture was created to implement the method proposed in [2] which assess a patient engagement by classifying sentences of a Patient Authored Text1 (PAT).

Summary description of the Multi-Agent Architecture to Measure Patient Engagement

A PAT Corpus is received by the Engagement Analyzer Agent (a1), which delegates the text processing task to the PAT Analyzer Agent (a2). Agent a2 communicates with the Semantic Analyzer Agent (a4), different language processing services and a lexical database, in order to break the original corpus into functional segments and perform a semantically-based annotation to each one of them. Having this information, Agent a1 asks to the Engagement Degree Classifier (a5) to calculate patient’s engagement degree. In its turn, Agent a5 asks to the Perceptions Classifier Agent (a3) to find and retrieve different types of patient’s perceptions on each functional segment, and based on this information, a5 weighs those perceptions and classify them as being positive or negative, later returning this information to a1. Agent a1 informs the engagement degree to agent a6, and requests information on the engagement stage (according to a previously chosen model of behavior change). Agent a6 returns the information to a1 alongside suggestions on clinical interventions that could improve patient engagement, retrieved by agent a7. Knowledge bases containing information about previous PAT classifications can be used to improve classification of new PAT. The system output is a complete report on patient's engagement information. Figure 1 shows the system architecture.

Conclusion

We introduce a multi-agent software architecture to measure a patient engagement on a health treatment, as the implementation of the theory method proposed by [2]. We consider this could help us to analyze a patient behavior over different perspectives and scenarios.

References


---

1 Medical text authored by patients
The Phenotype Execution and Modeling Architecture: A Roadmap Towards Next-Generation Phenotyping Using Electronic Health Records

Peter Speltz¹, Luke V. Rasmussen², Richard Kiefer³, Jennifer A. Pacheco², William K. Thompson², Guoqian Jiang³, Jyotishman Pathak⁴, Joshua C. Denny¹
¹ Vanderbilt University, Nashville, TN; ² Northwestern University, Chicago, IL; ³ Mayo Clinic College of Medicine, Rochester, MN; ⁴ Cornell University, New York, NY

Introduction
A limitation within the space of electronic health record (EHR)-based phenotype algorithm development is the lack of portability. Currently, phenotypes that are shared on websites such as the Phenotype Knowledgebase (PheKB.org) are implemented independently by different institutions using an amalgamation of custom code (such as SQL), natural language processing (NLP), and frameworks, which execute against various standard and custom data stores. Porting phenotypes implemented in this way to other institutions is cumbersome, time-consuming and error-prone. We have proposed the Phenotype Execution and Modeling Architecture (PhEMA) as a means to minimize the effort to streamline phenotype implementation by standardizing phenotype development and execution against i2b2 and OHDSI data repositories, as well as data stored in local, non-standard data models.

Methods and Results
PhEMA is comprised of multiple components supporting standards-based phenotyping (Figure 1). The Phenotype Authoring Tool (PhAT) is a web-based tool for creating phenotype algorithms utilizing the Quality Data Model (QDM), an information model for representing EHR-based electronic clinical quality measures (eCQMs). We have incorporated PhAT into PheKB.org, enabling PheKB users to launch PhAT to create and save computable phenotype definitions. PhAT also allows the reuse of value sets from the NLM’s Value Set Authority Center (VSAC), as well as the ability to compose new value sets in a local repository. PhAT phenotypes can be exported in various computable formats to be executed externally.

We are currently able to generate an HL7 Health Quality Measure Format (HQMF) representation of the phenotype definition logic, executable KNIME (knime.org) Analytic Platform workflows that may be configured against a local data repository or i2b2, and preliminary work has been completed in generating a JSON representation of VSAC value sets that may be run from within the OHDSI data analytic tools (e.g., ATLAS). Future integration with PheKB will automate publishing the executable formats on PheKB.org for others to download. We have demonstrated portable execution against both local databases and i2b2 instances at up to 8 sites with different EHRs for several phenotypes (e.g., T2DM, BPH), and have compared the results with the original phenotypes.

Conclusion
Portable phenotypes depend on standardization and interoperability. With the increased adoption of common data models for EHR research data warehouses, a set of common phenotype modeling and architecture tools will provide portability of phenotypes, speed up development of algorithms, and increase the availability of portable algorithms. This will in turn increase the overall quality and use of phenotypes in research. PhEMA aims to support standard phenotyping against the most used common data models while still being adaptable to local data models. The tools are still evolving and larger scale experiments and evaluations will be done.

References
Essential Components of a Patient Engagement Platform for Comprehensive Cancer Care

Daniel Stein, MD, PhD¹, Rimma Belenkaya, MS, MA¹, Erdeta Bani, MPA¹, Olga Strachna, MS¹, Peter D. Stetson, MD, MA¹

¹Division of Health Informatics, Memorial Sloan Kettering Cancer Center, New York, NY

Introduction
As healthcare organizations adopt patient-centered care models, they require tools to capture patient history, symptoms, and more formally constructed patient reported outcomes (PRO). While the number of these systems has increased, they are generally one-off tools and, “vary dramatically in their focus and features.”¹ Until such systems are standardized generally available, organizations must share experience in developing them to determine best practices around the tools to collect data from patients, and how to integrate into clinical systems and workflow.

Methods
Over the past 15 months, we developed and implemented a patient engagement platform, MSKEngage. The multidisciplinary team that worked to create this system consisted of software developers, informatics, clinician stakeholders, and patients. In addition to weekly development sprints, the interdisciplinary team met at least biweekly to review development progress and to collect and incorporate feedback. Beyond the core engagement platform that delivers surveys, we report on essential components required to make the collection of patient generated data automated, relevant, safe, and effective.

Results
Since our launch in January 2016, 3,234 patients have completed 8,554 PRO instrument submissions through MSKEngage. In the table below, we summarize the key components we discovered were essential for an implementation that integrated into existing systems and workflow:

Table 1. Categories and specific features of key components implemented in our patient engagement platform.

<table>
<thead>
<tr>
<th>Library</th>
<th>Library of questions/answer content; versioning control; embedded branching logic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Governance</td>
<td>Clinicians, informatics, biostatisticians, patients, determine best instruments, moderate new content requests, minimize “survey fatigue” (redundancy/repeated questions to same patients)</td>
</tr>
<tr>
<td>Patient Interface</td>
<td>Single sign-on with portal; accessible via web or app, in-clinic and/or between-visit reporting; automated reminders, scheduling/calendar; multi-language support; integrated help tool</td>
</tr>
<tr>
<td>Clinician Interface</td>
<td>Single-sign-on with EHR; numerical/graphical longitudinal data review; reporting with severity level; automated commit to EHR when viewed; selective integration into clinical documentation</td>
</tr>
<tr>
<td>Target Cohorts</td>
<td>Target cohort definitions and automated generation; versioning and reuse; scheduling and updates at cohort level; pipeline between cohort and survey management tools</td>
</tr>
<tr>
<td>Complex Scheduling</td>
<td>Accessible UI for scheduling engine built in application allows editing: complex scheduling rules such as multi-occurrence questionnaires, team notification rules, reminder rules and text</td>
</tr>
<tr>
<td>Smart Notifications</td>
<td>Real-time automated alerts to system-derived care team; threshold based notification for severity with instructions to patients and escalation to clinicians; notification of poor adherence</td>
</tr>
<tr>
<td>Advanced Reporting</td>
<td>In-application module; real-time calls for generation and display; configuration based on rules determined by clinician; advanced statistics models, normative data, and graphical display</td>
</tr>
</tbody>
</table>

Conclusion
Beyond the survey tool itself, there are numerous important components required when developing a comprehensive and integrated patient engagement platform, from governance of survey items, to multiple user interfaces, to advanced cohort management, scheduling, notification and reporting components. Specialized (e.g. cancer centers) and generalized healthcare organizations can benefit from sharing experiences and best practices as we work to integrate patient engagement systems into clinical practice.

References
Technology Integration Officer: A New Residency Leadership Position

Kevin U. Stephens Jr., M.D., M.B.A, Rachelle E. Toman, M.D., Ph.D., FABC
1Georgetown University-Providence Hospital, Washington, D.C.

Introduction
The widespread adoption of electronic health records has created an increasing demand for physicians who work with information technology departments to support clinical practice and health systems. This has led to the growth of positions such as Chief Medical Informatics Officer. The need to train physicians for these roles has culminated in the board-certified subspecialty of clinical informatics where physicians are trained to transform healthcare by “analyzing, designing, implementing, and evaluating information and communication systems that enhance individual and population health outcomes, improve patient care, and strengthen the clinician-patient relationship”1. However, many residents are unaware of this field and even less have structured meaningful experiences in clinical informatics. To increase the exposure of resident physicians to this field in a practical way, we created a new residency leadership position entitled Technology Integration Officer (TIO).

Technology Integration Officer: Role & Responsibilities
The official description of this role is to: identify and promote the use of technologies to support resident achievement, collaborate with appropriate technical staff to optimize available hardware and software to meet the needs of residents, and advocate for resident technological interests.

The TIO worked with the Medical Director to develop a longitudinal project to enhance the use of technology within the residency program. The project chosen for this past year was implementing secure texting among the residency program. The TIO was responsible for developing an implementation plan, collecting initial survey information, running a test pilot, and finally implementing secure texting among the residency program.

Additionally, the TIO was responsible for identifying and circulating an “App of the Month” among residents and faculty with a mobile app related to improving patient care.

Lastly, the TIO served as a liaison between the residency and the information technology staff/hospital leadership. During the first year of this position, there were two relevant committees that the TIO served on: The Physician Governance Committee and the Clinical Transformation Committee. Both committees were composed of key hospital leadership from different departments (e.g., pharmacy, laboratory, and nursing) with a goal of preparing hospital and medical departments to operate in an advanced clinical electronic environment.

Results
An end of the year survey sent to residents and faculty showed that 92% view the TIO as a good addition to residency leadership positions. 100% reported that providing resident input on EHR optimization and sending out an “App of the month,” were important roles for the TIO. 75% reported educating the residents on clinical informatics, 84% reported serving as a technology resource for residents, and 53% reported completing a longitudinal project, like secure texting were important. Only 23% of the residents and faculty were aware that clinical informatics is a board certified sub-specialty.

Conclusion
The role of TIO is a new residency leadership position that can offer resident physicians meaningful exposure to clinical informatics and project management early in their career. In the future, the TIO role will respond to the feedback of the survey by continuing to expand the TIO’s role as an advocate for EHR optimization (e.g., by having the resident undergo EHR super user training), providing formal lectures to residents on clinical informatics, and improving the perception of the longitudinal projects.

References
Content for Visualizations to Enhance HIV-Related Communication Between Patients and Health Care Providers
Samantha Stonbraker PhD, MPH, RN1,3, Sheyla Richards BA2, Mina Halpern MPH3, Suzanne Bakken PhD, RN1, Rebecca Schnall PhD, MPH, RN-BC1
1School of Nursing, 2College of Physicians and Surgeons, Columbia University New York, NY, 3Clínica de Familia La Romana, Dominican Republic

Introduction
Many persons living with HIV (PLWH) who attend Clínica de Familia in La Romana (CFLR), Dominican Republic may not understand the health information they receive. Information visualizations can help health care professionals provide the complex health information needed for self-management to patients with limited health literacy and recently, are being designed in various settings and to address the information needs of individuals living with various health conditions. A preliminary step in the design of visualizations is to determine necessary content for inclusion. The purpose of this study was to establish content for visualizations by combining patient and healthcare providers’ perspectives on what information is the most useful for effective self-management.

Methods
This study was guided by an adapted version of Wilson’s model of Information Behavior which depicts how information moves within a system. Data collection with 107 patients is reported elsewhere. Interviews with 40 healthcare providers who provide services to HIV positive patients at CFLR were conducted from Jan - Mar 2017. During both patient and provider interviews, participants were asked what information is the most useful for patients to improve self-management. Providers were also asked what information was the most important to teach patients during a clinical visit. Responses from patients and providers were coded and then frequencies of priority topics were compared to identify and prioritize the information that is the most important to share with PLWH. Identified priorities were then compared with findings from the scientific literature to confirm completeness and accuracy. The prioritized list of information needs formed the basis for what content to include in visualizations to promote HIV-related communication.

Results
Demographics of the 107 patient participants are reported elsewhere. The interdisciplinary team of providers included 13 physicians, 10 nurses, 4 community health promoters, 5 mental health counselors, 2 pharmacy technicians and 6 from other positions such as program directors or research assistants. The average age of providers was 38.2 years (Range 25-62) and the average length of time working at the clinic was 5.13 years (Range 0.25-25).

Patient responses (n=104) regarding what information is the most useful for self-management were medications/adherence (54.8%) methods of protection/prevention (13.5%), mental health support (9.6%), general health management (8.7%), healthy eating (4.8%), appointment adherence (2.9%), and other non-specific information (5.8%). Of healthcare providers’ responses (n=40), information about medications/adherence (30%) was also identified as the most useful for patients’ self-management. Providers identified other useful topics to be: free services at the Clinic (22.5%), mental health support (12.5%), HIV in general (12.5%), methods of protection/prevention (7.5%), transmission (5%), other health conditions (5%), laboratory values (2.5%) and family planning (2.5%). While patients and providers agreed on the most useful information for self-management, responses differed in that patients focused on a few central ideas whereas providers identified a wider range of information topics.

Conclusions
PLWH and their healthcare providers identified some similarities and some differences in what information is useful for effective self-management. Future work will be to conduct participatory design sessions to create the visualizations that incorporate the information that both patient and health care providers identified as necessary.

Acknowledgement: First author funded by Reducing Health Disparities through Informatics grant (T32NR007969).

References
Selection and On-boarding of Health Information Exchange Partners:
The DoD and VA Approach

Norman E. Stone III MD¹, Steven F. Kator, MD¹, Roger B. Price MPA², LeAnn Roling RN³, Crystal A. Baum MBA⁴, Kimberly A. Heermann-Do MHA², Margaret Donahue MD³, Lauren C. Thompson, PhD¹

¹ DoD/VA Interagency Program Office, Rosslyn, VA ² Defense Health Agency, Falls Church, VA ³ Veterans Health Agency, Washington, DC ⁴ Defense Health Management Systems, Rosslyn, VA

Introduction & Background
The DoD and VA operate two of the largest healthcare systems in the United States, with 200 hospitals, 1,842 clinics, 18.5 million beneficiaries, and combined budgets of over $221 billion.¹,²,³ Additionally, the Departments purchase over $25 billion in health care from civilian physicians and institutions.¹,⁴ The purpose of this study is to (a.) review the current status of health data exchange between the Departments and civilian partners, and (b.) describe the processes used to identify and on-board health data exchange partners.

Materials & Methods
DoD and VA process documentation and interoperability reports were reviewed, summarized, and compared. The Departments’ enterprise data repositories were queried for purchased care data.

Results
The DoD maintains connections to 23 community partner institutions, exchanging approximately 100,000 health documents per month. The VA maintains connections to 96 community partner institutions and regional HIEs, exchanging approximately 150,000 health documents per month. The DoD utilizes an “opt-out” policy, whereas the VA utilizes an “opt-in” policy. 185 DoD beneficiaries have opted-out of health information exchange (0.2% of 9.49M beneficiaries); 979,696 VA Veterans have opted-in (16% of 6.26M Veterans treated in FY2016; 11% of 9.05M enrolled Veterans).

The Departments evaluate candidate partners based on purchased care dollars spent and number of patients treated. Both Departments are signatories to the eHealth Exchange’s Data Use and Reciprocal Support Agreement (DURSA), and require eHealth Exchange participation by candidate partner institutions.

Conclusion
The DoD and VA exchange a substantial amount of health data with community partners, and regional HIEs. There are opportunities to expand participation with community partners and regional HIE’s, and to exchange a greater quantity and variety of documents. Expansion should be guided by cost and quantity of care purchased for each community partner.

References
Operationalizing Patient Generated Data Collection through Engagement and Iterative Visualization Design

Olga Strachna, MS¹, Erdeta Bani, MPA¹, Daniel Stein, MD, PhD¹, Peter Stetson, MD¹
¹ Division of Health Informatics, Memorial Sloan Kettering Cancer Center, New York, NY

Introduction and Background

Patient-generated data (PGD) in the form of patient reported outcomes (PRO) and care experience outcomes are becoming widely and routinely collected across various healthcare settings. The goal is to utilize this information to assess patients' well-being and perception of experience as related to their treatment, not only as a population but also as individuals. Enabling clinicians to truly leverage PGD within their established workflows remains a challenge. Beyond merely integrating PDG into the medical record, this new data must be translated into clinical insights and actionable knowledge at the right time. Furthermore, successful integration of PGD into clinical practice requires clinician engagement, as well as patient engagement within the process.

Methods

We deployed a human-centered design framework for visualizing PGD data as it is collected, and for converting PGD into actionable knowledge for clinicians. The design approach was conducted in iterative phases of complexity, where the clinical staff provided feedback throughout each cycle of dashboard deployment. One end goal was to increase clinician understanding of information collection bottlenecks so that they can better engage their patients. A second end goal was for the care team to analyze the patient responses at the population level, and at the individual level. This would allow clinicians to reflect on clinically meaningful differences to track over time across a population of patients, while at the same time allowing them to communicate more effectively with individual patients within the normal clinical workflow. We tested the design concepts by performing cognitive walk-through exercises, such as reviewing data across a panel of patients, sharing the data with care team members, and estimating the quality or meaningfulness of the data. We also collected feedback on visual design preferences, points of comparison as well as selecting a subset of filters to drill down to some common key questions.

Results

Through our electronic PGD collection tool we captured a wide breadth of data from patients, including quality of life data obtained before and after treatment, a surgical experience questionnaire and a 10-day daily symptom tracker of adverse events post-surgery, as examples. For every instrument deployed we engaged clinicians to develop an analytical dashboard for tracking high level summary measures and measures of patient and system engagement, such as portal sign up rate, survey opening rate, completion rate and patient response times. In addition, a population-level dashboard was developed that allowed care teams to review aggregate patient responses. We found that in order for the clinical dashboards to be successfully adopted by clinicians it was important to normalize scores, to track changes in scores over time as absolute and relative differences, and to maintain consistency in scoring results based on directionality of the data across all questions asked.

Discussion and Conclusion

We have found that the most important factor in successfully integrating PGD into clinical care is the early engagement of clinician stakeholders in the design, deployment and data analysis phases of each project. This was addressed by creating workgroups around reviewing data, which become the standard operating procedure for each instrument. Each workgroup established a regular meeting time to review operational data and tracking of patient engagement metrics. They reflected on the quality of the data and requested changes to the instruments, such as a reduction in questions asked or tuning of notifications received by the care team. In summary, early clinician involvement in deployment of PGD collection, and visualizations of the data are key factors for success of PGD projects.

References

Trend Analysis of EHR Events to Facilitate Hypothesis Generation and Data Quality Assessment

Yohan Sumithipala¹, Vojtech Huser¹
¹Lister Hill National Center for Biomedical Communications, National Library of Medicine, National Institutes of Health, Bethesda, MD

Introduction

Common Data Models (CDMs) for Electronic Health Record (EHR) data facilitated the creation of software tools that help with data characterization and quality assessment. In 2016, an informatics group named Data Quality Collaborative (DQC; repository.edm-forum.org/dqc) proposed a set of pre-computed and highly stratified data characterization measures that produce comparable and standardized outputs from multiple common data models, such as Sentinel, Patient-Centered Outcomes Research Network (PCORnet) and Observational Medical Outcome Partnership (OMOP). DQC measures can be further analyzed without requiring access to sensitive patient-level data. Using these DQC data measures, our study develops methods for analyzing temporal trends in EHR events to identify trend patterns, such as significant decline in event prevalence, that lead to hypothesis generation, epidemiologic analysis, or a data quality inquiry.

Methods

Using DQC compliant measures from the Achilles tool, created by the Observational Health Data Sciences and Informatics (OHDSI) consortium, we analyze temporal EHR event trends for structured data about drug ingredients, drug products, procedures, lab results, and diagnoses. We use rule-based, statistical, and machine learning approaches to classify each temporal trend (such as use of a certain drug ingredient) into several categories (e.g. significantly rising, relatively constant, significantly declining, or indeterminate). Because the population observed within the dataset can change over time, we stratified the analysis by age decade (0-9, 10-19, etc.) and normalized the trends to prevalence per 1000 people. We used 1 year temporal intervals, to eliminate seasonal variation.

Our rule-based approach uses goodness of linear fit to determine clearly rising or declining events. In the statistical approach, we used an ARIMA model that can account for inherent autocorrelation of time series data. We also piloted a machine learning approach.

Preliminary Results and Discussion

Our analysis (written in the R programming language) utilizes CSV input files with DQC measures data. For OMOP datasets, we also created data extraction script that uses standard OHDSI R packages. The results presented below were generated using a claims dataset from the Innovation in Medical Evidence Development and Surveillance program of the Reagan-Udall Foundation for the FDA. To eliminate rare events, we only considered event-age decade strata with a minimum prevalence of at least one person per ten thousand people in at least one calendar year. To illustrate the size on three event types, we analyzed 1,689 drug ingredients, 12,488 procedures and 8,654 diagnoses. Illustrative results for ingredients are the following: 642 drug ingredients passed the minimum prevalence threshold, and a pre-defined trend category was assigned to 577 of them (34% of all 1,689 ingredients). Examples of rising ingredients include ondansetron, cefdinir, nebivolol, mupirocin, or gadobutrol; declining ingredients include tegaserod, isomethptene, nabumethone, verapamil or triamterene. Additional results for other event types are available on the project website (github.com/OHDSI/StudyProtocolSandbox/tree/master/OHDSITrends). We further developed methods that allow trend examination of multiple events related to each other into event groups (e.g., ingredient-drug class, drug product-drug class, ingredient-disease indication or procedure-procedure group). Our work has limitations inherent to the EHR-only population (rather than general population). Because a single dataset may not cover all age-decades, we also created a framework that aggregates significant trends across multiple datasets.

References


Acknowledgement: This research was supported by the Intramural Research Program of NIH, NLM.
Clinical Machine Comprehension Using Case Reports

dr. Simon Šuster, prof. dr. Walter Daelemans
University of Antwerp, Antwerp, Belgium

Machine reading and text comprehension aim at complete large-scale understanding of natural language texts\(^1\). A system is expected to read a document and answer questions about it. The task, which resembles question answering, could have important applications in the clinical setting\(^2\). Recently, there has been growing interest in the development of machine comprehension algorithms and datasets, but these works are limited to the general newswire domain\(^3,4\). Our work represents a first step towards studying machine comprehension \textit{in the clinical domain}.

We set up a task in which a system is presented with a query (a sentence with a missing entity) and a document, and is then expected to infer the answer (the missing entity) from the unstructured document. Our dataset is constructed from BMJ Case Reports, the largest online collection of clinical case reports. Each case report includes a “Learning points” section which summarizes the document; we automatically create queries by taking each sentence in that section, identifying clinical concepts in it, and removing one concept. For example, from a report describing comorbid disorders of ADHD, we would obtain the query \textit{Patients with ADHD have higher incidence of \_\_\_\_\_\_\_.} The goal would be to correctly predict the missing entity \textit{enuresis}. As the phrasing in the summarizing sections of the case reports often differs from the one in the document, machine reading systems need paraphrasing and inference capabilities to provide the right answer, which makes the task challenging.

In the poster, we will report the performance of baseline vector-similarity matching and of a state-of-the-art reader based on recurrent neural networks. We will also show how evaluation poses a particular challenge: Often a good answer might be given, but which might not be exactly the same as the expected one. Returning to the example above, in addition to “enuresis”, an equally good answer would be “bedwetting”. We give credit to such answers by expanding the answer set with synonyms available from the UMLS\(^\text{TM}\) Metathesaurus\(^\text{TM}\). To find out how difficult answering is for humans, we carry out a manual analysis on a hundred document-query pairs. We report some of the dataset statistics in Table 1.

\begin{table}[h]
\centering
\begin{tabular}{|l|c|}
\hline
Number of clinical cases & 12300 \\
Average number of queries per case & 9 \\
Average length of a case (tokens) & 1414 \\
Number of “problem” answers & 74,000 (67\%) \\
Number of “treatment” answers & 24,000 (22\%) \\
Number of “test” answers & 12,000 (11\%) \\
\hline
\end{tabular}
\caption{Selected statistics of the case report dataset.}
\end{table}

References

Sharing exome sequencing data between Clinical Sequencing Labs and Healthcare Providers

Rajeswari Swaminathan, MS¹, Yungui Huang, PhD¹, Katherine Lintner, PhD¹, Matthew Pastore, MS, LGC², Sayaka Hashimoto, MS, LGC², Theodora Jacobson, MS, LGC², Danielle Mouhlas, MS, LGC², Simon Lin, MD, MBA¹

¹Research Information Solutions and Innovation, The Research Institute at Nationwide Children’s Hospital, Columbus, Ohio, USA; ²Institute of Genomic Medicine, Nationwide Children’s Hospital, Columbus, Ohio, USA

Introduction

Genome sequencing is being rapidly adopted for clinical diagnostics. Approximately 1 in 4 patients tested with whole exome sequencing gets diagnosed with a genetic disease. Clinical sequencing laboratories return a summary report of results to clinicians and in addition, with appropriate consent from the patient, may also provide the raw sequencing data in the form of FASTQ or VCF files. Although healthcare providers typically prefer to contact the ordering lab for a reanalysis of the data with emergence of new information in the knowledgebase, having access to the raw data provides them with the flexibility to take a deep dive into variants beyond those discussed in the clinical summary; this becomes important especially when patients decide to enroll in a research study. Currently, sequence data is either shipped through hard drives or transferred over the network and there is no strict protocol governing this process. We highlight the workflow of transfer established between our institution and a major clinical sequencing laboratory. We aim to understand the current state of genomic data exchange between healthcare providers and genetic laboratories and discuss some of the key lessons learned.

Methods

During the observation period of this study, from April 2016 to December 2016, the genetics clinic at our institution sent all clinical WES orders for their patients to one of the major clinical sequencing laboratories and tracked the entire process from consenting the patient to receiving the raw data. The Clinical Lab uses a “Managed File Transfer System” to facilitate the transmission of data between organizations. A detailed workflow of the transfer process is depicted in Figure 1

Results

The first lesson learnt is that the entire process is manually controlled, leaving room for errors. Email notifications are sent manually once the sequencing is complete and currently there is no easy way to know if all consent forms got received at the laboratory end. For 2 out of the 19 patients, data was available for download from Baylor but the consent forms were missing initially and took a while to be retrieved. Secondly, different patient identifiers were used through the entire process leading to confusion. 4 patients had a DNA Sample number mentioned in their consent forms, which is an identifier used by the clinical laboratory, whereas our institution uses Patient Name + DOB to uniquely identify a sample, thus demanding an extra mapping phase to complete the validation process.

Discussion

Consortiums like the Global Alliance for Genomics and Health provide a standardized interoperable genomics framework to assist data sharing between researchers and clinicians. [1]. In order to avoid the high costs associated with current genomic data transfer methods, data can be stored in centralized cloud based repositories and inter operable systems can be developed on top of it to provide seamless access to the data in a secure manner.

References

Identifying Disease Hierarchical Relation with Concept Embeddings and Distant Supervision

Lingyi Tang, MS, Yaoyun Zhang, PhD, Hua Xu, PhD
School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX, US

Introduction

The hierarchical relation, also called the hypernym-hyponym relation, is one type of essential relations to structure biomedical concepts in knowledge bases and ontologies. It is critical to support a wide range of biomedical applications. However, it is both time-consuming and labor-intensive to update the rapidly emerging biomedical concepts and their hierarchical relations manually. Moreover, current works in this field mainly leverage lexical-syntactic patterns and domain knowledge to extract hypernym-hyponym relations from biomedical text, which can only cover a small proportion of complex linguistic structures. To address these problems, this study proposed to apply a novel word embedding-based approach developed by Fu et al. 1,2 to identify hypernym-hyponym relations between concept pairs, with identifying disease hierarchical relations as the use case.

Method

As illustrated in Figure 1, our automatic system for disease hierarchical relation identification is built on the distant supervision framework. Firstly, distributional word representations (i.e., embeddings) trained from the abstracts in the MedLine corpus are used to generate distributional representations of disease concepts. Then we further expand the original method to accommodate multiple-word concepts, by integrating single-word embeddings to construct distributional representation of multiple-word concepts. Next, existing knowledge of hypernym-hyponym relations in the Disease ontology is employed to train the relation identification model, without any manual annotation. To further improve the relation identification model, samples are first clustered into different groups, each of which represents a similar disease category, and an identification model is then built for each cluster using the logistic regression algorithm based on concept embeddings.

Results & Conclusion

Evaluation with a test set of relations in the Disease ontology demonstrated that the performance of our proposed approach is promising, with a precision of 82.18%, a recall of 83.00% and a F-measure of 82.59%. Figure 2 illustrates the learning curve of the hierarchical relation identification model by tuning the relative error threshold of the logistic regression algorithm. We will leverage additional ontologies such as UMLS and SNOMED-CT to expand our training set in the next step. Furthermore, we will develop such embedding-based novel approaches for extracting hypernym-hyponym relations of other concept types such as chemicals and genes in the biomedical domain.

References

A Practical Scientific Workflow of Sharing Large Selected Dataset from Clinical Research Data Repository

Shiqiang Tao, PhD¹, Steve Roggenkamp, BS¹, Guo-Qiang Zhang, PhD¹
¹Institute for Biomedical Informatics, University of Kentucky, Lexington, KY

Abstract
Clinical research data repositories allow researchers to navigate and select target dataset and download it under certain data use agreement, but the data downloading efficiency and reliability become challenges when the selected data set is very large (terabyte level). To address these challenges, we implemented a data sharing workflow which integrates rsync for efficient and reliable file transferring and SSH for authentication. Our preliminary results showed that the new scientific workflow achieved more than 10 times of the reliable file transferring speed compared to previous HTTP file downloading.

Introduction
The Center for SUDEP Research (CSR) ¹ is a the NINDS-funded Center for SUDEP Research. CSR built a central data repository integrating epilepsy data captured from participating institutions. This central data repository provides a much larger patient data pool for Analysis. In addition, CSR informatics core created a comprehensive infrastructure called MEDCIS² to manage researchers’ access to the data repository. MEDCIS allows researchers to navigate patient data by characteristics such as demographics, diagnoses, medications, and symptoms. Researchers can build queries by customizing those characteristics and then retrieve the data set that matches the queries. MEDCIS previously provides a web interface for downloading the selected data set, but it is not ideal in neither speed or reliability. In this work, we describe a workflow designed for MEDCIS that can practically handle terabyte-level data downloading reliably and securely.

Methods
We designed a scientific workflow of 9 steps illustrated in Figure 1. First of all, we create a shared account on the CSR data server only for data downloading purpose. The workflow starts by generating a pair of SSH keys: a public key and a private key. Researchers keep the private key and upload the public key to MEDCIS. The public key is added to the authorized keys of the shared account on CSR data server. Then researchers continue to use MEDCIS to finish their data selection process. After that, researchers can generate a data downloading request letting MEDCIS know that they want to download data using the new workflow. A unique data-download token is generated together with the creation of the data downloading request. MEDCIS creates a folder with the token as the name and prepares all selected data files under that folder. At last, MEDCIS sends back a command to researchers with which they can start to download the data. The command uses rsync to handle data transfer and the authentication is done automatically by the SSH key configuration.

Results
CSR central data repository is located at Case Western Reserve University in Cleveland. We moved 215 files of total size 31.3 GB data to University of Kentucky in Lexington, which took 46.5 minutes with an average file transfer speed 11.2MB per second. Data transfers to University College London showed that data transfer among different continents could also achieve speed around 10 MB per second.

Conclusion
We design and implement a scientific workflow for sharing accurately selected data set from large clinical research data repository. Our preliminary test suggests that the workflow can achieve a practical transferring speed for terabyte-level. The workflow simplifies the authentication using SSH keys and with rsync, it can pick up from where it is left off if interrupted.

References
Developing a framework for a comprehensive data-sharing program
Asba Tasneem, PhD, Karen Chiswell, PhD, Brian McCourt, Matt Gross, Eric Peterson, MD, MPH, Michael Pencina, PhD
Duke Clinical Research Institute, Durham, NC, US

Introduction
There has been a surge in calls to improve transparency in clinical research.⁴ In response, many in academia and industry have developed platforms for sharing clinical trial data.⁴ The Duke Clinical Research Institute (DCRI) launched Supporting Open Access for Research (SOAR)—an initiative that includes the DCRI, academia, and industry designed to enable clinical research data sharing for the benefit of the broader research community. Under SOAR, Bristol-Myers Squibb (BMS) collaborated with the DCRI to enhance transparency of its clinical trial data-sharing program.⁵ In separate collaborations with SAS Institute, Inc. (SAS) and the American Heart Association (AHA), the DCRI has made accessible to researchers data from the Duke Databank for Cardiovascular Disease (DDCD).⁶,⁷

Methods
We developed a workflow to improve operational efficiency in processing data-sharing requests. Data requests (including data use agreement) are assessed for completeness, then reviewed by the SOAR-appointed Independent Review Committee (IRC) with expertise in biostatistics, research ethics, patient privacy, and the clinical specialty of the proposal. Upon approval from the IRC, the de-identified dataset(s) are provisioned for use via a secure analytics environment. The IRC also reviews the results prior to submission to assess concordance with the stated analysis plan.

Results: Analysis datasets, Secured environment, and Governance

Datasets: The DDCD is among the largest and oldest single-site cardiovascular databases in the world. The DukeCath analysis dataset, extracted from the DDCD and de-identified, includes ~150,000 cardiac catheterization procedures conducted in ~84,000 adult patients at Duke between 1985 and 2013. The DukeCath dataset is shared in two ways: 1) the de-identified research dataset is made available through SAS and AHA platforms; 2) an anonymized educational dataset is available via the SAS platform or direct distribution.⁵,⁷

Environment: The datasets are provisioned to investigator(s) in a secured analytics environment where access is controlled at user level and all exports of patient-level data are restricted without additional approvals. The analytics platform is equipped with statistical analysis tools to aid research analysis. The SAS environment provides SOAR sharing and analytical capabilities plus opportunity for collaborative analysis with industry datasets. The AHA Precision Medicine Institute platform (in development by Amazon Web Services) provides sharing and analytical capabilities plus prospective integration with related cardiovascular clinical and laboratory data.⁶

Governance: The Duke University Institutional Review Board approves all SOAR practices and use of the DukeCath data. The SOAR IRC reviews and evaluates each request for 1) statistical analysis plan, 2) plan to protect patient privacy, 3) dissemination plan, and 4) qualification of the investigator(s). BMS uses SOAR IRC services for reviewing in-scope proposals. As of February 2016, all eight BMS requests were approved by the IRC for data sharing and the first DukeCath researcher is accessing the production system.

Conclusion
The primary purpose of the SOAR program is to facilitate open sharing of clinical research data with responsible researchers to promote open science and allow investigators to verify reported results as well as pursue interesting secondary uses of existing data. This increased transparency will inform science and improve patient care. We plan to evaluate additional clinical research datasets and share eligible datasets with researchers and educators.

References
2. Executive Office of the President. Increasing access to the results of federally funded research. Link. Feb 2013
5. Pencina et al. Supporting open access to clinical trial data for researchers. Am Heart J. Feb 2016

2204
An Evaluation of Syntactic Dependency Parsers on Clinical Data
Stuart J. Taylor, BS\(^1\), Travis R. Goodwin, MS\(^1\), Sanda M. Harabagiu, PhD\(^1\)
\(^1\)University of Texas at Dallas, Richardson, Texas

Introduction
With electronic medical records becoming more widely adopted, the potential to extract information from the narratives in medical records is increasing. In clinical decision support systems, quality control, medication reconciliation, and other biomedical applications\(^1\), Natural Language Processing (NLP) techniques can be used to automatically discern knowledge from text. The accuracy of NLP techniques often heavily depends on the type of data used to train or develop them. Advanced NLP techniques – like semantic role labeling – require accurate syntactic parsing. Syntactic parsing reveals the syntactic structures (e.g. trees) in which words participate and discovers the syntactic roles of the words in a sentence. In this work we evaluate whether existing annotated clinical text is sufficient to train state-of-the-art neural parsing methods.

Methodology
We used the MiPACQ\(^1\) clinical text corpus with a random 17:1:2 split for training, development, and testing. We trained and evaluated two neural parsers: (1) Stanford’s greedy transition-based parser\(^2\); (2) SyntaxNet\(^5\), Google’s globally normalized transition-based parser; and two non-neural parsers: (3) OpenNLP’s\(^\ast\) chunking shift-reduce parser; (4) ClearParser\(^3\), a shift-pop transition-based dependency parser used by cTAKES\(^4\). Additionally, we evaluated the performance of each parser on clinical text when using their general (i.e., non-clinical) pre-trained models.

Results
We computed the Unlabeled Attachment Score (UAS) on the clinical test data to perform our evaluations. The UAS is the percentage of words which have the correct syntactic head. Syntactic heads are used to encode parent-child relations in a parsing constituent (e.g. in “stimulation increases flow” the noun “stimulation” is the child of the verb “increases” because the verb is the head of the phrase and there is a NOMINAL SUBJECT relation between the head and the noun). The results on the test set are shown in Table 1. The UAS-G column shows the UAS score for each parser when using the pre-trained general-purpose model, and the UAS-C column shows the UAS score after training each parser with the clinical training data.

<table>
<thead>
<tr>
<th>Parser</th>
<th>UAS-G</th>
<th>UAS-C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stanford</td>
<td>78.47%</td>
<td>80.66%</td>
</tr>
<tr>
<td>OpenNLP</td>
<td>74.04%</td>
<td>81.20%</td>
</tr>
<tr>
<td>ClearParser</td>
<td>39.11%</td>
<td>83.08%</td>
</tr>
<tr>
<td>SyntaxNet</td>
<td>63.76%</td>
<td>85.19%</td>
</tr>
</tbody>
</table>

Conclusion
While SyntaxNet performed the best after training on clinical text, all parsers showed improvement. These results suggest that the MiPACQ corpus is sufficient to train neural syntactic parsers. Interestingly, the parsers which performed the best with the pre-trained model performed the worst after being trained on clinical text. Directions for future research include: (1) tuning hyper-parameters; (2) incorporating more training data; and (3) evaluating on additional types of clinical data.

References

\(*https://opennlp.apache.org/\)
Overcoming Policy and Operational Challenges for Health Data Exchange Using a Validation Service for Patient to Provider Relationships

Sidney N. Thornton, PhD1,2, Shan He, PhD1, Darren K. Mann, BS1, Iona Thraen, PhD2,3, Matt Hoffman MD4, Wu Xu, PhD2,3
1Intermountain Healthcare, Salt Lake City, UT; 2University of Utah; 3Utah Department of Health, Salt Lake City, UT, 4Utah Health Information Exchange, Salt Lake City, UT

Introduction

The network of connectivity between healthcare systems broadens for Health Information Exchange, yet workflow barriers persist. Useful data do not always arrive at the right place in time to make informed decisions. Available data from external sources is frequently left unviewed and unincorporated in the provider’s system. Disclosures of patient data solely for undesignated future access is suboptimal for privacy and risk management policies. Intermountain Healthcare, an integrated delivery network serving Utah and southern Idaho, creates more than 18,000 clinical summaries every day for exchange, but until recently, has witnessed only a few dozen records accessed daily by the community care team. Workflow automation of data integration has been complicated by data quality concerns that can be rectified in part through relationship validation. Other policy barriers to external data usage include local rules prohibiting the open exchange of the state government held data. Recently, stakeholders in Utah described an exchange framework (ThSisU)1 to accommodate policy and operational concerns that would facilitate workflow automation for external data integration. The framework includes a foundational component to automatically validate patient-to-provider treatment relationships using central data governance. Standard-based exchange protocols initiate the relationship validation services that in turn call encounter or registry services convenient to the local system as trusted sources of patient and provider relationship data.

Methods

Three early adopter use cases are examined to determine how the relationship validation service improves the appropriate and timely uptake of inter-organizational data, while balancing local policies and privacy interests. The first use case uses an administrative scheduling process as a source of attestation of relationship to automatically pull clinical summary data (Cross Community Access/XCA) into the external organization EHR. The second use case leverages the patient matching service (Cross Community Patient Discovery/XCPD) to access a provider attestation list that in turn sends a filtered view of the facility encounter notifications to the external notification service. The third use case requires the relationship validation to link government Fact of Death information to the external provider organization (XCPD and Cross-Enterprise Reliable Document Interchange/XDR).

Results

In use case 1, the population of persons whose health care is shared across the exchanging organizations exceeds five thousand daily encounters. Appropriate data integration for patients with validated providers in both organizations improved from less than 1% to over 60% with improved workflow and data quality. In use case 2, nearly 2.5% of the facility encounters have attested relationships that are received by the external providers. With the implementation of the relationship validation service, nearly 24,000 daily PHI events are protected from inadvertent disclosure thereby supporting the local policies of minimum necessary data while still meeting the community purpose of the notification system. In use case 3, attestation of patient-to-provider relationship through the patient matching process has enabled the incorporation of Fact of Death data from 0% (since 2014) to over 99% in a manner acceptable to policy of the State’s Risk Management.

Discussion and Conclusion

Additional use cases are in various stages of development for care coordination, public health, and shared financial risk that will leverage the statewide service for patient-to-provider relationship validation. The centralized function of the relationship validation service adapts to a variety of technical platforms suggesting reasonable scalability. Simpler transactions and workflow automation would be supported by a single authoritative source of relationships with central data governance. Future investigations will study workflow behavioral tied to system enhancements.

References

Assessing Data Quality within Health Information Exchanges: 
A Case Study for Supporting Emergency Department Research

Margaret M Thorsen, BA1, Indra Neil Sarkar, PhD, MLIS1,3, Elizabeth Chen, PhD1,3, 
Elaine Fontaine, BS2, Gregory Walker, PhD2, Megan Ranney, MD, MPH1,4 
1Alpert Medical School, 3Center for Biomedical Informatics, 4Department of Emergency 
Medicine, Brown University, Providence, RI; 2Rhode Island Quality Institute

Introduction
Health information exchanges (HIEs) support the potential for improving healthcare quality, reducing unnecessary 
treatments, and promoting provider communication. Most HIE research to date has focused on issues of usability, 
implementation, and clinical decision making.1 The utility of HIE data for clinical research and population health 
analyses has yet to be fully explored.2 For instance, HIE data from state-based exchanges could be a source of 
robust, multi-source data to inform public health surveillance and interventions. However, little is known about the 
quality of HIE data for enabling their use for research purposes. Emergency Department (ED) data are well-suited 
for exploring HIE-based clinical research, as the ED is accessible to patients irrespective of health insurance status, 
referral, or ability to pay. Furthermore, EDs have started to engage with HIEs3 for conducting clinical or population 
health research. The objective of this study was to conduct a preliminary characterization of data quality within a 
single, statewide HIE and provide an early perspective on the challenges in using HIE data for supporting 
subsequent retrospective research studies. We also sought to compare the quality of HIE data to institutional EHR 
data (a typical source for research) – both in terms of representativeness and of meeting HIE standard requirements.

Methods
Rhode Island’s opt-in HIE, “CurrentCare,” is owned and operated by the Rhode Island Quality Institute (RIQI). 
Over 350 practice sites in Rhode Island, including all major hospital networks, use CurrentCare (comprised of 442 
data sources). Fifty percent of statewide ED visits in Rhode Island occur at a single hospital network. De-identified 
data from this hospital EHR and CurrentCare were analyzed in aggregate for all patients who had visited an ED in 
the hospital system between August 1, 2015 and March 31, 2016 focusing on: age, gender, race/ethnicity, diagnosis 
codes, problem list, Primary Care Physician status, and disposition. This study was deemed exempt by the IRB. The 
content of the CurrentCare HIE data fields of interest were compared to terminology standards (e.g., those defined 
for the HL7 Continuity of Care Document [CCD]).

Results & Discussion
Sample size, mean age, gender, and primary care access were determined. Disposition, race, and ethnicity were the 
most challenging to map to a standard. For instance, more than 40 non-standard values were found in the HIE data 
for race. This study revealed data standardization challenges, which are an artifact of multi-source HIEs, which can 
pose significant barriers to the prospective use of HIE data for research. Next steps will include the development of 
systematic methods for mapping heterogeneous values for similar concepts to recognized standards, further 
exploration and addressing data quality issues (e.g., accuracy and completeness), and examination of the use of HIE 
data in comparison to ED patient records from the electronic health record (EHR), a widely used source for research 
data. Collaboration between clinicians, hospitals, EHR vendors, clinical informaticians, and HIE curators is essential 
for the future success of HIEs as both a statewide clinical and research tool. Future efforts should quantify these data 
discrepancies and explore strategies for overcoming these barriers to enable HIE-based research without changing 
data collection processes.

References
2. Parker C, Weiner M, Reeves M. Health Information Exchanges – Unfulfilled Promise as a Data Source for 
3. Johnson K, Unertl K, Chen Q, Lorenzi N, Nian H, Bailey J, Frisse M. Health information exchange usage in 
Customization of a Commercial CPOE System to Improve Patient Safety

Clare L. Tolley MPharm1, 2, Neil W. Watson MBA MSc 2, Andrew Heed2, Andrew K. Husband PhD MSc4, David W. Bates MD MSc3, 4, 5, Sarah P. Slight PhD MPharm PGDip1, 2, 3

1School of Medicine, Pharmacy and Health, Durham University, U.K.; 2Newcastle upon Tyne hospitals NHS Foundation Trust, U.K.; 3The Center for Patient Safety Research and Practice, Division of General Internal Medicine, Brigham and Women’s Hospital, Boston, MA, USA; 4Harvard Medical School, Boston, MA, USA; 5Harvard School of Public Health, Boston, MA, USA.

Abstract: This study explored how a commercial CPOE system had been customized by both the organisation and users following implementation, and what benefits and challenges they faced. Certain work processes became more efficient and safer; however, some system design issues and a limited use of Clinical Decision Support (CDS) prevented the users from realizing its full potential.

Introduction: The Health Information Technology for Economic and Clinical Health (HITECH) Act encouraged widespread adoption of commercial Computerized Provider Order Entry (CPOE) in the U.S.1 Once implemented, these systems are often customized to improve their usability, patient safety and quality of care. The aim of this study was to explore how a world-leading, commercial CPOE system had been customized after its implementation in a large U.K. teaching hospital, and what the benefits and challenges were.

Methods: We obtained the necessary ethical and institutional approvals, and recruited staff across four adult wards (renal, cardiology, general medicine and surgery) in a U.K. teaching hospital and members of the implementation team. One researcher conducted 32 unique semi-structured interviews with doctors (n=16), nurses (n=8) and pharmacists (n=8) between Mar ‘15 and Aug ‘16, lasting between 17-70 minutes and performed 35 hours of ward-observations. Users were asked about their experiences of using the system and any customized features. All interviews were transcribed verbatim and the transcripts checked. These data were analysed using the framework approach, which supports an inductive approach to theory development, where there are pre-set aims and objectives. Qualitative data analysis software NVivo version 10 was used; a list of themes were developed inductively, using the ‘constant comparison’ technique and explanations for recurring patterns in these data were sought, discussed amongst the research team, refined and presented.

Results: Participants highlighted a number of key benefits and challenges with CPOE customization. 1) Some users changed the screen layout to improve the visibility of important information (e.g., a medication’s stop date). However, this came at a cost, as other information was now no longer visible; one nurse was concerned she was ‘missing drugs and doses’ (P4; Nurse). 2) The organization developed order sentences and order sets to improve the safety and efficiency of certain tasks, such as prescribing regimes that were error prone, which was considered “a really good way of preventing that error from happening again”(P30; Pharmacist). However, some users had difficulties remembering the ‘key trigger words’ to identify them, which limited their use. 3) Tailored alerts such as a warning issued to nurses if they were about to administer acetaminophen too early, improved the safety of the administration process. However, some users reported insufficient use of alert functionality and actively requested that more alerts e.g., drug-drug interactions for high-risk combinations be switched on (e.g., if multiple drugs that cause bleeding were prescribed). 4) The organisation modified the system to enable blood glucose levels to be reviewed and insulin to be prescribed electronically and reviewed from the same screen. However, if prescribing remotely, there was “a danger of just looking at a computer screen,(...) and going “oh right, yeah, the BM [blood glucose level] is high let’s give them something”, but you don’t actually know what’s going on’(P23; Pharmacist). 5) The organisation developed a pharmacy task list, which helped users efficiently target high-risk patients, but some warned it lacked the sensitivity to identify all cases and patients could be overlooked.

Conclusions: Users described several benefits and challenges in customizing this system and the effect on patient safety. Further research should focus on whether similar problems have emerged with other commercial systems and how to address these, and whether customisation had any measurable impact on patient outcomes.

References
A Random Forest Model for Detecting Somatic Mutations from Unpaired Cancer Samples

Sophie Tu, PhD¹, Bin Zhou, MS¹, Da Wei Huang, MD², Roland Schmitz, PhD², George W Wright, PhD², Louis M Staudt, MD, PhD², Calvin A Johnson, PhD¹

¹Center for Information Technology, National Institutes of Health, Bethesda, MD; ²National Cancer Institute, Bethesda, MD

Introduction

Due to the complexity of cancer genome rearrangement as well as sample impurity and sub-clonal mutations, regular SNV callers such as GATK and HaplotypeCaller, which rely on a ploidy assumption, do not work well on whole exome cancer samples. MuTect2¹, which was specifically developed for such a situation, albeit with paired-normal samples, is regarded as one of the most reliable cancer SNV callers². However, cancer genome studies without paired normal samples cannot take advantage of tools such as MuTect2. In this work, we hypothesize that a machine-learning model can be trained from MuTect2 scores on paired samples to detect somatic mutations from toxicity scores.

Methods and Results

We downloaded 37 cancer-normal pairs of diffuse large B-cell lymphoma (DLBCL) exon-seq datasets from The Cancer Genome Atlas (TCGA). The fastq sequencing files were first aligned to the human genome HG19 via BWA according to TCGA guidelines. MuTect2 was used to make both somatic and germline SNV calls. Those calls that passed statistical filters and fell into coding regions were added to the training and holdout sets. A random forest model was trained on a total of 4990 calls from these samples. The feature space comprised intrinsic data such as mutation rate, read depth, and duplication status as well as scores from 24 toxicity assessors (e.g. SIFT, PolyPhen2, etc.) Figure 1 compares the ROC of the random forest model with the sensitivity and specificity scores of the individual toxicity callers. Other classification methods similarly demonstrated the benefit of aggregating toxicity scores as a feature-space representation in discriminating between somatic and germline mutations.

Conclusion

As our results suggest that the random forest model can achieve a high degree of agreement with MuTect2, we believe that this classifier can be a useful tool to filter SNV lists derived from regular mutation callers on sequencing data from cancer samples when matched normal samples are not available.

References

A Decision Support Tool to Help Providers Address Perinatal Depression

Bengisu Tulu, PhD¹, Nicolas Adami-Sampson¹, Trivani Shahi¹, Nicholas Shannon¹, Leif Waugh¹, Drew Wethern¹, Nancy Byatt, DO², Tiffany A. Moore Simas, MD², Kathleen Biebel, PhD²

¹Worcester Polytechnic Institute, Worcester, MA; ²University of Massachusetts Medical School, Worcester, MA

Introduction

Perinatal depression, the most common complication related to childbirth, affects approximately 1 in 7 pregnant and postpartum women and is often under-recognized.¹ Currently, obstetric providers face challenges in diagnosing perinatal depression efficiently and effectively partially due to lack of access to validated decision support tools for diagnosis. We developed a mobile app that guides obstetric providers through the process of screening and assessing patients for perinatal depression, and identifying appropriate stepped treatment options for patients who screen positive.

Methods

Using Ionic open source framework and rapid application methodology, we developed a multi-platform (web, iOS, Android) app and evaluated its usability. After developing a paper prototype, we implemented the first prototype and conducted a usability study with seven participants who were obstetricians. During the interviews, participants were asked to use the app and comment on its features and ways the app can be improved. We developed a second prototype based on these comments and conducted a second usability study with 12 obstetricians in a focus group setting.

Results

The app is designed to provide evidence-based guidance to the obstetric providers based on the Edinburgh Postnatal Depression Scale (EPDS) score of a patient. After a user provides the EPDS score, the app navigates through the decision algorithms to determine the most appropriate course of action for the providers to suggest based on the severity range determined using the EPDS score and the evaluation conducted by the providers. We asked 12 participants in our focus group to spend some time using the app before they engaged in discussion and took a brief survey. Table 1 presents survey responses (5-point Likert Scale) we gathered from 12 participants.

Table 1. Usability Focus Group Survey Responses

<table>
<thead>
<tr>
<th>Questions</th>
<th>Mean</th>
<th>Std. Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>How valuable would this app be in clinical settings? (1-Not Valuable, 5- Very Valuable)</td>
<td>4.67</td>
<td>0.492</td>
</tr>
<tr>
<td>How useful would this app be to you? (1-Not Useful, 5- Very Useful)</td>
<td>4.75</td>
<td>0.622</td>
</tr>
<tr>
<td>How would you rate the app’s design? (1-Poorly designed, 5- Well Designed)</td>
<td>4.33</td>
<td>0.778</td>
</tr>
<tr>
<td>How often do you see patients for whom this app might be useful? (1-Almost Never, 5- Very Often)</td>
<td>3.75</td>
<td>0.866</td>
</tr>
</tbody>
</table>

Conclusion

Usability studies reported in this abstract showed that the app satisfies a need and the app can be useful in clinical settings. We received a number of requests to continue using the app after the focus groups. Given the demand for this decision support app, we are planning to evaluate the feasibility and effectiveness of using an easily accessible mobile health based pocket tool for perinatal depression decision support in clinical settings in our next study.

References

Creating RDF Data on Trauma Care Organizations from Questionnaires

Joseph R. Utecht, BA\textsuperscript{1}, Jonathan P. Bona, PhD\textsuperscript{1}, Mathias Brochhausen, PhD\textsuperscript{1}
\textsuperscript{1}University of Arkansas for Medical Sciences, Little Rock, AR, USA

Background

Assessment of trauma systems and trauma centers positively affects patient outcomes\textsuperscript{1,2}. However, assessment of these organizations often falls short because of a lack of comparability regarding organizational components and their implementation. The CAFE project (https://cafe-trauma.com) addresses this problem by enabling comparison of different roles (e.g. trauma medical director), the different components (e.g. multidisciplinary stakeholder group) and their requirements, privileges, and obligations, which we have found to be different from organization to organization. To allow the comparison of these properties in a web-based self-assessment environment we use semantic web technologies. To this end we develop a user-friendly questionnaire allowing respondents to enter data stored in RDF to compare organizational components in real time. This approach has two immediate benefits: a) we take advantage of the graph-based nature of RDF to easily produce graphical representations of organizational structures, and b) we use built-in automated reasoning in combination with OWL ontologies to classify and uncover differences between superficially identical organizational roles and components. A key ontology developed specifically for our applications is the Ontology for Organizational Structures of Trauma Centers and Trauma Systems\textsuperscript{3}.

Methods

To support our requirement of allowing user-friendly data entry, we developed a tool to create the appearance of a traditional web questionnaire. To prevent any interruption for the respondent while the server processes previous answers, we chose to build the interface as a separate web application in Angular2 (https://angular.io). To provide semantically-rich data, the server-side component will not only record the answer in a relational database, but simultaneously create a pre-configured RDF (https://www.w3.org/RDF) representation of the answer in a triplestore.

Results

The result of this work is available on Github (https://github.com/cafe-trauma/documentation). Fig. 1 shows the steps of acquiring RDF data through a questionnaire built using our tool. The first step is the creation of the questions, for example “Does your trauma center have a trauma registrar?” Next the administrative user defines the RDF representation for the answers, in this case, the representation of the trauma registrar, the trauma center, and the fact that the trauma registrar is an organizational member of that trauma center. The tool then builds a traditional web questionnaire with a yes/no question “Does your trauma center have a trauma registrar?” As respondents answer this question the RDF representation is created in a triplestore and their answer is recorded in a relational database. From the respondent’s point of view the questionnaire does not differ in any way from a typical web questionnaire.

Discussion

Once the CAFE project is finished our aim is to use the ontology and the data created to link it to outcome data. The methodology presented in this paper is not restricted to the domain of the CAFE project, but is applicable to multiple domains of biomedical research and medicine. In addition to trauma care, we have also successfully implemented questionnaires collecting data on potential drug-drug interactions using this methodology.

Acknowledgements

The research presented in this paper is funded by the National Institute of General Medical Sciences of the National Institutes of Health under award number 1R01GM111324.

References

Systematic Evaluation of Smartphone Applications on Spinal Cord Injury Self-Management

Gary Vargas¹, Spyros Kitsiou, PhD², Ben Gerber, MD., MPH¹
¹College of Medicine ²Department of Biomedical & Health Information Sciences, University of Illinois at Chicago, Chicago, IL 60612

Introduction

Smartphone applications (apps) have been gaining popularity in lifestyle and chronic disease self-management. However, little information is available on apps that address chronic spinal cord injury (SCI) management and prevention of complications. This study aims to conduct a content analysis of the functional characteristics, consumer interaction features, and usability of mHealth apps currently available in the Google Play store for self-management of chronic complications of spinal cord injury.

Methods

A comprehensive search was performed between October 22 and 29, 2016, in the Google Play store to identify SCI-related apps using relevant keywords (e.g. spinal cord injury, paraplegia, and autonomic dysreflexia). Inclusion criteria included apps that were free, provided education, symptom monitoring, or self-management of aspects relevant to SCI and were available in English or Spanish. Apps that were provider-focused, aimed at phone accessibility features, or were limited to only users with app-specific access rights were excluded. One author (GV) systematically assessed apps using an evaluation tool that incorporated multiple content and usability criteria. PubMed and clinicaltrials.gov were also queried to investigate whether any of the identified apps have been evaluated in pilot studies. iOS apps were not included due to time and resource constraints.

Results

Out of the 702 search results, twenty-eight apps met the inclusion criteria and were evaluated. 18% (n=5) of apps were specifically targeted at SCI pts. 14% (n=4) and 68% (n=19) were aimed at patients with disabilities and the general public, respectively, but had functionality that overlapped with SCI patients. Of the apps evaluated, 32% (n=9) provided SCI education, 57% (n=16) symptom monitoring options, and 54% (n=15) supported SCI self-management regimens. Multiple apps were dual-focused. The most common SCI topic included bowel and bladder education and management. Topics rarely addressed were respiratory, sexuality/fertility, pressure ulcer, and travel/transportation management. Only 3 education-focused apps cited organizations or scientific studies as references. Common monitoring features included bowel and bladder management (32%, n=9; 32%, n=9, respectively), followed by medication (18%, n=5), pain (14%, n=4), and blood pressure monitoring (14%, n=4). Forty percent (n=11) of apps included reminders and alarms for scheduled medications and bowel/bladder programs. 11% (n=3) also included treatment strategies for autonomic dysreflexia. Usability varied widely among all apps. Seven percent (n=2) of apps crashed at least once during testing. 71% (n=20) of the apps used images to support educational information or display user monitoring trends. Twenty-two percent (n=6) included graphs/images that lacked labeling of axes/lines or had minimal contextual information to support them. No associated pilot studies were found for any of the apps evaluated.

Conclusion

There are few Android apps designed specifically for SCI patients available. Also, there is much variability in usability, educational content, symptom monitoring, and self-management options in those that do exist. As no pilot studies were found for the apps evaluated, it remains unknown whether these apps can improve self-management for chronic SCI patients or support caregivers’ efforts in SCI management. Given that 20% of yearly recurring costs for chronic SCI patients represents home health regimens, and that preventable conditions such as urinary and skin complications contribute to other costly emergency hospitalizations, these self-management apps may present an additional means of improving self-management regimens, preventing complications and decreasing costs. Future work must evaluate the iOS and web apps available to SCI patients.

References


2212
Representation and Validation of Reference Content Sources Within a Commercially Available Knowledge Management System

Christopher Vitale PharmD,1 Priyaranjan Tokachichu MD,1 Daniel Kim,1 Margarita Sordo PhD,1,2,3 Saverio Maviglia MD, MS,1,2,3 Roberto A. Rocha MD, PhD1,2,3
1Clinical Informatics, Partners eCare, Partners Healthcare System, 2Brigham and Women’s Hospital, 3Harvard Medical School, Boston, MA

Introduction
The effective integration of reference terminology sources (e.g., SNOMED CT, LOINC) within knowledge and terminology engineering processes is a critical requirement to enable the creation and maintenance of interoperable knowledge assets [1]. At Partners, we implemented a multi-step process to normalize, import, and validate reference sources using a commercially available knowledge management system (KMS) [2]. Once integrated, reference terminology concepts can then be easily linked to local knowledge artifacts, allowing not only streamlined authoring, but also effective management of dependencies as sources evolve over time.

Methods
We developed a semi-automated process that (a) verifies the integrity of a new reference content release file, (b) identifies content changes by comparing to a prior release, (c) classifies identified changes as new, revision, or deprecated, (d) creates a new code system revision to bundle all identified changes within the new release, and (e) transforms changed content into the normalized meta-schema used by the KMS. At the end, this process generates an XML file that is used to import new release content into the KMS. The KMS implements a well-defined asset lifecycle and custom validation rules to manage dependencies between code system revisions. Gaps and inconsistencies are automatically verified during import based on preconfigured rules, generating error logs that guide improvements. Customized portions of the process are required for each content type due to source files that use non-standard formats, as well as to target specific previously identified error types.

Results
We have successfully integrated multiple reference terminology releases using the KMS tool (see Table 1). The integration of a new source includes an initial modeling effort to faithfully represent the reference content using the KMS meta-schema, plus the configuration of the software tools used during the semi-automated process. Once the first release of a given reference source is successfully imported into the KMS, subsequent releases require less effort, unless errors are identified. We have identified and reported over 4200 errors in various releases.

Table 1: Reference terminologies integrated with the knowledge management system (KMS)

<table>
<thead>
<tr>
<th>References source</th>
<th>Concepts loaded to KMS</th>
<th>Errors identified</th>
<th>Error type</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNOMED-CT US Jan 2014</td>
<td>2.9M</td>
<td>-</td>
<td>N/A</td>
</tr>
<tr>
<td>SNOMED-CT US Sept 2015</td>
<td>2.9M</td>
<td>-</td>
<td>N/A</td>
</tr>
<tr>
<td>SNOMED-CT US Sept 2016</td>
<td>2.8M</td>
<td>-</td>
<td>N/A</td>
</tr>
<tr>
<td>LOINC 2.54</td>
<td>84K</td>
<td>-</td>
<td>N/A</td>
</tr>
<tr>
<td>LOINC 2.56</td>
<td>79K</td>
<td>35</td>
<td>Incorrect property value syntax</td>
</tr>
<tr>
<td>LOINC 2.58</td>
<td>74K</td>
<td>4215</td>
<td>Inconsistent concept metadata</td>
</tr>
</tbody>
</table>

Conclusion
Integration of reference sources with local knowledge artifacts is necessary to support interoperability and knowledge sharing. Although the representation of reference terminologies within a KMS is a continuous and time-consuming process typically performed by content vendors, doing so locally allows for detection and resolution of content errors via customized validation rules, as well as prevention of future errors through ongoing dependency management. Implementing such processes can take an experienced clinical informatician weeks to months, but once in place takes minutes to execute. Efforts to normalize and validate reference sources would be greatly simplified if they were distributed using standard formats, such as Common Terminology Services 2 (CTS2) [3].

References
Toward Integrated Memory Assessment in Primary Care

Lisa M. Vizer, PhD¹, Daniel Kaufer, MD¹
¹University of North Carolina, Chapel Hill, NC

Introduction
Alzheimer’s disease (AD) and neurocognitive disorders (NCD) present a daunting economic and psychosocial challenge to families, healthcare systems, and society. We are seeing a rapid growth in those affected—estimated at more than 5 million in the US—coupled with a lack of therapeutic interventions. To achieve therapeutic success with NCD, patients’ cognitive decline must be identified early. Although cognitive screening for early detections is required during the Medicare Wellness Visit (MWV), we lack a standard protocol. Thus, primary care clinicians face a requirement with unclear standards, methods, and billing procedures, leading to inconsistency and inadequate documentation across the healthcare system. Additionally, scant resources are available to primary care physicians once they identify patients needing follow-up. Primary care physicians are currently unable to bill for cognitive testing as part of the annual MWV, and may not have access to guidance on interpretation of test results. Thus, we describe a proposal for combining cognitive screening and follow-up assessment for people with suspected NCD.

Proposed Solution
The senior author and colleagues conducted a randomized controlled trial evaluating an educational intervention on memory disorders care for primary care providers (Lathren et al., 2015). Data from this study show that while primary care providers could efficiently integrate cognitive screening into their practice flow, referring patients for a comprehensive memory disorders evaluation remained problematic. Ongoing work led by the senior author investigating a tablet-based cognitive testing protocol has also identified that many individuals who test positive on a memory disorder screen are not aware that they have a problem. To address these problems we propose a system for cognitive screening and triage along with a billable follow-up neurocognitive assessment protocol. This procedure is designed to support the annual MWV in primary care settings and provide remote consultative assistance to primary care physicians regarding need for follow-up. This approach will help distinguish AD and non-AD clinical profiles using a tablet-based screener; positive screens for possible AD will then return to the primary care setting for a remote-access, computer-based neurocognitive assessment similar to what is typically conducted in a memory disorders clinic. The tablet-based and computer-based testing will incorporate standard assessments adapted to administration in primary care along with novel assessments designed to provide a more holistic approach to cognitive screening. The multidimensional nature of the screening will help identify specific non-AD symptom profiles to prompt referral to a memory clinic and facilitate identification of extenuating factors contributing to cognitive complaints (e.g. sensory impairment, depression).

If screening indicates, the system will prompt a computer-based neurocognitive assessment that will meet Medicare guidelines for a billable procedure. A primary care clinic can use the new G0505 code for “Cognitive and functional status exam” or a clinic visit code for the primary care venue and a procedure billing code for the memory care clinic remote-testing service. Ultimately, we envision physician extenders (e.g. nurse practitioners) collaborating on test supervision and follow-up planning in the primary care setting, with test scoring, interpretation, and feedback recommendations coming from the memory clinic. This approach, utilizing primary care providers as “first-responders” in memory care, could readily be scaled up to provide coverage for an entire healthcare system, and in turn, provide a platform for implementing and disseminating “best practices” and research advances in the field.

Conclusion
Shifting from the current “passive referral” model for all suspected NCD patients to a focused “active triage” model through a billable remote assessment integrated into primary care may positively disrupt screening and follow-up care for neurocognitive disorders. Refining the standard assessments to give finer-grained insight into extenuating factors such as mood and sleep disturbances, differentiate amongst dementias, and incorporate feedback from a memory care nurse practitioner enables more support for primary care providers, improved care for patients, and more efficient use of memory care resources. This system is also scalable throughout the health system and facilitates rapid institution of cutting edge research and new treatment protocols.

Reference
A Diabetes Management Platform to Reduce Severe Hypoglycemia in Older Adults with Type 1 Diabetes

Jonathan S. Wald, MD, MPH1, Elena Toschi, MD2,5, Garry M. Steil, PhD3,5, Howard Wolpert, MD2,5, Christine Slyne, BA2, Tara MacNeil, BA2, Astrid Atakov-Castillo, BA2, Stephanie Edwards, MPH2, Jennifer R. Webb, MA1, Alexa M. Ortiz, MSN, RN1, Maria Alva, PhD1, Robert D. Furberg, PhD1, Medha Munshi, MD2,4,5

1RTI International, Raleigh, NC; 2Joslin Diabetes Center, Boston, MA; 3Boston Children’s Hospital, Boston, MA; 4Beth Israel Deaconess Medical Center, Boston, MA; 5Harvard Medical School, Boston, MA

Introduction

Older adults (65+ years) with longstanding type 1 diabetes (T1D) have more frequent and severe hypoglycemia for many reasons including comorbid conditions that impact glucose variability; changes in mobility, vision, dexterity, and cognitive ability that impact problem-solving skills and self-management; hypoglycemia unawareness; and reduced social support. Munshi1 found communication with an educator by telephone between visits to be important when working with older T1D adults. Weinstock2 found hypoglycemic unawareness and glucose variability to be associated with low blood sugar, and called for more research to assess interventions for severe hypoglycemia.

Technology such as continuous glucose monitoring (CGM) found to help younger individuals may be particularly useful in older adults experiencing high glucose variability and hypoglycemic unawareness. However, older adults may experience barriers adopting the technology. In this poster we describe the diabetes management platform (DMP) and research study we designed to address these barriers. We are currently enrolling participants into a randomized controlled trial to determine if the DMP reduces hypoglycemia in older adults with T1D.

Platform

The DMP (Figure 1) incorporates glucose, insulin, and activity monitoring, direct patient feedback, and automated monitoring of continuous data to enhance provider decision-making and feedback to the patient and caregiver.

Figure 1. Diabetes Management Platform: 1) Tablet with preloaded software collects data from blood glucose (BG) meter, CGM, insulin pump or pen, and activity tracker; 2) patient having hypoglycemia enters contextual information about hypoglycemic event; 3) Clinical decision support (CDS) system analyzes DMP data to suggest insulin dose adjustments for clinician review; 4) Geriatric-specific education and clinician messages sent to the patient.

Conclusion

This poster describes the DMP components, study approach, and preliminary findings concerning barriers and enablers based on interviews with early participants. Acknowledgement: This work is supported by NIH/NIDDK grant DP3DK112214 and is approved by the Joslin institutional review board; registration with Clinicaltrials.gov is pending.

References

Novel Pedigree Analysis Allows for Genetic Heterogeneity to Aid in Gene Discovery for Common, Complex Diseases

Rosalie G. Waller, HBS1, Todd M. Darlington, PhD1, Myke Madsen, MS1, Robert J. Klein, PhD2, Steven M. Lipkin, MD, PhD3, Celine M. Vachon, PhD4, Nicola J. Camp, PhD1

1University of Utah School of Medicine, Salt Lake, UT; 2Icahn School of Medicine at Mount Sinai, New York, NY; 3Weill Cornell Medical College, New York, NY; 4Mayo Clinic, Rochester, MN

Introduction

Together, common, complex diseases are the leading cause of morbidity, yet the genetic factors for these conditions are least understood. Missing heritability exists for the majority of complex diseases, demonstrating genetic factors contribute significantly to risk. Thus, understanding the genetic factors involved in common, complex disease is important for treatment decisions in precision medicine. Rare, risk variants have been suggested as a source for missing heritability; however, identifying these variants remains a challenge due to genetic heterogeneity and complex inheritance patterns. Localizing inherited, chromosomal regions to target the search for rare-risk variants is instrumental to discovering genetic factors involved in complex disease.

Methods

We developed a gene mapping strategy, based on the Shared Genome Segment (SGS) method, that utilizes high-risk pedigrees (HRPs) sufficiently large to singularly identify likely-inherited, chromosomal segments. The strategy addresses intra-familial, genetic heterogeneity by optimizing over all possible subsets of cases in a HRP. Key to the strategy is the derivation of significance thresholds for segment interpretation. These thresholds address the genome-wide search and the multiple testing, inherent from the optimization, through use of distribution fitting and the Theory of Large Deviations. We applied this new SGS strategy to 11 extended, Utah, multiple myeloma (MM) HRPs, and subsequent whole-exome sequencing (WES) in SGS regions of interest in 1064 MM/MGUS (monoclonal gammopathy of undetermined significance—a precursor to MM) cases and 964 controls from a jointly-called collaborative resource, including cases from the initial 11 HRPs and from an additional 57 families. MM is a complex cancer of plasma cells with 30,330 new cases annually. Thus far, no inherited risk variants have been identified for MM.

Results

Application of our new SGS strategy to the 11, MM HRPs identified a genome-wide significant, 1.8 megabase (Mb), shared segment at 6q16 in one HRP. WES in this region revealed a stop gain variant and a missense variant in USP45, a gene known to influence DNA repair through the nucleotide-excision repair (NER) complex, in two families. Additionally, a 1.2 Mb, shared segment at 1p36.11 is inherited in two Utah HRPs. WES in this region revealed 2 missense variants in ARID1A, a key gene in the SWI/SNF chromatin remodeling complex, in one of the HRPs sharing the segment and in another MM/MGUS family. We extended our WES analysis to the complexes of USP45 and ARID1A. In the analysis for the NER complex, we identified a genome-wide suggestive, 0.8 Mb, shared segment at 19q13 containing 2 NER genes, and ClinVar pathogenic variants in 2 additional NER genes in early-onset, sporadic MM cases. In the analysis for the SWI/SNF chromatin remodeling complex, we identified a marginal, genome-wide suggestive, 1.5 Mb, shared segment at 3p21 containing 1 SWI/SNF gene. Finally, burden testing was significant for 7 of the 22 NER complex genes and 8 of the 15 SWI/SNF chromatin remodeling complex genes.

Discussion

Our results provide compelling, segregating risk genes and variants for MM and demonstrate a novel strategy to use large HRPs for risk-variant discovery more generally in complex disease. Specifically, our results support a role for NER and chromatin remodeling in MM risk. Prior literature suggests both complexes are essential to cancer prevention and are mutated in multiple malignancies. Our HRP analysis strategy can be mutually beneficial in a collaborative setting containing both extended HRPs and smaller families. Post-hoc, additional value could be gained from demographic and/or clinical data on the sharing subsets shedding light on other shared characteristics that may aid future mapping. In sum, our strategy addresses gene-discovery challenges in common, complex diseases and aids in risk-gene mapping to inform precision medicine therapies.
Visual Exploration of Metabolic Networks to Provide Context for Metabolic Profiling in Precision Medicine

T. Cameron Waller, B.S.\textsuperscript{1,2}, Jared Rutter, Ph.D.\textsuperscript{1,3}, Alexander Lex, Ph.D.\textsuperscript{2,4}

\textsuperscript{1}Department of Biochemistry, University of Utah School of Medicine, Salt Lake City, Utah; \textsuperscript{2}Scientific Computing and Imaging Institute, University of Utah, Salt Lake City, Utah; \textsuperscript{3}Howard Hughes Medical Institute, University of Utah School of Medicine, Salt Lake City, Utah; \textsuperscript{4}School of Computing, University of Utah, Salt Lake City, Utah

Metabolism is an integral part of the biological systems of cells, tissues, and organisms. It is a cooperative and continuous system with tens of thousands of individual genes, transcripts, proteins, and metabolites. Catabolic processes degrade food to liberate energy and materials while anabolic processes harness this energy and utilize these materials to synthesize and assemble all cellular components. In these ways, metabolism sustains life and all of its diverse processes. It is little wonder that metabolic defects contribute to the pathology of many human diseases, including obesity, diabetes, cardiovascular disease, and cancer.

In their study of metabolism, investigators need tools to access the vast complexity of the metabolic system. As current interest in metabolism is increasingly interdisciplinary, this accessibility is a challenge for novices and experts alike. Scientists and engineers from diverse backgrounds consider metabolic phenotypes and mechanisms, and metabolic profiling in the clinic is an important aspect of precision medicine. Another challenge is the shift in experiments from reductionist to systemic scales. Modern technologies, especially the omics technologies (genomics, transcriptomics, proteomics, and metabolomics), measure biological entities at nearly comprehensive scales. Both the reliable design of these experiments and the accurate interpretation of their results requires consideration of context, including an holistic perspective of the entire metabolic system.

Current analyses risk over-simplifying metabolism and neglecting its real context. Most available tools represent metabolism as a collection of separate pathways, such as Glycolysis, Pentose Phosphate, and the Citrate Cycle. While these pathways are conceptually simple, they neglect metabolism’s extensive connectivity and continuity across pathways. Experimental results from omics technologies are frequently relevant to multiple pathways. Identifying relations between these is laborious, and there is a great risk of arbitrarily or subjectively ignoring relevant parts of the system.

We aim to provide a versatile tool for the objective and dynamic exploration of metabolic networks. We adopt a common model for these networks—with reactions as relations between metabolites as entities—and derive information from reconstructions of systems biology. As this network for human metabolism is large (roughly 13000 nodes and 43000 links), our tool supports dynamic, custom queries to select portions of the network that match the user’s interest. Queries consider both attributes and topology of the network. To query by attributes, a user might select metabolites and reactions in a specific metabolic pathway or cellular compartment. To query by topology, a user might select metabolites and reactions either within a specific proximity to a single metabolite or within shortest simple paths between multiple metabolites. We find that specific queries with multiple criteria are necessary to select subnetworks that are sufficiently simple for exploration. We are developing a visual interface to facilitate the construction of these queries and to communicate relevant information about the network concisely and clearly. We hope that our tool will promote basic discovery and translational interpretation to support metabolic profiling in precision medicine.
Matching Consumer Health Vocabulary with Professional Medical Terms Through Concept Embedding

Yue Wang, MS1, Jian Tang, PhD2, V.G. Vinod Vydiswaran, PhD3, 4, Kai Zheng, PhD5, Hua Xu, PhD5, Qiaozhu Mei, PhD2, 1
1Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI; 2School of Information, University of Michigan, Ann Arbor, MI; 3Department of Learning Health Sciences, University of Michigan, Ann Arbor, MI; 4Department of Informatics, University of California, Irvine, CA; 5School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX

Introduction: Matching consumer health vocabulary with professional medical terms is an important task for medical information retrieval, knowledge sharing, and effective health communication. Previous approaches studied mining consumer health vocabularies (CHV) with hand-crafted heuristics on medical search logs1 and community-generated content2. Though the professional-consumer concept pairs mined through these approaches are very accurate, many relevant pairs could be missing. In other words, the recall is low. In this pilot study, we propose to match professional-consumer concept pairs through text embedding approaches. Text embedding approaches have proven to be very effective in capturing the similarity between words and phrases3, which can yield high recall of professional-consumer concept pairs. We first learn the representation of the medical concepts with a large amount of unlabeled text. Afterwards, the professional-consumer concept pairs are matched according to the similarities of their embeddings.

Methods: We first learn word embeddings from large-scale text corpus and then represent a medical concept as the average of word vectors it contains. Three corpora are used to train word embeddings: 2.2 million MedHelp posts (consumer-generated text), 7 million Medline abstracts (professional-generated text), and 4.7 million Wikipedia articles (general-domain text). We use the Skip-gram model1 to learn word embeddings and calculate the similarities between concepts as the cosine similarity of their word embeddings. The results are evaluated using a ranking-based metric, the mean reciprocal rank (MRR). Given a professional (consumer) concept in a CHV pair, we rank candidate medical concepts by cosine similarity to the query. Ideally, the corresponding concept in the pair is ranked high.

We use a large collection of candidate medical concepts and ground truth professional-consumer concept pairs. We collect all English medical concepts in the UMLS as the concept collection and use two existing CHVs as the ground truth pairs: CHV-1 from Vydiswaran et al.2 and CHV-2 from Zeng et al.1 To ensure fair comparison between word embeddings learned on different corpus, we filter out concepts that contain out-of-vocabulary words for any corpus. This gives us 494K medical concepts from UMLS, 892 concept pairs from CHV-1 and 124K concept pairs from CHV-2. We randomly select 1000 pairs from CHV-2 to make the sizes of CHV-1 and CHV-2 comparable for evaluation.

Results and Discussion: Our initial results show that it is feasible to identify alternative medical concepts by using professional or consumer concepts as queries in the concept embedding space. The MRR values indicate that on average, a professional or consumer concept is about 14th closest to its counterpart in CHV-1 and about 7th closest in CHV-2. Note that this is only the result on a small subset of the large medical concept collection where we have ground truth. We will generate an extended list of related concept pairs and pseudo-label them as professional- or consumer-oriented using the propensity measure2. This list can then be checked by medical experts and expand current CHVs. This will open up the opportunity to build more extensive CHVs with minimal effort from practitioners.

To make the concept embedding specialize in professional-consumer vocabulary translation and continuously improve itself, we are developing a semi-supervised representation learning algorithm. The algorithm uses a small set of known CHV pairs as the supervision signal, and learns word embeddings (and the way to compose word embeddings into a concept embedding) such that consumer concepts and professional counterparts are close to each other. With more and more supervision pairs, the recommended alternative concepts will become more precise.

References
Utility of Anti-hypertension Prescription Orders in Predicting Future Hypertensive Instability Events

Xuan Wang, MS1, Lujie Chen, MS1,2, Marilyn Hravnak, PhD3, Gilles Clermont, MD4, Michael R. Pinsky, MD4, Artur Dubrawski, PhD1

1School of Computer Science and 2Heinz College, Carnegie Mellon University; School of Nursing3 and School of Medicine4, University of Pittsburgh, Pittsburgh, PA

Introduction

Accurate prediction of the cardiorespiratory instability (CRI) risk in monitored stepdown unit (SDU) patients may prepare clinical teams to anticipate CRI more accurately and respond sooner to improve outcome. While previous work has explored methods for non-specific CRI risk prediction using static information available at SDU admission time, we hypothesize that provider’s prescriptions accruing dynamically over the admission stay (specifically for anti-hypertension medication [anti-HTN]) may have utility in predicting specific CRI events (hypertension [HTN-CRI]), if we view the prescription as a proxy of the dynamic information collected from multiple sources which might not otherwise be captured or leveraged.

Methods

Noninvasive vital signs (VS) were collected at 1/20Hz frequency over 15 months from bedside monitors in a 24-bed SDU, including heart rate (HR), respiratory rate (RR) and peripheral oximetry (SpO2); systolic (SBP) and diastolic (DBP) blood pressure measurements were intermittently recorded. From Electronic Health Records, we extracted providers’ prescription orders during SDU stays, patients’ medical history for hypertension, and demographics. We identified 382 SDU events (cases) during which patients developed at least one instance of HTN-CRI (VS exceeded SBP >200 or DBP>110mmHg) and 2074 controls without any HTN-CRI. We recorded whether or not these patients had anti-HTN medications prescribed prior to their initial HTN-CRI during their SDU stays. We excluded anti-HTN prescribed within the hour before HTN-CRI to reduce the bias from concomitant reactivity, e.g. anti-HTN prescribed from HTN-CRI events at sub-threshold levels.

Results

56% of HTN-CRI cases and 38% of controls had an anti-HTN prescribed (p-value <0.001). We evaluated the utility of anti-HTN orders predicting HTN-CRI risk by constructing models incorporating potential confounding covariates. Table 1 shows the output from three logistic regression models when covariates are sequentially added: starting age on admission, then HTN history on admission, and finally accruing anti-HTN orders over admission. All coefficients for covariates of interest are positive and statistically significant (p-values in parentheses). In addition, the likelihood ratio tests (last column) comparing subsequently enriched models show improvement at each step, supporting our hypothesis that accruing anti-HTN prescriptions carries additional information in predicting future HTN-CRI events over information available at time of SDU admission alone.

Table 1. Output from logistic regression models with factors sequentially added.

<table>
<thead>
<tr>
<th>Model</th>
<th>Age</th>
<th>HTN History</th>
<th>ANTI-HTN</th>
<th>Likelihood Ratio Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) HTN-CRI ~ age</td>
<td>0.020 (&lt;0.001)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(2) HTN-CRI ~ age + HTN history</td>
<td>0.014 (&lt;0.001)</td>
<td>0.556 (&lt;0.001)</td>
<td></td>
<td>Model 2 vs Model 1 p-value &lt;0.001</td>
</tr>
<tr>
<td>(3) HTN-CRI ~ age + HTN history + ANTI-HTN</td>
<td>0.013 (&lt;0.001)</td>
<td>0.439 (&lt;0.001)</td>
<td>0.551(&lt;0.001)</td>
<td>Model 3 vs Model 2 P-value &lt;0.001</td>
</tr>
</tbody>
</table>

Conclusion

Assuming anti-HTN is a surrogate for pre-existing hypertension, this analysis demonstrates the potential utility of leveraging the history of concomitant cardiovascular modulating therapies as an identifier in predicting future HTN-CRI events. Thus, dynamic information encoded in providers’ prescription orders may be used to update the baseline prediction of relevant CRI events using only historical information. Future work will explore prescription related markers for predicting other CRI events such as tachycardia and hypotension. We envision this approach will open avenues for building models that could provide more actionable and specific information of future CRI events, helping facilitate the personalized monitoring paradigm.
Semantic Representations of Medical Terms: A Comparison Study of Word Embeddings Trained on Clinical Notes and PubMed Articles

Yanshan Wang, PhD\(^1\), Naveed Afzal, PhD\(^1\), Liwei Wang, PhD\(^1\), Feichen Shen, PhD\(^1\), Hongfang Liu, PhD\(^1\)

\(^1\)Division of Biomedical Statistics and Informatics, Department of Health Sciences Research, Mayo Clinic, Rochester, MN

Introduction

With the availability of large quantities of health data, deep learning has become more and more prevalent. The word embeddings learned by deep neural networks have seen tremendous success in various natural language processing (NLP) applications. Word embeddings trained on common domain data have been shown to carry syntactic and semantic meanings of words. Training word embeddings usually requires an adequately large corpus. In medical domain, clinical notes and biomedical articles are two resources that could serve as training data. In this study, we answer the question: Do the word embeddings trained by these two resources perform differently on representing semantics of medical terms?

Methods

In our studies, we utilized the word2vec model, skip-gram, to evaluate the word embeddings. Word2vec is based on the assumption that words occurring in similar contexts tend to have similar meanings. Based on a training corpus, it uses a deep neural network to embed words into a continuous vector space. In our implementation, we set the minimum number of times a word must appear in the corpus to 7, the window size to 5 words, and the number of dimensions in feature space to 60.

Two corpora were utilized to train the word embeddings: EHR Corpus and PubMed Corpus. EHR Corpus contains clinical notes for a cohort of 113k patients receiving their primary care at Mayo Clinic, spanning a period of 15 years (1998-2013). PubMed Corpus contains 1.25 million articles from PubMed Central. Minimum pre-processing was conducted including punctuation removal, digits replacement, etc. We also compared the trained word embeddings with two public pre-trained embeddings: Google News and Wikipedia.

Four datasets were utilized to evaluate the word embeddings for capturing medical term semantics. Dataset 1 contains 30 medical term pairs with manually generated similarity scores; Dataset 2 contains 34 medical term pairs with similarity scores; Data 3 consists of 101 clinical term pairs whose similarity was determined by physicians from the Mayo Clinic; Data 4 consists of 725 clinical term pairs whose similarity was determined by residents from the University of Minnesota Medical School.

On each dataset, the word embeddings were evaluated based on their Pearson correlation coefficient with the similarity scores determined by human experts.

Results

The results of Pearson correlation coefficient on four datasets are listed in Table 1. Overall, the similarity results using word embeddings trained on PubMed Corpus are the closest to human experts’ results. However, there is no statistical significance between the word embeddings trained on EHR Corpus and those on PubMed Corpus. Both word embeddings are significantly superior to the pre-trained word embeddings on Wikipedia and Google News.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>EHR Corpus</th>
<th>PubMed Corpus</th>
<th>Wikipedia</th>
<th>Google News</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.542</td>
<td>0.569</td>
<td>0.334</td>
<td>0.357</td>
</tr>
<tr>
<td>2</td>
<td>0.416</td>
<td>0.311</td>
<td>0.159</td>
<td>0.243</td>
</tr>
<tr>
<td>3</td>
<td>0.295</td>
<td>0.300</td>
<td>0.001</td>
<td>0.084</td>
</tr>
<tr>
<td>4</td>
<td>0.374</td>
<td>0.404</td>
<td>0.190</td>
<td>0.154</td>
</tr>
</tbody>
</table>

Table 1. Pearson correlation coefficient of four datasets.

Conclusion

There is no statistical significance between the word embeddings trained on EHR Corpus and PubMed Corpus in regards of semantic representations of medical terms. However, the word embeddings trained on medical domain data are superior to those trained on common domain data.
Neural Network Word Embeddings for Text Classification of MeSH terms.

Walter Wang, MSc, Yindalon Aphinyanaphongs, MD, PhD
NYU Langone School of Medicine, New York City, New York

Introduction
In recent years, neural network and deep learning approaches have redefined benchmarks in vision, natural language, classification, and prediction. A common biomedical task is assigning MeSH terms to MEDLINE documents. In this work, we apply two popular neural network classifiers\(^1,2\) to the benchmark OHSUMED dataset, a collection of 50,000 Medline abstracts, to a 23 label classification task established in prior work\(^3\). One of the key contributions of the neural network approach is their ability to generate unsupervised semantically relevant word embeddings. For every token, an \(N\) dimensional vector is created such that similar words appear close together in the Euclidian space\(^4\). This work evaluates \(1\) the usefulness of semantic word embeddings for classification performance on a benchmark task and \(2\) the value of general vs domain specific embeddings for a task.

Methods
For pretraining, we tested three different sets of word representations. The first set are word vectors built using the skip-gram model\(^5\) on the full OHSUMED dataset. The second set is a pretrained word vector dataset from Google News\(^6\). The third set is a pretrained word2vec dataset from Pubmed and Pubmed Central articles\(^6\).

Using FastText for text classification and a Tensorflow implemented convolutional neural network\(^7\), we report the recall-precision breakeven point. The loss function is an approximation of the sigmoid cross entropy with logits optimized for the multi-label classification task.

\[
\text{loss} = \max(x, 0) - x \ast z + \log(1 + e^{z}), \quad x = \text{logits, } z = \text{labels}
\]

Results
The reported result is an average of the breakeven points across the 23 label benchmark classification tasks.

<table>
<thead>
<tr>
<th></th>
<th>No pretraining</th>
<th>Skip-gram (OHSUMED)</th>
<th>Word2vec (Google)</th>
<th>Word2vec (Pubmed)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Support Vector Machine</td>
<td>66(^3)</td>
<td>----</td>
<td>---</td>
<td>----</td>
</tr>
<tr>
<td>FastText (unigram)</td>
<td>58</td>
<td>.70</td>
<td>697</td>
<td>703</td>
</tr>
<tr>
<td>FastText (bigram)</td>
<td>62</td>
<td>.736</td>
<td>702</td>
<td>708</td>
</tr>
<tr>
<td>ConvNet</td>
<td>64</td>
<td>.718</td>
<td>698</td>
<td>715</td>
</tr>
</tbody>
</table>

Discussion
Word vector based pretraining are useful and produce significant improvements in classification. Both types of pretrained neural networks outperform Support Vector Machines. The specific embeddings from the actual dataset surprisingly perform better than general purpose word embeddings either produced from Google or from Pubmed.

References:
6. Moen S, Ananiadou TSS. Distributional semantics resources for biomedical text processing [Internet]. LBM; 2013.
Come Together: a multi-hospital transition in reporting strategy from untethered databases to integrated warehouse solutions

Emily C. Webber, MD¹,², Phillip S. Cadle³, Cara L. Willey RN, BSN³, Michael Schwarz MBA³, Rebecca Rose, MD¹
¹Indiana University School of Medicine, Indianapolis, IN; ²Regenstrief Institute, Indianapolis, IN; ³Indiana University Health, Indianapolis, IN

Introduction
Access and management of clinical data is an important part of leveraging the electronic health record (EHR) for operational reporting, quality improvement efforts and research. In a large health care organization, standards for operational reports is important for accuracy and longitudinal outcome tracking. The objective of this project was to evaluate and reconcile several sources of operational and clinical reports for a multi-hospital neonatology group into the health care organization data warehouse.

Methods
Evaluation of reports. Prior to the project, clinical data was manually transcribed from the clinical EHR into a neonatal database housed outside of the enterprise data warehouse. The database was found to have approximately 400 custom reports, the majority of which were operational reporting, quality reporting and one time reports (frequently to determine a cohort of patients with a specific profile). Despite this large number of reports, only about 12 reports were routinely obtained by different facilities, and slightly different sources (physician, nursing or EHR numbers) were used by each facility (Table 1). In addition to the primary neonatal database, different locations used variable sources for standard reporting and there were additional manual steps to reconcile gaps in standard reports.

EHR and Data Warehouse design. Using the previous database and clinical pathways as a guide, a standard form for targeted clinical assessments and quality elements was implemented in the EHR. The form contained approximately 300 unique clinical data fields; the fields as well as other selected clinical data and diagnoses is loaded daily into a custom data mart in the enterprise data warehouse. Curating a data mart of neonatology patients allowed for strategic mapping to other databases of hospital data and use of analytical resources and tools already adopted across the health care organization. Analysts and neonatologists created web-based dashboards which present a catalog of standard reports as well as elements for creating custom queries in the clinical data in the data mart.

Results
Since implementation, facilities utilize the same reports leading to both collaborative validation and improved ease of shared objectives. Use of the reports has led to improvements in the EHR form. Additional routine reports have been identified and redundant reporting processes reduced. Clinicians and other members of the medical team utilize validated query tools to run shared reports allowing for comparisons between facilities.

Conclusions
By working directly with clinicians, warehouse and informatics analysts were able to curate a high impact and clinically relevant data mart, which can adapt to include new elements as needed. By integrating into the shared data warehouse, the clinical teams at different facilities now have improved data integrity and depth.
Resident Physician Efficiency in the Electronic Medical Record (EMR): sustaining improvements
Emily C. Webber, MD1,2, Ellen Klink, MD1, Andrew O’Brien, MD1, Courtney Hines3
1Indiana University School of Medicine, Indianapolis, IN; 2Regenstrief Institute, Indianapolis, IN, 3Indiana University Health, Indianapolis, IN

Introduction: Resident physicians comprise a large portion of the physician workforce utilizing the EMR in many health centers and have a different workflow from non-resident physicians. The objective of the project was to redesign the resident EMR position to promote efficiency in finding information in the chart as well as perception in ease of use.

Methods: The project involved a redesign of the resident EMR view with a focus group, EMR process reports (click counts and time in different parts of the chart) and a survey before and after the change.

- Update of resident physician view. A multidisciplinary group of residents redesigned the resident EMR view. This was done by analyzing the clicks for buttons or links, removing and reordering the remaining links, and adding the ability to change between views based on the specialty needed that month.

- Evaluation of EMR process reports. The vendor EMR utilization analytics tool evaluated process outcomes for all residents in all settings for 5 months prior to the change and 6 months after including: the number of times they needed to navigate between different sections of the chart to find data (“tab hops”), time spent in chart review and time required to find a patient (patient discovery).

- Survey assessment. A voluntary, anonymous electronic survey was conducted pre- and post- the EMR change.

Results: The resident EMR view was changed for 1564 residents in May 2016 (Figure 1, month 5). Tab hops, chart review and patient discovery time all decreased after the change. The tab hop control chart (Figure 1) revealed the average number of tab hops/patient declined by 2.03/patient. Upper control limit (UCL) and LCL (lower control limits) seen below. Patient discovery declined by 6 seconds/patient, and chart search time by 23 seconds/patient on average.

Residents responding to the survey (171 pre-EMR change, 198 approximately post-EMR change) reported more often being able to find information, as well as improvements in productivity and workflow (Table 1).

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>I can find information in the EMR</td>
<td>3.75</td>
<td>3.86</td>
<td>0.11</td>
</tr>
<tr>
<td>The EMR helps me be productive</td>
<td>3.26</td>
<td>3.42</td>
<td>0.16</td>
</tr>
<tr>
<td>The EMR is consistent with my workflow</td>
<td>2.99</td>
<td>3.69</td>
<td>0.7</td>
</tr>
</tbody>
</table>

Table 1. Resident survey responses pre and post the EMR change.

Conclusions: Early results show a trend in the decrease in the overall amount of time by residents spent searching in the EMR as well as discrete navigation (ie. tab hops). Additional months of data collection are required to evaluate whether this represents a more permanent shift in these measures. The survey of perceptions reflected positive changes (ie. residents responded that the EMR performed searching and other functions more consistently in the post survey). Both objective measures of time as well as perceptions of efficiency are needed to evaluate the success and value of optimization in the EMR. Although promising, these metrics should continue to be evaluated to see if the time savings are sustained.
Measures of Improvement: Combining Satisfaction Survey and Time Measures to Identify Impactful Optimization

Emily C. Webber, MD1,2, Wanita Kumar, MD1, Jeff Aldrich3, Cara Willey, RN3, Zach Gordon, MD1,2, Erika Green, MD2, Matthias Kochmann, MD2, Dan Seitz, MD1,2, and Jason Schaffer, MD1
1Indiana University School of Medicine, Indianapolis, IN; 2Regenstrief Institute, Indianapolis, IN; 3Indiana University Health, Indianapolis, IN

Background Electronic health record (EHR) implementation has been widespread over the last decade; however, there is little published data regarding evaluation of continuous EHR improvements. Some EHRs can provide data from embedded tools to measure time spent on ordering, number of clicks and other process measures. Surveys and interviews with users provide important context regarding perceived value and satisfaction of these measures. The objective of this project was to compare EHR measures of time and actions with survey responses pre- and post- EHR enhancements. This would allow us to determine if improvements in time savings correlated with perceptions of efficiency and higher satisfaction.

Methods A project team of physicians and clinical informatics analysts devised a strategy for using the vendor EHR analytic database and survey data to evaluate the effectiveness of updated EHR functionality.

Vendor analytic database The vendor EHR database was used to track key performance indicators such as total time in chart per patient, time for chart review, ordering and documentation and navigation before and after the EHR updates. The database is administered by the EHR; it analyzes active time and activities in the chart by applying a proprietary analysis of mouse and keyboard input. The data can be filtered by provider position and other elements.

Survey A five question, web-based, IRB-approved survey was administered to all providers who received the updated functionality. Questions were modelled on validated end-user computing satisfaction1. All providers were provided a link to a pre-survey for two weeks prior to the updated EHR functionality and a post-survey 1 month after the update. Results were reported on a Likert scale with 1 indicating least satisfied and 5 indicating most satisfied.

Results A summary of the efficacy of the implementation for each provider group using both the EHR process metric and survey was provided (Fig 1). The documentation time per patient (graph on left) was determined by averaging the total time spent in the documentation tasks per patient/per number of providers in the group. Further filtering by individual provider was also provided.

Figure 1. Sample report including EHR vendor process metric and survey results.

Results 1659 providers in 12 different service lines received updated functionality in their EHR, with approximately 10% responding to the pre and post survey. All 12 service lines showed decreased times for chart search, order and documentation in the EHR database report. On the survey, 10 out of the 12 service lines increased scores on all five questions on their post-survey. Two service lines had lower satisfaction responses on their post-survey.

Conclusions The survey responses for the majority of the providers reflected higher satisfaction on all five questions; however, two groups were less satisfied despite objective measures of improved efficiency (as reflected by EHR database reports of process outcomes). Process measures like time metrics are insufficient as a sole determinant of the effectiveness and value of improvements in EHR usability. Evaluations are more comprehensive if they include assessments like surveys to reflected perceived value.

References
Using Statistical Models to Predict Senior Patients’ Use of Smartphone Apps for Managing Parkinson’s Disease – A Preliminary Study from A Service Learning Project

Duo (Helen) Wei
Computer Science and Information Systems – BUSN
Stockton University, Galloway, NJ 08205

Introduction
Approximately four million people worldwide suffer from Parkinson’s disease and nearly one million people in the U.S. are living with Parkinson’s disease (PD), with between 50,000 and 60,000 new cases diagnosed each year, according to the National Parkinson’s Foundation. PD is incurable but medications and appropriate interventions can control symptoms for many years. With the advent of internet and mobile phone technologies, people are increasingly interested in using Smartphone technology to manage their health. Although senior people are those groups that are in urgent need of help, due to the limitation of the resources, personal health status, and the lack of technical background, seldom do they expose themselves to the application of technology to manage their health. In this project, we aim to investigate: (1) what are the major factors that influence the senior PD patient’s inclination to use Smartphone apps for managing their health? (2) how to predict a PD patients’ inclination of using Smartphone apps?

Method
Students in Medical Informatics class, taught by the author (DW), organized a workshop for senior patients with Parkinson’s disease focusing on health management through technology in the spring of 2016. The workshop was held at the Capital Health Medical Center as a service learning project for the course. Student utilized knowledge acquired during the class to introduce one of the most popular Apps called “Parkinson’s Central Toolkit” including informational apps, medication and symptom management, with the goal of helping those struggling with PD increase their knowledge of resources that they had in an effort to promote a better quality of life. After conducting the workshop, surveys were distributed, which were designed to learn patients’ demographic information and their habits of using technology. Based on the survey results, we extracted five most important factors: gender, age, the perception of the severity of PD patients, operating systems (Android vs. iOS), and daily used ratios. We first used the logistic regression model to see each factor against their inclinations of using the Smartphone app. Then we applied the multiple regression model to predict the use of Smartphone app based on the above mentioned five factors.

Results
Twenty PD patients were involved in the workshop and among which 18 patients turned in the survey. There were fourteen completed surveys and the four incomplete surveys (<50% complete) were not included in the statistical analysis. Among the fourteen patients, there were eight males and six females. The survey, containing 15 mandatory questions and five optional questions, was designed to use paper-based format and was distributed after the workshop. In the survey, patients were asked questions about their demographic information, their perceptions of the severity of the disease, their level of technology backgrounds, and their uses of the Smartphone. Logistic regressions and multiple regressions have been applied for the modeling. According to the statistical model, there were no statistically significant association between the response variable (the inclination to uses) and the terms (such as age, gender, and the severity of the disease). By applying the goodness-of-fit tests, it is determined that the predicted probabilities deviate from the observed probabilities in a way that the binomial distribution does not predict (p-value >= 0.05). The multiple regression models show that there is a statistically significant association between the inclination of usage and patient’s gender. That being said, male patients are more likely to choose to use Smartphone app that their female counterparts.

Discussion and Conclusion
Smartphone apps lead to a revolution in patient health management. In this research, PD patients are informed of the technological tools available to them and ways to utilize the apps. Statistical models are employed to analyze and predict patients’ usage of Smartphone apps. Results show that male patients have higher inclination of Smartphone app usage than female, which is partially due to the fact that females tend to be less interested in technology than male according to early research. Therefore, educators should involve more female in STEM on educational level. Besides, Smartphone app developer may consider to develop more female friendly products. A larger sample set is needed to get a more precise result and a better-fitted model. We are going to organize two workshops to engage more patients in the community. In the future research, we are going to investigate not only PD patients but patients with other diseases. The research will also focus on which Smartphone app features are most important to a particular patient community.

2225
Using a Convolutional Neural Network model for Automatic Medical Subject Headings (MeSH) Assignment

Wei Wei, PhD, Zhanglong Ji, PhD, Lucila Ohno-Machado, MD, PhD
University of California, San Diego, La Jolla, CA, USA

Introduction
Medical Subject Headings (MeSH) is a controlled vocabulary thesaurus used primarily for indexing biomedical publications. Assigning MeSH terms enables efficient retrieval of biomedical publications via PubMed. Most assignment methods use manually extracted features, including the state-of-art model. Inspired by the performance of the Convolutional Neural Network (CNN) model for feature extraction in the text mining research, we studied the contribution of automatically generated features from CNN for the MeSH term assignment task.

Methods
Kim’s CNN model was adopted to generate distributed representations (i.e., CNN features) from abstracts and MeSH term candidates. The model by Huang et al. served as the baseline. The baseline model includes three features (i.e., additional features) that do not require extra training data:

- neighborhood feature: the number of similar articles in which a MeSH term candidate appears,
- overlap feature: number of unigrams/bigrams overlapping between MeSH term candidate and the abstract, and
- translation feature: the probability of translating the abstract into the MeSH term candidate.

The neighborhood feature has a dominant impact on the performance (see Table 1, row 2), while the other features are valuable supplements. We analyzed the contribution of CNN features alone, together with the neighborhood feature, and with all the three features. The performance was evaluated on NLM2007, a National Library of Medicine (NLM) benchmarking dataset, and measured by Mean Average Precision (MAP), which is the average precision values at ranks of relevant items (i.e., MeSH terms). The scripts and the configuration of CNN are available from https://github.com/w2wei/ConvNet_MeSH.

Results
A combination of the CNN and additional features achieved good MAP in this dataset, comparable to just using the 11 manual features in the baseline model (see Table 1).

Table 1. The performance of CNN features alone and with the additional features.

<table>
<thead>
<tr>
<th>Features</th>
<th>MAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline (11 manually extracted features)</td>
<td>0.626</td>
</tr>
<tr>
<td>neighborhood</td>
<td>0.602</td>
</tr>
<tr>
<td>CNN</td>
<td>0.335</td>
</tr>
<tr>
<td>neighborhood + CNN</td>
<td>0.602</td>
</tr>
<tr>
<td>neighborhood + CNN + overlap + translation</td>
<td>0.646</td>
</tr>
</tbody>
</table>

Conclusion
The model built with CNN and the three features helped improve MAP in the NLM2007 data set. Neither the CNN features nor the selected additional features require extra training data.

References
Computing Performance Analysis on Clinical Document-level Classification
Wei-Hung Weng, MD1,2, Kavishwar B. Wagholikar, MBBS, PhD2,3, Henry C. Chueh, MD2,3
1Department of Biomedical Informatics, Harvard Medical School, Boston, MA;
2Laboratory of Computer Science, Massachusetts General Hospital, Boston, MA;
3Department of Medicine, Massachusetts General Hospital, Boston, MA.

Effective computing is critical for developing machine learning-based solutions. In particular, machine learning tasks related to the clinical decision often require real-time modeling and evaluation to provide timely treatment and management. Machine learning based-natural language processing (NLP) techniques allow researchers to construct automated document classifiers to categorize clinical documents based on their content and document-level information without manual processing. Nevertheless, the NLP approach usually creates a higher dimensional feature space, which takes an abundant time to train the classifier if the training corpus is large. There is also the lack of studies on performance analysis on clinical document classification tasks.

We performed document classification tasks based on the medical subdomain of the clinical document using two datasets -- 431 publicly available clinical notes and reports from the Integrating Data for Analysis, Anonymization, and Sharing (iDASH) data repository, as well as 91,237 Massachusetts General Hospital (MGH) clinical notes from the Research Patient Data Registry (RPDR) data repository of Partners HealthCare system. For each dataset, we applied seven clinical feature representation methods and five supervised learning algorithms to construct 35 classifiers. For all classifiers with different size of feature spaces (Table) and algorithms, we evaluated the model performance using performance score (PS), which was defined by 

$PS = \text{sigmoid}(\text{standardized } F_1 \text{ score}) \times [1 - \text{sigmoid}(\text{standardized processing time})]$. 

The standardization and sigmoid transformation were applied to $F_1$ score (classifier performance) and processing time (machine performance) to ensure that scales of $F_1$ score and processing time are consistent and comparable. For all classification tasks, we used the virtual machines on the Partners ERISOne Linux Cluster, with four CPU cores (Intel Xeon 3.47GHz) and 16GB RAM memory.

We found that using the classifiers trained by support vector machine (SVM) with linear kernel yielded the most effective computing performance across two datasets. The appropriate feature representation method, such as using medical concepts with semantic restriction, can additionally augment the performance. We found that the clinical document classification tasks based on the medical subdomain of the document could be done within the acceptable effectiveness, regarding model performance and speed in the scale of ninety thousand documents. We plan to execute and evaluate the same task on a larger analytical platform for scaling up.

<table>
<thead>
<tr>
<th>Feature Representation Method</th>
<th>iDASH</th>
<th>MGH</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bag-of-words (Vocabulary size)</td>
<td>10150</td>
<td>160097</td>
</tr>
<tr>
<td>UMLS concepts</td>
<td>4750</td>
<td>25456</td>
</tr>
<tr>
<td>UMLS concepts (5 semantic groups)</td>
<td>4531</td>
<td>24457</td>
</tr>
<tr>
<td>UMLS concepts (15 semantic types)</td>
<td>3634</td>
<td>18520</td>
</tr>
<tr>
<td>Bag-of-words + UMLS concepts</td>
<td>14900</td>
<td>185553</td>
</tr>
<tr>
<td>Bag-of-words + UMLS concepts (5 semantic groups)</td>
<td>14681</td>
<td>184554</td>
</tr>
<tr>
<td>Bag-of-words + UMLS concepts (15 semantic types)</td>
<td>13784</td>
<td>161949</td>
</tr>
</tbody>
</table>

Developing an Evaluation Plan for Nationwide Data Infrastructure Programs

S. West, PhD1; J. Amoozegar, MSPH1; M. Ashok, PhD1; J.E. Richardson, PhD, MS, MLIS1; B.H. Blumenfeld, MD, MS1; S. Lumsden, RN, MS2; C. Cravens, MPH2

1RTI International, Durham, NC
2HHS/Office of the Assistant Secretary for Policy and Evaluation, Washington, DC

Introduction

Between 2003 and 2010, Congress enacted three laws that authorized research and funding to expand the number of studies on the outcomes and effectiveness of treatments and interventions used in health care. Most recently, the Patient Protection and Affordable Care Act (ACA) of 2010 established a trust fund for patient-centered outcomes research (PCORTF), with a small portion (4%) for HHS to coordinate relevant federal health programs to build data capacity for research. Through a coordinated effort, federal agencies have a undertaken a portfolio of projects (N=30) focused on collecting, linking, and analyzing research data in a manner that supports research networks and the interoperability of health data. Evaluators from RTI International designed a multilevel evaluation to assess the impact of the work done on the portfolio of data infrastructure projects and to identify gaps for future development activities.

Methods

We developed a multilevel evaluation plan for PCORTF data infrastructure initiatives incorporating an HHS provided strategic framework. The framework contains five “functionalities”, which represent required core functions for enhancing data infrastructure for research: 1) use of clinical data for research; 2) standardized collection of standardized clinical data; 3) linkage of clinical and other data for research; 4) collection of participant-provided information; and 5) use of enhanced publicly funded data systems for research. Achieving these functionalities should make clinical data more usable for research. The five functionalities are organized by “components”: standards, services, and policies and governance (SSPG); these components are further broken down into more specific sub-components referred to as developmental components (DCs). The EP’s measurement strategy integrates the Evaluation Planning Incorporating Context (EPIC) model, a maturity schema, and input from a technical panel with program evaluation, informatics, and outcomes research expertise. The maturity schema guides data collection and analysis, with an emphasis on mapping project goals and activities to DCs, gauging their level of achievement and assessing their status within and across projects. In aggregate, the EP measures the degree to which each data infrastructure project and the entire project portfolio, enable the five functionalities.

Results

The EP is driven by three overarching evaluation questions: 1) What specific contributions did each project towards creating clinical data infrastructure for conducting research? 2) How has the project portfolio enabled the functionalities to address the collection, linkage, and analysis of data? and 3) How have the funded projects (to date) built data capacity and advanced researchers’ ability to capture, store, access, link, exchange, and analyze data securely and efficiently? The maturity schema delineates the increasing levels of maturity (signifying a progressively increasing degree of sophistication) required for DCs to enable functionalities from a process perspective. We determined which levels of achievement were required to fully operationalize each functionality. To determine how well the functionalities are supported by levels of achievement across the portfolio, we are reviewing project documents to assess project activity status. This will allow us to determine progress at each level of achievement for each functionality’s DCs. We will compile findings across all projects before Fall 2017.

Discussion

The primary challenge we noted in conducting the EP is the need for additional information beyond what exists in the project’s statements of work and progress reports from the federal agencies overseeing the projects to HHS, especially for projects in progress at the time of evaluation. We have addressed some of these informational gaps by conducting project-specific literature searches, which we will supplement with calls to project directors. The key anticipated challenge in executing the plan is providing results in an actionable way that allows us to assess the comprehensiveness of the project portfolio and identify gaps in the infrastructure for collecting, linking, and analyzing electronic data for research. The evaluators are using a heat map model to graphically depict levels of maturity in an intuitive way to present results to stakeholders.

2228
Project A-G-T-C: Adding Genetic Test data in Clinical data warehouse

Jonathan Wickes; Hyeoneui Kim, RN, MPH, PhD; Hyun-Dae Kim; Jihoon Kim, MS; Wenhong Zu, PhD; Lucila Ohno-Machado, MD, MBA, PhD; Olivier Harismendy, PhD

Department of Biomedical Informatics, University of California, San Diego, La Jolla, CA

Introduction and Background
An increasing number of patients are now undergoing molecular testing as part of their clinical care and those test data should be made available for patient care and patient centered outcome research. However at UCSD Medical Center, the data is currently deposited in a highly heterogeneous format in the Electronic Health records (EHR). For example, the results of several tests performed by the UCSD laboratory are deposited into the laboratory test results table in a structured format while those of many tests performed by third party vendors are upload as a PDF report. Not surprisingly, the vendor produced test results have not been incorporated into UCSD’s Clinical Data Warehouse (CDW). Therefore, none of useful queries, which can be as simple as identifying which patient underwent genetic testing and as refined as knowing the patients who carry a specific mutation in a specific gene, were possible neither in EHR nor in CDW. To address this challenge, we started developing the procedure that parses and standardizes vendor produced genetic test results to make them available in the CDW. As a feasibility study, we processed the test results from a genetic test vendor, Foundation Medicine (FM), and incorporated them into the CDW.

Methods
Upon IRB approval, we obtained genetic test results produced by FM for the past 4.5 years (08/2012-08/2016) in XML (before PDF rendering) as well as in PDF. This includes 1,669 reports of 1,594 patients. By reviewing several reports in PDF, we identified the fields that hold meaningful information thus need to be included in the CDW. Next we wrote an XML parser that takes the XML formatted test results and extracts those fields with meaningful information. Following the recommendation of the Clinical Genomics Working Group of HL7 we standardized gene names with HUGO Gene Nomenclature Committee (HGNC) standard; variant names with the variant description nomenclature of Human Genome Variation Society (HGVS); variant name, variant types (inferred based on the results) and specimen types with Logical Observation Identifier, Names, and Codes (LOINC); and specimen collection sites with Systematized Nomenclature of Medicine - Clinical Terms (SNOMED-CT).

Results
The overall workflow is summarized in Figure 1. A new Genetic Test Results table (Box 1) was added to CDW and linked to the Patient table using a medical record number. Six dictionary tables were implemented to accommodate the standardized code sets. Majority of the test results (93%) were successfully standardized with the data standards listed above. We created our own local code sets as an interim solution to incorporate a few variant types reported by FM that were not included in the LOINC by the time we completed this study. The fidelity of this implementation was validated with 3 data query use cases: (1) number of patients with TP53 nonsense tumor mutation, (2) number of male and female with any tumor mutation in PIK3CA, and (3) number of patients with cancer carrying a non-VUS NF1 tumor mutation.

Discussion
We successfully processed the genetic test results from a third party vendor laboratory Foundation Medicine and imported them into UCSD’s CDW in a standardized manner. This feasibility study showed that the current data standards provide quite good coverage for encoding the important information on the genetic test data. Standard development for genetic test information is an on-going effort thus we anticipate that the revision of the current mapping will be necessary. As this feasibility testing was done using the data from a single vendor, the generalizability of the resulting process and pipeline is not guaranteed. We plan to refine our approach by applying it to the test results from various vendor laboratories.

Box 1. Data fields in the genetic test result table
(* coded value)

- Medical Record Number
- Report ID
- Report Date
- Test Name
- Laboratory Name
- Gene Name*
- Variant Name*
- Variant Type*
- VUS
- Is Equivocal
- Is Subclonal
- Medical Dx (ICD9CM)*
- Specimen Types*
- Specimen Collection Site*
- Specimen Collection Date

Figure 1 Overall workflow
Secure Text Messaging to Improve Workflow Efficiency and Provider Satisfaction in Administration of Radiation Therapy

Aaron Wild, MD1, Erik Anderson, MD, PhD1, Natalie Lockney, MD1, Suchit Patel, MD, PhD1, Robert Samstein MD, PhD1, Amy Xu, MD, PhD1, Keri Martin2, Aaron Ostrovsky, MPH2, Chhavi Kumar, MD3, Daniel Stein, MD, PhD2

1Department of Radiation Oncology; 2Division of Health Informatics; 3Department of Medicine; Memorial Sloan Kettering Cancer Center, New York, NY

Introduction
A key to delivering high quality, high dose radiation treatments is optimal patient set-up on the treatment machine on each day of treatment. A secure, two-way text messaging (STTM) system was implemented for use between the radiation therapy technologists (RTTs) at treatment machines, and attending physicians, who previously relied on one-way text paging, email, and phone. Previous studies of app-based communication systems in other clinical domains have shown improvements in clinician efficacy and satisfaction1-2. Our objectives were to determine if this also would apply in Radiation Oncology by: 1) reducing the time from patient setup films to radiation treatment and 2) improving provider satisfaction with the communication system between RTTs and attending physicians.

Methods
The system used was the Voalte platform (Voalte, Inc., Sarasota, FL). A smartphone-based app was installed on the attendings’ personal phones, and a web-based client was installed on the RTTs’ workstations. The participants could send and receive text messages in real time. The software also displayed the users’ roles, availability, and message status. During the two-month study period, STTM was the preferred method of communication for: (1) notifying attendings that patient setup films were ready for review and (2) notifying RTTs that the films had been reviewed and signed by the attending physician, permitting radiation treatment to proceed. We recorded length of time from films acquisition to approval for 20 treatments before and 20 treatments after the new STTM system was instituted. Mean times were compared between pre- and post-intervention using a two-tailed paired t-test for normally distributed data. To evaluate satisfaction, we conducted pre- and post-intervention surveys of 20 RTTs. Survey response scores were statistically assessed using the Wilcoxon rank sum test for datasets without normal distribution.

Results
Mean time from patient setup films to signing of films by the attending physician was not reduced by the STTM system at 4.8 mins (SD, 3.1) pre-intervention vs. 8.2 mins (SD, 5.2) post-intervention. There was a trend towards reduction in time after implementation of the STTM system when films were signed before/after normal business hours by a non-service attending at 7.2 mins (SD 3.7) pre-intervention vs. 4.7 mins (SD, 1.4) post-intervention (p=0.15). Survey data indicated that when compared with standard communication between RTTs and attending physicians, the implemented STTM system was more effective in: (1) reducing the need to be in a specific hospital location to initiate or receive communication (p=0.015); (2) reducing the time between sending a message and receiving the desired response (p<0.001); (3) allowing for both sending and receiving communication (p<0.001); (4) allowing for ease of delegation of tasks to colleagues (p<0.001); (5) decreasing dependence on telephone calls (p<0.001); and providing general satisfaction with communication tools (p=0.002).

Conclusion
Based on the study results, we conclude that implementation of the STTM system provides a valuable means of significantly enhanced communication between RTTs at treatment machines and attending physicians. We expect that as familiarity with the new STTM system increases over time, objective improvements in workflow efficiency and treatment times will become manifest. This study adds to the literature on the utility of secure text messaging and we will continue to explore the ways that app-based communication can enhance clinical workflow and care.

References
Data Cleaning for Peace Corps’ Population Health Surveillance

Thomas M. Wilkinson, MD, Rennie W. Ferguson, MHS, Daniel Murphy, Adrianna Evans, Susan J. Henderson, MD, MPH; United States Peace Corps, 1111 20th St. NW, Washington DC 20526.

The problem: Using the Peace Corps electronic medical record (EMR) system, Peace Corps Medical Officers (PCMOs) have an opportunity to select ‘Epi codes’ to add epidemiological context about clinical findings. PCMOs are in the best position to provide accurate data for population health surveillance, since they provide direct care to Peace Corps Volunteers (PCVs) and are aware of the clinical context; however, given other demands on PCMO time, epidemiologic data collection can become a lower-priority task. In a survey conducted by the Centers for Disease Control and Prevention in 2016, a large number of PCMOs (41%) reported that it was ‘somewhat difficult’ to find an appropriate Epi code in the EMR. The Peace Corps Office of Health Services (OHS) Epidemiology and Surveillance Unit (Epi Unit) extracts epidemiological data from the EMR on a monthly basis, and the Epi codes are reviewed for accuracy based on manual chart review. Given the limitations on PCMO time and constraints within the current release of the EMR, we estimate that approximately one-third of appropriate Epi codes (28-40%) were not recorded between January and November 2016. Correcting this deficiency during data cleaning places an untenable burden on the Epi Unit and likely yields incomplete and possibly inaccurate datasets.

The solution: The primary data error we are seeking to correct is the absence of an appropriate Epi code. The code represents contextual information surrounding a patient encounter, it is only occasionally predictable from a related diagnosis code, and its absence constitutes a type of error that is difficult to detect or correct through automation. We designed a number of interventions to improve the initial capture of Epi codes by PCMOs at the point of care, but our focus was to support the data cleaning process itself. We decided to use simple methods to leverage existing unstructured data that PCMOs were already entering into the EMR, and to combine that with structured and semi-structured data from elsewhere in the chart into a single pipelined process for data cleaning.

1. Pre-analyze, annotate, and store textual data from sources such as History of Present Illness or consultation discussion threads, using regular expressions to match keywords, their spelling variants, and synonyms.
2. Assemble comprehensive sets of related data which all stem from a given patient encounter, including the textual data, any associated scanned documents, and structured data elements from the chart, such lab orders, the problem list, and the medication list.
3. Using Microsoft Access, create a data cleaning console, and present the datasets in a single stream for manual review, with highlighted keywords and URL links to supporting documents according to which Epi codes are being reviewed at the time.
4. Finally, record the human cleaner's decision as to whether a presented dataset in fact satisfies criteria for adding the Epi code in question, or not, so as to accumulate a corpus of text, documents, and data elements that can be used in the future as a reference set for machine learning.

The above process produces a moderately high recall with a moderately low precision, and was designed to simplify the manual review process, rather than try to replace it. Epi codes can be added or removed directly, based on a pre-assembled and annotated dataset, minimizing the need for detailed, time-intensive chart reviews. Looking ahead, we will use the accumulated corpus of datasets and Epi code decisions to build dynamic, context-sensitive, suggested coding for the PCMOs.

Exploration of Contextual Digital Data as a Source of Patient Safety Insight

Dana M. Womack, MS, RN, Michelle R. Hribar, Ph.D., Paul N. Gorman, MD
Oregon Health & Science University, Portland, OR

Abstract
Although analysis of operational data has led to safety improvements in non-clinical domains, patient safety research in informatics traditionally prioritizes electronic health record data over contextual workplace data. Preliminary findings from an exploratory analysis of contextual data include verification that these data reflect daily rhythms of care and may contain actionable insights. Current research seeks to identify hidden patterns in contextual digital data that may serve as early indicators of patient safety risk.

Introduction
Elimination of preventable harm in healthcare settings is an urgent priority. Although high reliability organizations “recognize that the earliest indicators of threats to organizational performance typically appear in small changes in the organization’s operations”1, hospitals currently lack real-time visibility into the moment-to-moment activities of clinical staff and patients. Transactional records and log files produced by workplace tools such as medication dispensing, nurse call, and communications systems contain temporal information about granular patient care activities. These contextual digital data are not included in the electronic health record, however, they may serve as a source of data in healthcare improvement research and expand the types of data used in future early warning systems.

Exploratory Analysis of Contextual Digital Data
Although contextual digital data is imperfect, “when data about workers and work environments come to exist in sufficient volume, and are increasingly ubiquitous, there is the potential for significant changes not only in the work we study, but also in how we study it and the depth of understanding that can be achieved”2. To explore this possibility in a healthcare setting, data for a 30-day period were extracted from five hospital operational systems. Over 40,000 time-stamped events were collected with an overall mean of 60 events per hour. Locked door entry was most frequent at 17.5 events/hour overall, followed by staff communication (16.1/hour) and medication dispensing (13.2/hour).

![Figure 1. Medication dispensing by hour of day and day of week reveals twice daily peak administration times](image)

Figure 1. Medication dispensing by hour of day and day of week reveals twice daily peak administration times

Exploratory data analysis confirms that contextual digital data reflects recognized daily rhythms, such as higher activity during daytime hours than nighttime hours. Peak medication administration times mirrored the hospital’s twice daily medication schedule (Figure 1). Bed exit alarms were seen in the exploratory dataset at irregular intervals. As an example of future application to patient safety research, bed exit alarm data could contain patterns, that when combined with additional contextual data, may inform the development of proactive interventions to reduce patient falls. Evaluation of relationships between contextual activity patterns and safety risk is the next step in this research.

Conclusion
The use of contextual digital data in safety research is new to the field of clinical informatics. Preliminary exploration has confirmed that operational data reflects daily rhythms. Current research seeks to find hidden patterns as early indicators of safety risk. Through exploration of relationships between temporal activity patterns and patient safety, this study will lay a foundation for proactive monitoring of latent safety conditions in the hospital setting.

References
Factors Affecting the Decision of Heart Failure Patients to Accept Telehealth Services in the Home: An Integrative Review

Kyungmi Woo, MA, BSN, RN, Ph.D.(c) 1, and Dawn Dowding, Ph.D., RN, FAAN 1, 2
1Columbia University School of Nursing, New York, NY, USA; 2Center for Home Care Policy & Research, Visiting Nurse Service of New York, New York, NY, USA

Abstract

Telehealth is an effective intervention for helping heart failure patients manage their symptoms at home. However, the refusal rate for telehealth services among eligible patients is reported to be between 24-70%. The aim of this review was to identify factors that may affect patients’ decisions to adopt or refuse telehealth services. Factors identified included concerns over equipment or technology, concerns over service change, ease of use, access to care, knowledge of telehealth and its benefits, cost, and privacy.

Introduction

Heart failure (HF) affects about 5.7 million adults in the USA and related deaths increased from 81.4 per 100,000 in 2012 to 84.0 in 2014.1 2 Telehealth (TH) is one intervention that can assist patients with heart failure manage their symptoms at home and has been shown to reduce all-cause 30-day readmissions from 19.3% to 5.2%.3 However, it has been reported that between 24-70% of eligible patients refuse telehealth.4 The purpose of this integrative review was to identify factors affecting HF patients’ decision making to accept or refuse TH services in the home.

Method

Electronic databases (Medline, CINAHL, Cochrane Library, Embase, Scopus, Web of Science, OpenGrey) were searched using both medical subject headings (MeSH terms) and key words. Studies were included if they reported original data related to the acceptance of TH among HF patients in home care. The initial search yielded 208 articles; after screening of titles, abstracts and full text 5 studies were included in the final review. All studies were appraised for quality using the Mixed Methods Appraisal Tool (MMAT).

Results

Four out of the five included studies reported positive patient attitudes toward the use of TH. Refusal rates reported in the included articles ranged from 29% to 37%. Several factors were found to be related to decision-making regarding TH acceptance. These include: concerns over equipment or technology, concerns over service change, ease of use, access to care, knowledge of telehealth and its benefits, cost, and privacy.

Conclusion

Results of the review identified specific patient concerns about the use of TH and factors that could boost HF patient’s adoption of TH. Overall general perceptions toward TH are positive in this population. Building upon these positive views, healthcare providers can create and implement practices that further promote the use of TH in HF patients.

References

1. DHDSP. Heart Failure Fact Sheet. Division for Heart Disease and Stroke Prevention: Centers for Disease Control and Prevention; 2016.
A Phenome-Wide Association Study (PheWAS) of Iron Deficiency in a Large Electronic Health Record Database

Patrick Wu, BS1; Joshua C. Denny, MD1,2; Wei-Qi Wei, MD, PhD1
1 Department of Biomedical Informatics, Vanderbilt University School of Medicine, Nashville, TN 37203
2 Department of Medicine, Vanderbilt University School of Medicine, Nashville, TN 37232

Introduction: The phenome-wide association study (PheWAS) method was originally developed to discover phenotype associations for target genetic variants, but its utilization has been further extended to find links between laboratory data and phenotypes1. A recent study suggested an association between iron-deficiency anemia & hearing loss2. We sought to replicate their findings and to explore the breadth of biological insults associated with iron deficiency using PheWAS.

Methods: We conducted a PheWAS analysis using data from the Electronic Health Record (EHR) at Vanderbilt University Medical Center (VUMC). We have identified 1,862 cases (male: Total Iron Binding Capacity (TIBC) >450 µg/dL & Ferritin <24 ng/ml; female: TIBC >450 µg/dL & Ferritin <15 ng/ml) and 30,277 controls (male: TIBC = 250-450 µg/dL & Ferritin = 24-336 ng/ml; female: TIBC = 250-450 µg/dL & Ferritin = 15-204 ng/ml) from the cohort that met our medical home criteria3. We performed the analysis using the R PheWAS package4.

Results: In this study, we found highly significant associations between iron deficiency with anemias and digestive disorders. Further, in addition to replicating the finding of combined hearing loss in males, we observed a new clinical association between iron deficiency with depression in female individuals.

Conclusion: In this PheWAS analysis, iron deficiency was clinically associated with hearing loss and depression in male and female individuals, respectively.

References:
**Toward Reading Comprehension of Online Health Information: An Initial Annotation**

Danny T. Y. Wu, PhD, MS¹, Seung-Hyun B. Ko¹, Chris A. Feak, PhD²

¹University of Cincinnati Department of Biomedical Informatics, Cincinnati, OH; ²University of Michigan English Language Institute, Ann Arbor, MI

**Introduction**

Online health information is widely accessible and can provide patients and others with medical knowledge. It is also beneficial in facilitating patient-provider communications. Despite these expected benefits, the full potential of online health information cannot be achieved without individuals being able to read the text and comprehend the content. This highlights the importance of having a reliable means of ensuring that online health information is produced in a highly comprehensive manner to patients and others.

While computational methods for assessing the readability of online health information has been developed and widely used, such as the Flesch-Kincaid Ease Formula, very few exist that accurately and reliably quantify the elements contributing to reading comprehension of online health information. The goal of this study is therefore to develop such a computational method and improve the comprehensibility of online health information in the long run. The research team utilized a scientific writing technique called “old-new” information flow¹ to assess the comprehensibility of online health information. This is consistent with research in Applied Linguistics showing that when a text is written such that the information in a sentence proceeds from given or familiar to new, this is a strong indicator of text comprehensibility.

**Method**

MedlinePlus health topics were selected as a highly readable online health information based on our previous research². Since this corpus is claimed to be carefully maintained “in language that laypeople can understand”, it is an ideal source to examine how old-new information is followed. We hypothesized that the Medline Plus health topics strictly follow this progression of information to achieve their high level of reading comprehension. To facilitate the annotation process, we extracted medical concepts from the health topics using MetaMap and presented the concepts along with the original text using the brat rapid annotation tool (http://brat.nlplab.org/).

**Results**

A total of 993 Medline Plus health topics were downloaded from MedlinePlus on Nov 2016 as one XML file. Each health topic and its attributes were parsed using the Python BeautifulSoup library and stored in a SQLite database. Through the brat rapid annotation tool, two native English-speaking college students are able to highlight medical concepts in each sentence and link the concepts between sentences to indicate the old-new information flow. Moreover, human annotated concepts can be compared with those identified by MetaMap to explore which semantic types of medical concepts play a significant role in information flow (Figure 1). A simple measure based on the number of linkages between sentences has been developed to assess the validity of this approach. This annotation task will be extended to other online health information, such as ClinicalTrials.gov and PubMed abstracts, to ensure the effectiveness of the computational method.

**Conclusion**

This manuscript presents the initial effort to collect human annotations in preparation for the development of a computational method capable of quantifying the old-new information flow of online health information, which has implications on generating writing suggestions to improve the comprehensibility of online health information.

**Reference**

Classification method of dental electronic health record data potentially improves precision treatment

Di Wu, PhD
Department of Biostatistics, Department of Periodontology, University of North Carolina, Chapel Hill, North Carolina 27599

Abstract: The statistical method Latent (Hidden) Class Analysis, derived from the available dental DARIC study, was used in the local electronic health records collected in periodontal clinics at UNC, for novel classification of dental patients and teeth to subtypes, which results suggest different treatment effects for patients in the different newly classified dental clinical subtypes.

Introduction: The rationale for creating new personalized dental therapies is based upon the observation that not all patients respond favorably to a standardized care. The Latent (Hidden) Class Analysis (LCA) classification model has previously built [1], that provides parameters for other dental cohort to classify individuals into seven latent periodontal profile classes (PPC) classes based on 224 dichotomous clinical variables (from 32 teeth). The tooth-level LCA classified teeth into seven latent tooth profile classes (TPC), based on 14 categorical clinical parameters [2]. The combination of PPC and TPC allows the dental treatment evaluation at both the patient and tooth levels. We hypothesized that the LCA method with multiple clinical measures can better refine clinical phenotypes and provide insight into creating effective measures of treatment outcomes.

Methods: We accessed the dental electronic health records (EHR) for patients over the past 5 years, from the periodontal clinics at UNC, that include the clinical variables required to calculate PPC and TPC as well as the demographic variables and the treatment information. 3,704 subjects were placed into the novel LCA classifications (7 PPCs) and each of their teeth into one of the 7 TPCs. The criteria on the baseline data for follow-up exam data period and tolerance of missing teeth were applied. Treatment responses for scaling and root planning vs surgery were compared across PPCs and TPCs.

Results: For the teeth with severe TPC, significant difference of improvement at probing depths and bleeding on probing levels was observed for the patients who received surgery as compared to those who received scaling and root planing. By comparison, using the conventional American Dental Association (ADA) definition of classification of severe disease, no significant difference was observed when comparing scaling and root planing vs following surgery.

Conclusion: This LCA-based classification method may be more sensitive than the conventional classification method for detecting certain types of treatment responses. Furthermore, it suggests that measuring treatment outcomes might be streamlined based upon LCA classifications.

References:
Detecting Contradictory and Consistent Citations in Biomedical Literature

Jun Xu, Ph.D., Yonghui Wu, Ph.D., Yaoyun Zhang, Ph.D., Qiang Wei, M.S., Hua Xu, Ph.D.
School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, Texas, USA

INTRODUCTION
Lack of reproducibility in scientific studies is an increasing concern in biomedical research.¹ The scientific community has made many concerted efforts to improve reproducibility, such as raising research standards, promoting data sharing and re-use, and examining results through replication etc.² However, the current efforts require substantial resources and take long time to establish. Here we propose to develop novel citation analysis approaches to identify irreproducible studies. Scientists use citations to compare research findings with others, build strong and multi-faceted arguments, and report consistencies and contradictions among relevant studies. Hot research topics often bring many follow-up studies to examine new findings and compare with existing findings to report consistencies or raise contradictions. Figure 1 shows an example of the consistencies and contradictions among citations, where findings from the follow-up studies (reference ids: 12-15) contradict the findings reported in previous studies (reference ids: 9-11). Automatic detection of consistent and contradictory relations among biomedical literature citations could help researchers to understand author’s stances on specific scientific claims, which may be used to identify possible irreproducible studies. However, very few studies have explored the automatic detection of contradictions and consistencies through citation content analysis. To the best of our knowledge, this is the first study that investigates automatic methods to detect contradictory and consistent citations based on their stances on a specific claim.

MATERIALS AND METHODS
Data Set This study started with 34 eligible clinical studies collected by a review paper². We collected all the publications citing these 34 studies from the Open Access subset of PMC and extracted paragraphs containing citations referring to the above referenced 34 studies. Then, we developed annotation guidelines and manually annotated these paragraphs using three categories: Consistent, Contradictory, and NA. Figure 2 shows examples from our manual annotations. Through annotation, we constructed a corpus containing 2,210 relations (#Consistent: 632; # NA: 1,429; #Contradictory: 149) identified from 1,013 sentences.

Machine Learning Approach We extracted words, biomedical effects related lexicons (e.g. “increase” and ‘negative’) and syntactic components combining positions of the words (before, between or after the target pair) as features from the citing sentence. We developed a machine learning based relation classification system using the Support Vector Machines algorithm.

RESULTS

Table 1 shows the 10-fold cross-validation results for citation relation classification. The model combining all features achieved the best performance (Accuracy: 0.896, Micro-F: 0.896 and Macro-F: 0.781).

DISCUSSION and CONCLUSION
Automatic detection of contradictory and consistent citations could help biomedical researchers to understand author’s stances on specific scientific claims, which may be used to further identify possible irreproducible studies. We developed annotation guidelines and constructed the first data set for biomedical citation relation analysis. Preliminary results indicate that the citation relation classification is a challenging problem. To the best of our knowledge, this is the first study exploring automatic detection of contradictions and consistencies from biomedical literature through citation content analysis.

REFERENCES
Evaluation of Automated Spectrographic Seizure Detection Using Scale Invariant Feature Transform and Support Vector Machines

Peter Yan, MD1,2, Fei Wang, PhD2, Ramin Zabih, PhD3, Zachary Grinspan, MD, MS2,4

1Dept. of Neurology, Weill Cornell Medicine, NYC, NY; 2Dept. of Health Policy & Research, Weill Cornell Medical College, NYC, NY; 3Dept. of Computer Science, Cornell Tech, NYC, NY; 4Dept. of Pediatrics, Weill Cornell Medicine, NYC, NY

Introduction. Subtle and non-convulsive seizures (NCS) are difficult to clinically diagnose. The bedside clinician must rely on the electroencephalogram (EEG) as the primary diagnostic tool. EEG is complex. It contains 18-channels and requires neurophysiologist interpretation (gold-standard). In critically ill patients, the EEG workflow is laborious and may delay treatment by hours.[1] Thus, automated seizure detection would be valuable. EEGs can be processed into spectrograms, colored images that represent the EEG’s frequency content changing over time. One popular technique is the Compressed Spectral Array (CSA), but it remains complex, consisting of eight spectrograms.[2] We have previously developed the Median Power Spectrogram (MPS) that not only provided improved spectral resolution, but was less complex (1-3 spectrograms) and maintained seizure detection comparable to CSAs (mean 77% sensitivity and 72% specificity when used by bedside clinicians). Furthermore, users required only 5-minutes of training.[3] We hypothesize that Scale Invariant Feature Transform (SIFT) descriptors, a common image descriptor in computer vision[4], may be used with Support Vector Machines (SVM) to auto-detect seizures in spectrograms, and SVMs using SIFT descriptors from the MPS will outperform those from the CSA.

Methods. EEGs: 282 randomly selected 2-minute EEG segments, 50% containing seizures. Two minutes is the typical clinical window for treatment decisions. Seizures were annotated by neurophysiologists from Boston Children’s Hospital and Weill Cornell Medical Center. EEGs and annotations were acquired from the PhysioNet EEG database.[5] Spectrograms: CSAs and MPS images were generated via short time Fourier Transform (1024 point, 4s window, 1s step) from EEGs. 8-spectrogram CSAs were created using a cosine taper.[2] Single spectrogram MPS was created with multi-taper spectral estimation. Image Processing: SIFT descriptors were created over a 2D 4x4 grid (16 bins) moving across the image at 50% bin width. To determine the optimal bin size (in pixels [px]) for calculating SIFT descriptors, SIFT descriptors were calculated at various bin widths (8px by 2px incrementing to 24-px) for all images. Classification: Quadratic kernel C-SVMs were trained using SIFT descriptors. Hyper-parameters (kernel scale & C) were optimized using grid search. SVM performance was assessed using 10-fold cross validation in all optimization iterations. MATLAB v2017a was used in all steps.

Results. The optimal SIFT descriptor bin widths were 10px for CSA and 20px for MPS images. Using the optimized SIFT descriptors hyper-parameters, the MPS SVM performance was 91% sensitivity, 92% specificity, 0.98 AUC. The mean CSA SVM performance (across 8 spectrograms, all using optimized SIFT descriptors and hyper-parameters) was 66% sensitivity, 88% specificity, and 0.86 AUC.

Conclusion. Seizures can be identified on EEG spectrograms using SIFT descriptors and SVMs. The MPS trained SVMs outperformed CSA trained SVMs. Seizure detection using MPS based SVMs can provide rapid bedside seizure detection with high performance compared to current clinical practice, where inter-neurophysiologist agreement on EEG interpretations is at best modest and 20% may be inaccurate.[6]

References.
Preliminary Study of the Portability of an Automated Text Categorization Model to Identify Unsuspected Lung Nodule Findings in Radiology Reports

Jianji Yang, PhD1, Theodore B. Wright, MD1,2, James Morrison, MD1,2
1VA Portland Healthcare System; 2Oregon Health & Sciences University, Portland, OR

Introduction
The Unsuspected Radiologic Findings (URF) Registry, a web-based tool developed at The VA Portland Healthcare System (VAPORHCS), helps expedite the diagnosis of lung cancers by ensuring that incidental lung nodules are identified and managed in a timely manner1. During dictation of a radiology report, a unique URF code is added to the results of the study at the discretion of the radiologist. The presence of this code is used as the inclusion criterion for the Registry. To improve the accuracy of the registry, a text categorization model was developed to identify nodule-positive reports with missing URF codes by analyzing the report text2. The model was developed using a linear-kernel Support Vector Machine trained on a bag-of-words representation of the Impression text of fifty-thousand radiology reports. It performed well with F1 & F5 scores of 0.98 in evaluation. In November 2016, VA Puget Sound Health Care System (VAPSHCS) implemented the URF Registry and their radiologists began adding URF codes to nodule-positive cases. Before implementing the text categorization model to identify cases with missing URF codes at VAPSHCS, the model’s portability was evaluated by comparing its performances at VAPORHCS and VAPSHCS.

Methods
We applied the model to both VAPORHCS and VAPSHCS lung-related radiology reports from 1/1 to 3/2/2017 to classify each case as either nodule-positive or nodule-negative. We defined true positives as cases that had a URF code assigned by radiologists, while false positives had no associated URF code. We then calculated the performance measures of recall, precision, F1 and F5. As the model aims to provide a safety net to capture all nodule-positive cases, we report F5 scores (emphasizing recall) as the key measure for performance. Finally, false positives from VAPSHCS were further reviewed by an Internist at VAPORHCS (TBW) to evaluate the error rate of URF coding.

Results
At VAPORHCS, 766 (6%) out of 12,499 lung-related imaging studies had a URF code. At VAPSHCS, 172 (2%) out of 7,816 were URF-coded. Compared to VAPORHCS, model performance dropped in all categories at VAPSHCS, with the precision for nodule-positive suffering the most. Please see Table 1 for a summary and comparison of model performance. After expert review, 71 (43%) out of the 167 false positive cases at VAPSHCS were found to be nodule-positive actually. This is a significantly higher rate than reported by prior studies at VAPORHCS (10%)2.

Table 1. Model performance at VAPORHCS (left) and VAPSHCS (right). Nodule-Positive/Negative: imaging studies classified by the text classification model as with/without the presence of nodule(s).

<table>
<thead>
<tr>
<th>VAPORHCS</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
<th>F5</th>
<th>VAPSHCS</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
<th>F5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nodule-Positive</td>
<td>0.68</td>
<td>0.94</td>
<td>0.79</td>
<td>0.93</td>
<td>Nodule-Positive</td>
<td>0.22</td>
<td>0.76</td>
<td>0.34</td>
<td>0.69</td>
</tr>
<tr>
<td>Nodule-Negative</td>
<td>1.00</td>
<td>0.97</td>
<td>0.98</td>
<td>0.97</td>
<td>Nodule-Negative</td>
<td>0.99</td>
<td>0.94</td>
<td>0.97</td>
<td>0.94</td>
</tr>
<tr>
<td>Weighted Avg.</td>
<td>0.98</td>
<td>0.97</td>
<td>0.97</td>
<td>0.97</td>
<td>Weighted Avg.</td>
<td>0.98</td>
<td>0.93</td>
<td>0.95</td>
<td>0.94</td>
</tr>
</tbody>
</table>

Discussions and Conclusions
Model performance at VAPORHCS remained high and consistent with past results2. The prior probability for nodule-positivity appears much lower at VAPSHCS than VAPORHCS. This might be due to URF under-coding early in the implementation, or could reflect a true lower incidence of URF. The additional 71 node-positive cases found after expert review support the first possibility. They also partially explain the drop in precision (0.68 to 0.22). Other factors affecting the portability, e.g. difference in narrative style and URF coding guidelines, will need further exploration.

In conclusion, we show that porting an automated text categorization model to a new corpus led to an expected drop in performance. We also show that the model is capable of uncovering human error and finding room for improvement in an early implementation of the URF registry. However, while the baseline model might be used to identify reports with missing URF codes and provide feedback for improving clinician coding adherence early in an implementation, re-training of the model using a local corpus should be performed as this should further improve model performance.

References
Visualizing Social Support Mechanisms in English and Spanish Tweets to Meet Self-Management Needs of Family Dementia Caregivers

Sunmoo Yoon, PhD¹, Niurka Suero-Tejeda, MA¹, Robert Lucero, PhD², Suzanne Bakken, PhD¹

¹Columbia University School of Nursing, New York, NY, ²University of Florida College of Nursing, FL, USA

Abstract
We analyzed 314,000 dementia related tweets collected on World Alzheimer’s Day in 2015 and 2016 applying network analysis, topic modeling, and visualization techniques to gain insights about dementia caregiving.

Introduction
The incidence of dementia is increasing. Although belongingness is a human basic need, in our previous studies we observed many isolates in Twitter social networks who do not receive feedback from others ¹. The study purpose is to visualize social support networks and topics to gain insights about dementia caregiving.

Methods
Using NodeXL and Ncapture, we extracted 314,000 Tweets mentioning Alzheimer’s and dementia/demencia related keywords from Tweet corpora collected on World Alzheimer’s Day in 2015 and 2016. We applied Latent Dirichlet allocation (LDA)-based topic modeling (Figure 1)², network analysis, and novel visualization techniques.¹

Results
48 different languages were detected. LDA revealed self-management concerns of dementia caregivers including role (i.e. finding providers), medical (i.e sleep, symptom awareness) and emotional (i.e. stress, depression) management. The proportion of isolates, who express their opinions and feelings but receive no feedback, was substantially decreased in 2016 (P< .0001) as compared to 2015(Figure 2).

Conclusion
Application of mining and visualization techniques provided knowledge for insights regarding dementia caregiving and self-management needs for the person with dementia and the dementia caregiver. The topic modeling findings such as early diagnosis from English and family caregiving supports from Spanish corpus will be used for a targeted intervention development for English- and Spanish-speaking Hispanic dementia caregivers.

Acknowledgement: U.S. National Institute of Health grant #R01NR014430-03S2; P30 NR016587-01

References
Novel Visualization of *Clostridium difficile* Infections in Intensive Care Units

Sean C. Yu, MS\(^1\), Courtney L. Hebert, MD, MS\(^1\), Albert M. Lai, PhD\(^2\), Justin Smyer, MLS(ASCP)CM, MPH, CIC\(^1\), Jennifer Flaherty, RN, MPH, CIC\(^1\), Julie Mangino, MD\(^1\), Susan Moffatt-Bruce MD, PhD, MBA\(^1\)

\(^1\)Ohio State University Wexner Medical Center, Columbus, OH, USA; \(^2\)Washington University, St. Louis, MO, USA

**Introduction**

Approximately 1 in 25 inpatients in U.S. acute care hospitals develops at least one healthcare-associated infection (HAI), and *Clostridium difficile* (CDI) is the most common pathogen\(^1\). CDC National Healthcare Safety Network (NHSN) publishes reporting guidelines to track HAIs including CDI. It is widely known that HAI CDI can be transmitted via contact with contaminated hospital sources, and there have been efforts to incorporate geospatial and temporal dimensions for CDI surveillance\(^2\). However, most studies do not leverage geographic information systems (GIS) nor do they have a path towards operationalization of real-time analytics or visualization tools to track HAIs.

**Methods**

Lab results and admission discharge transfer (ADT) data on all patients admitted to the Ohio State Wexner Medical Center (OSUWMC) from 1/1/2015–12/1/15 were extracted from the information warehouse and processed in accordance with CDC NSHN guidelines to identify HAI CDI cases. Results were validated against the infection preventionist (IP)-curated surveillance database as a gold standard. We obtained GIS files for the hospital and used them to prototype a dashboard in Tableau (Seattle, WA). The dashboard shows OSUWMC intensive care unit floor plans and allows the user to visualize HAI CDI cases within an adjustable time period as dots in the room of attribution. Rooms that a patient was housed in from 2 days before to 7 days after diagnosis are shaded, so the user can see rooms that may have been contaminated during that time period. The user can also select single or multiple rooms to drill down and identify details of patient(s) with CDI in those rooms. We performed a formal usability study for this dashboard with four Clinical Epidemiology staff using a think-aloud protocol\(^3\) and anonymous survey.

**Results**

Using integrated data, we were able to automate HAI CDI surveillance and display results on an interactive map of the ICUs. During the study period, we recorded 253 HAI CDI cases throughout our multi-hospital system. Our simple algorithm correctly identified all 253 HAI CDI cases with a PPV of 97.31%. On the survey, all users “strongly agreed” that the dashboard would be a positive addition to Clinical Epidemiology and would allow them to present HAI information to others more effectively. All “agreed” or "strongly agreed" that they felt confident in manipulating the dashboard to demonstrate to others and that it was easy to learn and use. Respondents also suggested improvements to specific features including improving the intuitiveness of changing the date ranges, adding isolation information, and showing present on admission CDI cases.

**Conclusion**

Given success with surveillance automation and visualization and positive responses from the initial round of usability testing, we are now developing a second web application-version for improved flexibility, extensibility, and the capacity to integrate with existing systems including the EHR. We anticipate that this application will be leveraged for use with other HAIs and hospital events that would benefit from spatiotemporal visualization.

*This project was supported by the Institute for the Design of Environments Aligned for Patient Safety (IDEA4PS) at The Ohio State University which is sponsored by the Agency for Healthcare Research & Quality (P30HS024379). The authors’ views do not necessarily represent the views of AHRQ.*

**References**

Subtype Discovery for Autism Spectrum Disorder by Comorbidity Analysis

Kun-Hsing Yu, MD, PhD¹, Nathan Palmer, PhD¹, Isaac S. Kohane, MD, PhD¹
¹Department of Biomedical Informatics, Harvard Medical School

Introduction

More than 3.5 million Americans suffer from autism spectrum disorder (ASD), costing $236 billion annually in medical and behavioral interventions. ASD patients have a wide range of clinical manifestations and prognosis. However, the current classification guideline (Diagnostic and Statistical Manual of Mental Disorders, 5th Edition; DSM-5) only accounts for disease severity and intellectual or language impairment. A richer and clinically relevant subtyping system will better distinguish ASD patients with different medical needs and different underlying biology. To address this challenge, we studied the comorbidities of ASD patients in a nationwide health insurance database and identified the phenotypical differences between patients with different patterns of comorbidities.

Materials and Methods

Using the unidentifiable member claims data from Aetna, a total of 132,211 ASD patients were found using the PheWAS code 313.3, and a 1:1 age-and-gender-matched control cohort was selected from this dataset. To identify the phenotypical differences between ASD patients with and without certain comorbidities, all PheWAS codes between the comorbidity groups were examined. Specifically, for each of the known ASD comorbidities (Cᵢ), such as epilepsy, inflammatory bowel disease (IBD), and obsessive-compulsive disorder (OCD), ASD patients with and without Cᵢ were separated into two groups (Gᵢ₊ and Gᵢ₋), and the prevalence rate of each PheWAS phenotype (Pⱼ) between age 0 and 19 was calculated for both groups. For each Pⱼ, Wilcoxon rank sum test was employed to identify its prevalence rate difference between Gᵢ₊ and Gᵢ₋, and the odds ratio (OR) was estimated to determine the effect size. The same procedure was repeated for the control cohort. Benjamini-Hochberg procedure was employed to control the false discovery rate (FDR) at 0.05. All analyses were performed using R version 3.3 with R packages icd and NMF.

Results

We found that ASD patients with epilepsy were more likely to suffer from primary pulmonary hypertension (OR=4.80), somatoform disorder (OR=3.89), aphasia (OR=3.58), esophagitis (OR=2.96), and cystic fibrosis (OR=2.37), comparing with those without epilepsy. Similarly, ASD patients with IBD were more likely to have celiac disease (OR=9.87), immunity deficiency (OR=8.33), and chronic sinusitis (OR=2.25) comparing with ASD without IBD. Moreover, ASD patients with OCD were more likely to develop tics and chorea (OR=4.16) and irritable bowel syndrome (OR=2.46) but are less likely to have cystic fibrosis (OR=0.40). These differences (FDR<0.05) were not observed between non-ASD participants with and without epilepsy, IBD, or OCD, respectively, and the ORs in ASD patients with two comorbidities were not simple multiplications of the ORs obtained from patients with one comorbidity. Comparing to ASD patients, non-ASD children were more likely to have influenza and clavicular fracture. The modularity of these comorbidities indicated that ASD might constitute many dissimilar disease subtypes.

Discussion

To our knowledge, this is the first systematic analysis of the ASD comorbidity patterns in a large national cohort. We demonstrated that patients with different comorbidities have dissimilar clinical phenotypes. For instance, ASD patients with IBD had higher risks of many immune-related disorders. However, these immune disorders were not enriched in ASD patients without IBD or IBD patients without ASD, indicating that ASD-IBD could be a distinct ASD subtype with unique clinical manifestations. These findings have direct implications in the clinical managements of ASD patients. Further studies are needed to explore the biological mechanisms underpinning these phenotypes.

References

Utilizing EHR-based Recruitment Methods

Marianne Zachariah1,2, Carina V. Hampp1,2, Ruby Y. Wan, MS1,2, Amanda L. Do, MPH1,2, and Douglas S. Bell, MD, PhD1,2

1UCLA Clinical & Translational Science Institute (CTSI), Los Angeles, CA; 2UCLA David Geffen School of Medicine, Los Angeles, CA

Introduction

Traditionally, studies involving patient recruitment have relied on methods such as flyers and advertisements, physician referrals, and in-clinic recruitment. Unfortunately, these methods typically result in low enrollment rates. Electronic health record (EHR)–based recruitment methods offer an alternative, and possibly more effective approach. For this study, we evaluated: 1) the extent to which research studies involving patient recruitment at UCLA utilized EHR-based methods, 2) considerations that may be needed when adopting EHR-based recruitment methods, and 3) the success of EHR-based recruitment methods.

Methods

All studies for which data has been provisioned by the UCLA CTSI Informatics Program for recruitment efforts were selected for review. Each study’s IRB application was analyzed, and information related to study methods and recruitment was summarized. Variables of interest were identified, including whether traditional recruitment methods were used in conjunction with EHR-based methods vs. EHR-based methods only, the outreach methods (in-clinic, flyers/ads, referrals letters, telephone, email), on whose behalf outreach was conducted, if physician approval to approach patients was obtained, the department of principal investigator (PI), the study’s level of sensitivity, and whether any patient complaints were received. Enrollment statistics were also requested for each study and compared to recruitment methods and recruitment language.

Results

Of the 27 studies reviewed to date, 78% (21) used a combination of traditional recruitment methods (e.g. ads/flyers, physician referral) and EHR-based recruitment methods while 22% (6) relied solely on EHR-based recruitment methods. Of the studies that used traditional recruitment approaches, 86% (18/21) involved in-clinic recruitment or physician referral, 71% (15/21) used ads/flyers, and 14% (3/21) reached out to patients who participated in other studies. Focusing only on EHR-based methods, 52% (14/27) of studies obtained physician approval before contacting patients, 70% (19/27) sent letters to a list of eligible patients, 11% (3/27) sent emails or patient portal messages, 67% (18/27) telephoned patients, and 4% (1/27) provided physicians with a list of their eligible patients and asked for assistance with recruitment. For the studies that used EHR-based methods to identify patients for outreach via a letter, email, portal message, or telephone call (25/27 studies), patients were informed that they were being contacted on behalf of either the study’s PI (72%, 18), a department chair (8%, 2), the patient’s own physician (16%, 4), or it was ambiguous (24%, 6). The studies spanned 13 departments or divisions (e.g. 19% nephrology, 15% rheumatology, 11% general internal medicine), and 15% (4/27) were on a sensitive topic (e.g. drug abuse) while 85% (23/27) were not sensitive. Patients from 2 studies (7%) complained about the outreach to the IRB. One complaint was related to a violation of the IRB-approved outreach protocol. The other was related to contacting a patient who had died. Results for enrollment statistics are pending as data collection is still ongoing.

Discussion & Conclusion

Preliminary analysis indicates that most studies (78%) use a combination of traditional recruitment methods and EHR-based recruitment methods, suggesting that researchers still believe that traditional recruitment methods are effective. The most common traditional method of recruitment is in-clinic recruitment or physician referral (86%), which suggests that researchers believe that face-to-face interaction is important. The most common EHR-based outreach method is mailed letters (70%), which is costly compared to email, but email is utilized much less (11%). Low use may be attributed to security/privacy concerns from the IRB and the fact that some patient subgroups would be underestimated. When patients were contacted by letter, email, portal message, or telephone call, the EHR-supported outreach was most often conducted on behalf of the PI (72%). Most studies requesting data for recruitment purposes were not related to a sensitive topic (85%). Further analysis of results from these recruitment efforts will investigate how enrollment rates were affected by the factors listed above.
Recommending education materials for diabetic questions using information retrieval approaches

Yuqun Zeng$^{1, 2}$, Yanshan Wang$^2$, Feichen Shen$^2$, Sijia Liu$^2$, Liwei Wang$^2$, Majid Rastegar Mojarad$^2$, Xusheng Liu$^1$, Hongfang Liu$^2$
$^1$The Second Clinical College, Guangzhou University of Chinese Medicine, Guangzhou, Guangdong, P.R. China $^2$Division of Biomedical Statistics and Informatics, Mayo Clinic, Rochester, MN, USA.

Background
Self-management is crucial to diabetes care and providing expert-vetted content for patients’ questions is crucial in facilitating patient self-management.

Objective
To investigate the use of information retrieval (IR) techniques in recommending patient education materials for diabetic questions of patients.

Methods
We compared two retrieval algorithms, one based on LDA topic modeling (TMB) and one based on semantic group (SGB), with the baseline retrieval models, vector space model (VSM) in recommending diabetic patient education materials (144 materials from patient education database at Mayo Clinic) to diabetic questions posted on the Tudiabetes forum (7,510 diabetic questions). The evaluation was based on a gold standard made by two experts with high agreement, and the experts manually assigned all diabetic education materials with the relevancy (0-2 ranking) to give 50 randomly selected diabetic questions. The performance was assessed using a precision algorithm of top ranked documents.

Results
Corpus analysis indicates language used by diabetic questions is different from patient education materials, the Heatmap of forum question category and PEM topic shows in Figure 1. TMB outperformed other retrieval algorithms (Figure 2). For example, for the top retrieved document, the precision of TMB, SGB, and VSM model is 67.0%, 62.8%, and 54.3% respectively.

Conclusion
The study demonstrated topic modeling can mitigate the vocabulary difference and achieved the best performance in recommending education materials to questions. One direction for future work is to assess the generalizability of our findings and extend our study to other disease areas, other patient education material resources, and online forums.

Figure 1
![Heatmap of forum question category and PEM topic based on cosine similarity of word vectors weighted using TF-IDF or topic word distribution. The clustering is based on Euclidean distance.](image1)

Figure 2
![Figure 2. Precision at rank 1 to 20 for TMB, SGB, and VSM models.](image2)
Towards Improving the Quality of Medicaid Data: Identifying Data Defects in a Large Healthcare Administration System

Yili Zhang, MS, Pratik Tamakuwala, BE, and Güneş Koru, PhD
University of Maryland, Baltimore County, Baltimore, MD

Introduction
As the use of information systems in healthcare increases, data quality problems are encountered at an increasing rate in healthcare organizations.¹ A lack of data quality can result in imprecise, useless, or even misleading results which detract from the quality of reports produced and decisions made.² Medicaid data have served as a useful resource for various healthcare research activities, however, the deficiencies in the Medicaid data often increase the difficulty in obtaining useful results.³ Thus, effective data maintenance and cleaning become crucial to improve the quality and utility of the Medicaid data. In this research study, we focused on the first step to improve the quality of data in the Provider, Procedure, and Diagnosis subsystems of a Medicare and Medicaid Management Information System (MMIS). Specifically, we identified the types of data defects, created a taxonomy of data defects, and finally detected all of the data defects in the MMIS data automatically with respect to the taxonomy created. In this context, a data defect refers to a discrepancy between the actual and expected values held by a data item which requires a corrective change.

Methods
As the first step, all of the available and relevant documents kept in the organization were reviewed. Next, a descriptive analysis was performed to better understand the data and create a data defect taxonomy with essential data defect categories. Then, an extensive literature review was performed to supplement more categories to the taxonomy of the data defects. A data quality toolkit (DQT) was developed to detect the data defects automatically and efficiently according to the created taxonomy. Finally, DQT was run on the datasets to detect the existing defects in various categories. The datasets subject to defect detection consist of 1.5 million, 700 thousand, and 200 thousand records for the Provider, Procedure, and Diagnosis subsystems, respectively.

Results
The taxonomy of data defects includes six major categories: Missing data, wrong value, syntax violation, semantic violation, duplicate, and ambiguous value. These major categories are further divided into eighteen subcategories. Corresponding to this defect taxonomy, DQT detected more than three million data defects in the MMIS data, which is more than the number of records since each record has multiple columns and the value in each column could violate one or more rules. The syntax violation category had the highest number of defects, followed by the missing data and semantic violation categories. Most of the syntax violation defects were about the invalid values of certain Medicaid codes. Wrong value usually occurred due to large amount of implausible high or low values. These problem areas should be the focus of future organizational initiatives for data quality improvement.

Conclusion
The results contribute to the body of knowledge in this area by providing (i) refined and detailed taxonomy of data defects, (ii) effective and efficient methods to detect data defects, and (iii) the prevalence and distribution of data defects in a large MMIS. So far, there has been no study which created a taxonomy of data defects for the healthcare data and detected the data defects automatically. This research takes the first step to make the Medicaid data an even more useful resource for healthcare research. More generally, it makes a timely contribution for improving data quality as we make progress towards learning health systems.

References
A Twitter Study of the Relationship between the Geographic Variations of HPV Vaccination Rates and Online Information in the United States

Hansi Zhang, BS1, Yunpeng Zhao, BS1, Christopher Wheldon, PhD2, Xinsong Du, BS1, Terrell Brown, PhD3, Yi Guo, PhD1, Stephanie Staras, PhD1, Jingcheng Du, BS4, Cui Tao, PhD4, Jiang Bian, PhD1

1University of Florida, Gainesville, FL; 2National Cancer Institute, Rockville, MD; 3Florida A&M University, Tallahassee, FL; 4UT Health Science Center at Houston, Houston, TX

Introduction: Human papillomavirus (HPV) is responsible for more than 90% of anal and cervical cancers. On the other hand, the current HPV vaccine can prevent 90% of cancers caused by HPV infections. The HPV vaccination coverage in US remains low, and has significant variations among states. Nationally, coverage rates of HPV vaccination have been increasing since its recommendation. Online information and online discussions of HPV-related issues may have a considerable influence on public’s perception and acceptance of HPV vaccination. Thus, the geographic variations may be related to the availability and the quality of online HPV information. Further, the use of social media such as Twitter has dramatically increased over the past decade. Nowadays, social media is one of the most popular ways to get health information and exchange health-related experiences. In this work, we use Twitter data to study the relationship between the state-level variations of HPV vaccination rates and online information.

Methods: First, we collected tweets related to HPV using keywords and their variations such as “hpv”, “human papillomavirus”, and “gardasil”. We first preprocessed the collected data to eliminate tweets that were not written in English. Using our previous geocoding method, we assigned each tweet to a US state based on the location string in the user’s profile, and eliminated those that were outside of US. We used a set of simple rules to categorize these HPV tweets into two groups: 1) online information such as news and expert opinions, and 2) lay people’s discussions and reactions to this online information. Second, we used the TextBlob tool to assess the sentiments (i.e., positive, negative and neutral) of the tweets tagged as lay people discussions. We then aggregated the sentiment analysis results at the state level. Third, we extracted topics from the tweets tagged as online information using the Latent Dirichlet Allocation (LDA) model for topic modeling. Last, we selected several opposite states based on their adolescent HPV vaccination rates (low vs. high), and qualitatively studied the relationship among the vaccination rates, the topics and lay people’s sentiments among these states.

Results: We collected 242,685 tweets from November 2015 to December 2016. After eliminating non-English and non-US tweets, we retained 62,396 tweets for further analysis. Based on a CDC report, we selected 2 states with high vaccination rates (DC, PA), and 2 with low rates (FL, GA). The set of rules (e.g., a tweet is tagged as online information if it contains an URL or from an organization) we used performed reasonably well to categorize the tweets (precision:0.84, recall:0.72) on a manually reviewed sample (~100). We also evaluated the sentiment analysis tool on another manually labeled sample (~100), and obtained good performance (precision: 0.80, recall:0.63). The proportion of positive sentiments is consistently higher than negative sentiments across all states, which reflects the rising trend of HPV vaccination uptakes. Further, we observed geographic variations in terms of HPV vaccination rates and lay people’s sentiments towards HPV. States with high vaccination rates also have high positive to negative sentiment ratios (DC: 3.38; PA: 2.53); while, states with low vaccination rates have low sentiment ratios (FL: 1.57, GA: 1.69). Moreover, using LDA, we extracted topics from online information of two opposite states, DC with a high vaccination rate vs. FL with a low vaccination rate. As shown in Figure 1, there were more negative words in FL than in DC, such as ‘adverse effects’, ‘deadly’, and ‘scam’. While in DC, the words like ‘good’, ‘effective’, ‘safety’ appeared more frequently.

Conclusion: Our study found that the lay people’s opinions towards HPV on social media in a state are associated with the state’s HPV vaccination rate. Promoting positive HPV vaccine discussions on social media, including Twitter, may be a promising intervention tool to increase HPV vaccination and, in turn, cancer prevention.

References
Effects of Clinical Decision Support Implementation on Clinical Guideline Compliance

Yongkang Zhang, MS, Mark L. Diana, PhD, Allison B. McCoy, PhD

Tulane University School of Public Health and Tropical Medicine, New Orleans, LA

Introduction
Strong evidence indicates that implementation of protocols that standardize practices may reduce the risk of surgical infection. The Surgical Care Improvement Project (SCIP) was created to reduce the rate of surgical complications through increased compliance with infection prevention measures. Prior research has documented wide adoption of clinical decision support (CDS) incorporating clinical guidelines among hospitals. However, few studies have tested the association between CDS adoption and clinical guideline compliance. In this study, we sought to evaluate the effects of CDS implementation on hospitals’ performance on SCIP measures.

Methods
We performed a cross-sectional study of hospital CDS adoption, SCIP measures, and hospital characteristics in 2014. We retrieved hospital characteristics and CDS adoption data from the 2014 American Hospital Association (AHA) Annual Survey and Information Technology Supplement, and SCIP measures from the 2014 Hospital Compare data set; 2,699 hospitals were included in the study. Based on hospitals’ self-reported information, we categorized adoption of CDS with clinical guidelines into three levels: fully implemented across all units, fully implemented in at least one unit, and other (i.e., plan to implement or not considering implementing). SCIP measures included percent of surgical patients who received prophylactic antibiotics within 1 hour prior to surgical incision (INF-1), percent who received recommended prophylactic antibiotics (INF-2), percent whose prophylactic antibiotics were discontinued within 24 hours after surgery end time (INF-3), percent with urinary catheter removed on postoperative day 1 or day 2 (INF-9), percent on beta-blocker therapy prior to admission who received a beta-blocker during the perioperative period (CARD-2), and percent who received appropriate venous thromboembolism (VTE) prophylaxis within 24 hours prior to anesthesia start time to 24 hours after anesthesia end time (VTE-2). Hospital characteristics included ownership, number of staffed beds, teaching status, health care system affiliation, proportion of Medicare patients, proportion of Medicaid patients, and rural/urban location. We used linear regression to test the association between CDS adoption and SCIP measure scores.

Results
Among all hospitals, 1,920 (71.94%) had full implementation of CDS across all units, 426 (15.96%) had full implementation in at least one unit, and 323 (12.10%) had other implementation status. Not-for-profit hospitals, hospitals with health care system affiliations, teaching hospitals, hospitals with more staffed beds, hospitals with lower Medicare and Medicaid shares, and hospitals in urban areas were more likely to fully implement CDS with clinical guidelines. Multivariate analyses indicated that the full implementation of CDS with clinical guidelines across all units was associated with higher scores in CARD-2 (β=1.18, p=0.003) and VTE-2 (β=0.617, p=0.015). No other significant associations were found.

| Table 1. SCIP measure scores (mean and standard deviation) by CDS implementation status |
|---------------------------------|--------------------------------|---------------------------------|------------------|------------------|
| Measure          | Fully implemented across all units | Fully implemented in at least one unit | Other            | Overall          |
| INF-1            | 98.58 (4.39)                  | 97.78 (4.42)                  | 97.84 (4.68)     | 98.40 (4.43)     |
| INF-2            | 98.36 (3.23)                  | 97.62 (4.56)                  | 97.47 (4.30)     | 98.17 (3.58)     |
| INF-3            | 98.04 (3.40)                  | 97.08 (4.80)                  | 97.01 (4.87)     | 97.81 (3.80)     |
| INF-9            | 97.83 (4.16)                  | 96.98 (4.66)                  | 97.28 (4.81)     | 97.66 (4.30)     |
| CARD-2           | 97.88 (4.12) *                | 96.40 (5.36)                  | 96.37 (7.30)     | 97.55 (3.13)     |
| VTE-2            | 99.70 (2.45) *                | 99.12 (5.49)                  | 99.30 (2.90)     | 99.58 (3.13)     |

Conclusion
We found that full implementation of CDS was associated with better compliance with clinical guidelines in two SCIP measures. Further research using longitudinal data and additional measures is warranted and underway to provide more evidence on the effects of CDS implementation on clinical guideline compliance and other quality outcomes.
Evaluating Word Embeddings from Multiple Domains for Symptom Recognition in Psychiatric Notes
Yaoyun Zhang, Ph.D., Hee-Jin Li, Ph.D., Yonghui Wu Ph.D., Hua Xu Ph.D.
School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX, USA

Introduction & Background
Recently there is a rapid growth of deep learning (DL) methods and their application in various natural language processing (NLP) tasks, promoted by word embeddings pre-trained from large-scale corpora of multiple domains. However, few NLP studies have used DL methods in the biomedical domain, potentially due to the lack of large amounts of data to generate word embeddings for different sub-domains. The goal of this study is to investigate and exploit word embeddings of external domains for clinical NLP tasks. Specifically, we take the psychiatric symptom recognition task as a typical use case. Psychiatric symptoms consist of subjective descriptions of the patient’s experience, as well as the nature and severity of mental disorders, which are critical for qualitative investigation into mental disorders. However, extracting psychiatric symptoms from clinical text suffers from the data sparseness problem and a low coverage by existing biomedical lexicons, making it challenging to apply traditional named entity recognition methods to this task. We will evaluate the performance of word embeddings from four domains – intensive care, biomedical literature, newswire and social media – to recognize entities of psychiatric symptoms using DL methods. To the best of our knowledge, this is the first work of evaluating multiple word embeddings for psychiatric symptom recognition.

Methods
Datasets: In addition to word embeddings trained from 1,000 psychiatric notes provided by the CEGS N-GRID 2016 challenge, word embeddings pre-trained from corpora of four external domains are investigated, including (1) MIMIC II with clinical text of intensive care units, (2) Medline with abstracts of biomedical literature, (3) Gigaword with newswire from the open domain, and (4) psychiatric forum data from WebMD. Algorithms: Word embeddings are generated by the skip-gram model in the word2vec package by default settings. The recurrent neural network (RNN) algorithm proposed by Lample et al. is employed to build the DL-based symptom recognition model. The pre-trained word embeddings are used to initiate the word embedding vectors in RNN at the beginning of the training process. Evaluation: 400 psychiatric notes are annotated with symptoms and are split into training (60%), development (20%) and test datasets (20%) for experiment. The performance of psychiatric symptom recognition is evaluated using precision (P), recall (R) and F-measure (F1). The result of exact match with the gold standard annotations is reported.

Results & Discussion
As displayed in Table 1, the baseline of using random initiation of RNN without any pre-trained embedding yielded the lowest result, with a F-measure of 68.94%. Surprisingly, the top 2 F-measures of 73.71% and 73.42% were achieved by word embeddings from Gigaword and the Psychiatric Forum respectively, which outperformed the three biomedical corpora, psychiatric notes, MIMIC and MEDLINE. Besides, directly concatenating the five word embeddings gained a median F-measure of 72.71%.

One potential reason of the higher performance obtained using word embeddings from the newswire and social media corpora is due to the essential characteristics of psychiatric symptoms. Psychiatric symptoms consist of words more commonly present in the open domain text, instead of the biomedical terminologies in clinical notes of intensive care and biomedical literature. More complex methods beyond direct concatenation to integrate word embeddings from multiple domains will be explored in the next step.

Reference

Table 1 Results for RNN-based psychiatric symptom recognition using word embeddings from each domain.

<table>
<thead>
<tr>
<th>Corpus</th>
<th>P</th>
<th>R</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>72.84</td>
<td>65.45</td>
<td>68.94</td>
</tr>
<tr>
<td>Psychiatric Notes</td>
<td>73.57</td>
<td>70.26</td>
<td>71.88</td>
</tr>
<tr>
<td>MIMIC</td>
<td>76.22</td>
<td>69.17</td>
<td>72.53</td>
</tr>
<tr>
<td>MEDLINE</td>
<td>69.66</td>
<td>70.78</td>
<td>70.21</td>
</tr>
<tr>
<td>Gigaword</td>
<td>75.87</td>
<td>71.68</td>
<td>73.71</td>
</tr>
<tr>
<td>Psychiatric Forum</td>
<td>74.68</td>
<td>72.19</td>
<td>73.42</td>
</tr>
<tr>
<td>Concatenation</td>
<td>73.84</td>
<td>71.61</td>
<td>72.71</td>
</tr>
</tbody>
</table>
Detecting Off-label Uses of Prescription Drugs with Meta-Path-Based Mining
Mengnan Zhao¹, Christopher C. Yang¹
¹College of Computing and Informatics, Drexel University, Philadelphia PA 19104

Background
Off-label drug uses are very common in clinical practice. Up to one-fifth prescriptions are off-label¹. To some extent, off-label drug uses provide a pathway for clinical innovation, however, they could cause serious adverse effects due to lacking scientific research and tests. Identifying those off-label uses can provide a clue to the stakeholders to further the investigation on drug use and safety. Therefore, there is the need to develop an automated method to detect off-label drug uses.

Objective
We focus on developing a data mining approach to identify off-label drug uses systematically, rather than relying on human annotations or surveys.

Methods
• Dataset
Large numbers of health consumers exchange health information on online health communities (OHCs); MedHelp, as a pioneer in OHCs, empowers over 12 million people each month to discuss health related issues on the site, which provide huge volumes of valuable information. Considering the fact that many people describe their drug uses in OHCs, we collected data from MedHelp and retrieved more than 70,000 threads.

• Heterogeneous Network Construction and Mining
A heterogeneous network is defined as a graph consisting of nodes connected by links, with at least two types of nodes and at least two types of links, capturing more essential and accurate information about the network. Based on the medical entities we identified from OHC data, we constructed a heterogeneous network that contains three types of nodes (i.e. disease, drug, ADR) and six types of links.

A meta-path is a path defined on the network schema in the form of \( A_1 \rightarrow A_2 \rightarrow \ldots \rightarrow A_{i+1} \), which composes the relations between nodes in the heterogeneous network. Here we exploited meta-path-based networking mining approaches to detect all the possible relationships between the disease and drug\(^*\). \( s(d_i, r_j) \) denotes the relation between nodes \( d_i \) and \( r_j \); \( f(p,d_i,r_j) = 1 \) when there is a meta-path \( p \) connecting \( d_i \) and \( r_j \):

\[
s(d_i, r_j) = \begin{cases} 1, & (\forall p \in P, f(p, d_i, r_j) = 1) \\ 0 & \end{cases}
\]

• Classification
By setting the maximum distance to 3, we defined 13 meta paths in total, and we used the calculated value on each meta path as an input feature for the classifier. There are 2,087 positive disease-drug pairs suggested by Drugbank and PharmGKB and 28,000 generated negative pairs in the dataset. Then we built the binary classification model on Random Forest algorithm and tuned the parameters by applying 10-fold cross validation on the training sets.

Results & Evaluation

Conclusion & Future Plan
This initial research shows the possibility of using OHC data and network mining approaches for off-label drug use detection, as well as the potential of developing an automated detection method. The result confirmed the hypothesis that the connections between medical entities and the network related information are important for detecting off-label drug uses. Therefore, in the future, we plan to extract more network information by using heterogeneous network mining skills or other text mining skills.

References
HyDeXT: A Hybrid De-identification and Extraction Tool for Health Text

Xinyan Zhao, MSI, V.G.Vinod Vydiswaran, PhD
Department of Learning Health Sciences, University of Michigan, Ann Arbor, MI

Introduction: Identifying Protected Health Information (PHI) from clinical notes is critical to protect the privacy of individuals before their data could be shared publicly. The goal of de-identification is to identify and remove PHI from health records while protecting the integrity of the data as much as possible. In this poster abstract, we present HyDeXT, a Hybrid De-identification and eXtraction tool. HyDeXT combines the labels derived from a machine learning-based sequential tagger and a pattern-based labeler, to effectively de-identify eighteen categories of PHI and outperform existing state-of-the-art approaches.

Methods: The hybrid de-identification and extraction tool, HyDeXT, consists of a text processing pipeline with the following four stages: (i) Data cleaning and pre-processing; where the original text is modified by parsing the raw text format, adding missing whitespace characters, and identifying sentence boundaries and section headers. (ii) Pattern-based and Gazetteer-based labeling; where each labeler performs a high precision pattern-based or list-based matching to generate PHI candidates. (iii) Machine learning-based sequence tagger; which takes in information about section headings and previously identified candidates to derive additional contextual features and learn a sequential tagging model. (iv) The final Combiner; where the labels assigned in stages (ii) and (iii) are combined based on their confidence values to generate the final result.

The machine learning-based sequence tagger is a linear-chain Conditional Random Field (CRF) model defined over five classes of features; namely (a) Word form features: features derived from a word itself such as capitalization or suffixes; (b) Part-of-speech features: the syntactic role of each word in the text string; (c) Neighborhood features: features (a) and (b) derived from words within a contextual window of size 7 for each word; (d) Regex features: pattern-matching based features to capture commonly occurring patterns for certain PHI fields; and (e) IsInX features: features that check if a word matches elements in a list or gazette for certain closed form PHI fields. For example, if a word matches both an organization and a city, then we would activate two features – say, IsInCity and IsInOrganization – and allow the CRF model to assign them appropriate weights based on other contextual features.

Results: The HyDeXT tool was evaluated on the 2016 i2b2 Research Domain Criteria dataset. The dataset contains 1000 psychiatric clinical notes annotated for all instances of PHI fields. Of the thirty different fields annotated in the clinical notes, the tool focuses on the eighteen most frequent fields. The twelve excluded fields are very rare in the dataset and jointly represent only 0.13% (n=46) of all labeled PHI fields.

We compared our system with the current state-of-the-art medical de-identification system, MIST. The micro-averaged accuracy of MIST on the cleaned text version (after stage (i)) is 0.851. HyDeXT achieves an overall accuracy of 0.898, a relative improvement of 5.5% (error reduction of 31.5%)

Discussion and Future Work: The de-identification and extraction tool, HyDeXT, presented in this abstract, contributes from three perspectives: (1) the tool is able to identify eighteen PHI fields using a machine learning-based sequential tagger combined with a pattern-based labeler, which results in a highly efficient and effective PHI de-identification system and improving over a widely used de-identification tool; (2) the individual stages of the proposed pipeline can be utilized separately, in addition to allowing for a plug-and-play architecture to support further improvements; and (3) HyDeXT provides a reliable name recognition and extraction on unseen health data, including social media data. The tool has been used to successfully identify person names and other PHI-like fields from Facebook posts. The HyDeXT tool will be made publicly available for research benchmarking and use. In future, we plan to further develop HyDeXT as a health information extraction tool. Additional work is needed to handle noise common in social media data in terms of limited data and typographic and grammatical errors.

References
2. 2016 i2b2 CEGS NGRID Shared Task and Workshop on Challenges in NLP for Clinical Data Guidelines. https://www.i2b2.org/NLP/RDoCforPsychiatry
The Use of Natural Language Processing to Identify Tdap Related Local Reaction at Five Health Care Systems in the Vaccine Safety Datalink

Chengyi Zheng, PhD¹, Wei Yu, MS¹, Fagen Xie, PhD¹, Wansu Chen, PhD¹, Cheryl Mercado, MPH¹, Lina S. Sy, MPH¹, Lei Qian, PhD¹, Sungching Glenn, MS¹, Hung Fu Tseng, PhD¹, Jonathan Duffy, MD², Lisa A Jackson, MD³, Matthew F Daley, MD⁴, Brad Crane, MS⁵, Huong McLean, PhD⁶, Steven J Jacobsen, MD PhD¹


Introduction: Local reaction is the most common vaccine related adverse event.¹ Although local reaction is generally mild, patients may seek medical attention for more severe manifestations of local reactions.² There is no specific diagnosis code for local reaction due to vaccine. Previous vaccine safety studies³ used a mixture of non-specific diagnosis codes to identify potential local reaction cases and confirmed the cases through chart review. However, due to the high incidence, manual chart review of all suspected local reaction cases would require significant resources. In this study, a Natural Language Processing (NLP) algorithm was developed and implemented to identify medically attended local reaction associated with tetanus-diphtheria-acellular pertussis (Tdap) vaccine at five integrated health care systems in the Vaccine Safety Datalink.

Methods: Presumptive cases of local reactions were identified among members ≥ 11 years of age using ICD-9 codes in outpatient, urgent care, emergency and inpatient settings in the 1-6 days following a Tdap vaccination between 2012 and 2014. The NLP system was developed at the lead study site, Kaiser Permanente Southern California (KPSC), based on NLTK, pyConText/NegEx, and Stanford NLP. The algorithm was developed based on 250 chart reviewed training cases from KPSC, and implemented at all five participating sites. A stratified validation sample of 500 patients (250 at KPSC and 250 at other sites) was randomly selected and manually chart reviewed at each site. Chart review, which served as the reference standard, was performed according to a chart abstraction manual. A quality assurance plan was implemented to ensure abstraction quality. Keywords and phrases were compiled based on published case definitions, other ontologies such as SNOMED and MedDRA, and enriched by the training data. The clinical notes containing the keywords were searched for signs and symptoms consistent with local reaction. Information on the timing and the location of a sign or symptom was also extracted to help determine whether or not the sign or symptom was vaccine related. Reactions triggered by causes other than Tdap vaccination were excluded. A confirmed case satisfied the following criteria: local reaction symptom location matched the Tdap injection site, symptom onset was not prior to vaccination, and there were no other possible causes for the symptom.

Results: There were 7,123 vaccinated patients identified with the presumptive ICD-9 codes. Only 26.8% of cases identified by ICD-9 codes were confirmed by chart review. The clinical notes generated on days 1-6 following a Tdap vaccination were processed. The NLP algorithm achieved an overall sensitivity of 86.6%, specificity of 92.3%, positive predictive value (PPV) of 80.6% and negative predictive value (NPV) of 94.9%. Compared to KPSC, lower PPV at two other KP sites and lower sensitivity at two non-KP sites were observed. Based on the error analysis of the 18 false negative and 28 false positive cases, we identified three major types of errors: data sources (11), abstraction (10), and NLP (25).

Discussion: These results suggest that NLP can reduce the efforts of manual chart review in vaccine safety studies. However, they also suggest that further refinement of the NLP algorithm should be done when applying an NLP system developed at one site to other sites. Medically attended local reaction leads to an unscheduled clinical visit which adds burden to both patients and the healthcare system. Studying vaccine related local reaction can help clinicians provide better instructions to patients regarding possible vaccine adverse events.

Heart on FHIR: Integrating Patient Generated Data into Clinical Care to Reduce 30 Day Heart Failure Readmissions (Extended Abstract)

Timothy Bergquist¹, Ronald W Buie, MTCM¹, Kevin Li¹, Pascal Brandt, MSc¹

¹University of Washington, Seattle, WA

Problem Definition

30-day readmissions are extremely costly for hospitals (fig. 1¹). Heart failure (HF) patients are particularly vulnerable following an exacerbation event, and so need to be tracked closely following discharge. However, data tend to be siloed, the patient may fail to retain or accurately interpret home care instructions, and the self-management plan may not be fully communicated to the patient. Additionally, data collected by the patient may not be accurately reported, and are often collected on paper, which makes interpretation during the clinical encounter difficult. Finally, the receipt of patient generated data often occurs through oral history taken during the clinical encounter, requiring substantial time on the part of the provider. Patient follow-up is conducted on a predefined schedule, limiting opportunities for providers to identify failure, and adjust treatment protocols.

Design Process

Starting in March 2017, our design process started with a series of team meetings to isolate the target problem. Over the next 4 months, we went through a series of common design procedures to develop a digital solution to our defined problem (fig. 17). To start our design process, we interviewed a number of cardiologists and nurses to gain a deeper understanding of the challenges that caring for heart failure patients brings. Through interviews and literature surveys we developed representative personas for whom we optimized our proposal: the patient, Mr. Liang, the home care assistant and daughter, Mrs. Liang, the patient’s doctor, Dr. Wyatt, and the cardiology nurse, Karen (fig. 2-5). A diagram is provided to better understand how these personas interact (fig. 12). Since we are developing digital technology with user interfaces, we built interface mockups to visualize how each persona would interact with the technology (fig. 6-9). Along with these mockups, we developed storyboards to develop a deeper picture of how users would interact with the application in a longitudinal fashion (fig. 13-16).

Description of the proposed solution

To address the challenges defined above, we propose the development of an integrated patient generated data (PGD) data collection platform, and an application embedded in the EHR using the FHIR and SMART technologies. An intuitive interface allows providers to select the clinically meaningful data to be collected, and to view annotated reports integrating clinical data with the collected PGD within the EHR before meeting the patient. Data collected outside of the point of care will be gathered through telematics and survey/log records. Patients will receive feedback to improve compliance and motivation and be able to view longitudinal summaries of the collected data and health events. In the following sections, we present the major components of our proposed design, including patient enrollment, data collection, communication and the various user interfaces. We also include a brief discussion of the technology choices we considered while designing Heart on FHIR. We present our designs organized by feature.

Enrollment

At discharge from the hospital, patients are given their home care instructions in the form of verbal description, paper packets with information, and medication prescription orders. Application enrollment would occur during this period so as to minimize disruption of established workflow. Instead of packets of information, the patient would be instructed on the application and how to use it, since Heart on FHIR includes care plan details. This would also

¹ All figures in supplemental materials.
facilitate enrollment at the earliest possible time, allowing for close monitoring in the crucial early days post discharge.

In order to begin using the system, a patient must complete the enrollment process. This process is initiated by the provider at the time of discharge after hospitalization resulting from an HF exacerbation event. When this happens, an account is created for the patient on the Heart on FHIR servers. Simple demographic data will be pulled from the EHR to avoid duplicate data entry. The patient will then receive an email or text message with an activation link they can use to enroll and complete the creation of their account. This enables shared decision making and empowers the patient to choose whether or not they would like to enroll in the system.

**Patient Data Collection**

At University of Washington Medical Center (UWMC), Microsoft Excel is the primary patient data collection system. It is used when a patient takes initiative to collect their own data and visualize it themselves. Most of the time, patients either don’t track their own health progress or write all their data onto paper and present this during their clinical consult. Though useful, data presented this way is very difficult to interpret. Most heart failure patients are told to track similar data, however, providers need to have the ability to customize which data is important for each individual to collect, and at what frequency. During our interviews, when asked about what data would be useful for care, a common answer from physicians was, “It depends.” It became apparent through our interviews that the concept of precision medicine applies to the collection of PGD. By allowing the provider to tailor the data collection of the application to only the clinically necessary data elements specific to the patient, the patient data collection burden and the provider information overload are both reduced.

The Chief Technology Officer of Wellpepper² reported in our interview with him that continued engagement can be driven by the personal desire for tracking/monitoring improvements. Their experience showed that patients like to visualize their own data, and as a result input it regularly. In this design, we visualize the PGD to help encourage care plan adherence. In our application, data that the patient is to collect is assigned by the physician. There are many data collection options that the application will facilitate. Only selected modules will be displayed to the patient, and will include instructional materials, customized alerts set by the doctor, and forms for data input (fig. 6).

**Data collection options**

Daily diet, blood pressure, heart rate, step count/activity, and weight directly logged into the app

- Medication compliance measured through a daily reminder/question of whether the patient took their medication. Since the app would know the medication schedule of the patient which it pulls from the host EHR, it could be setup to alert the patient when their next round of medications is supposed to be taken. Although the effectiveness of these reminders will be hard to evaluate, we hope that these reminders will simply make the patient more aware of their medication compliance.

- Survey questions and frequency are chosen by the physician to match patient needs. The provider will be able to customize their own questions, but default questions will include common health assessment questions like “Do you ever feel short of breath?”; “When do you feel short of breath?”; or “Were you able to take your medications today?”

- Health literacy assessment surveys will be implemented early in the home care process. These include questions like, “Why is tracking your weight important?”; “Do you know how to perform a six minute walk test?”; or “Do you know how to measure your heart rate?”

All data collected will be visualized as trend lines to allow the patient to have a view of their health history.

**Patient Care Team Communication**

It takes a team to care for a heart failure patient, and this team, which includes the patient, needs to communicate. Based on our stakeholder interviews, we found that this communication happens in a number of ways. The primary

---

² Wellpepper ([wellpepper.com](http://wellpepper.com)) is a commercial patient engagement platform.
way is synchronous communication in phone calls and clinic visits which is supplemented by asynchronous secure messaging via the patient portal.

The first communication event happens during enrollment, when the care plan is discussed between the care team members and made available on the patient interface. During execution of the care plan, the team communicates through a built-in secure messaging feature, which allows all team members to send direct, private messages, as well engage in group conversation. This type of communication supplements phone calls and clinic visits, and improves upon patient portal messaging by sending push notifications on new messages, as well as by supporting group chats.

**Provider Interface**

One of the main themes that we found in our interviews with cardiologists and nurses is that large amounts of time are spent on prying information from patients about their health. Physicians often described using 20 minutes of a 30 minute consults collecting and validating data from the patient. Often patients who did log their information, did it on paper, limiting the usefulness of the information due to lack of interpretability at the point of care. Inaccuracies in reported prescription adherence or symptom reporting further reduced information quality. These inefficiencies slow down the assessment by the physician. Misinformation can inhibit the ability of the physician to make an informed and effective decision about the patient’s health.

In the interviews, a theme that appeared was the need for the physician to customize what data patients collected. Not only is every bit of information not universally useful, but it may cause stress in some patients to have to collect a lot of daily information. Our clinician interface design facilitates this customization.

The provider interface would be built using SMART, a health application standard that uses FHIR (Fast Healthcare Interoperability Resources) for data access. This would allow the application to directly interact with the EHR, and the physician to open the application in the EHR window. They would be presented with a screen like figure 6, allowing them to select any patient enrolled in the system. Our application is modular, with each module managing a different data collection feature. For instance, adding the weight module will activate the weight data collection module on the patient interface, sending a reminder, at the physician defined frequency, to the patient to log their weight using the intuitive weight data entry form. Each data type can be added or removed from the patient profile as the physician sees fit.

After the patient has input data, the physician can pull up their data and trends at any time. The application would take all the input data and visualize it into a clean and easily digestible format (Fig. 6). Ideally, this would be used during a scheduled appointment and would allow the clinician to have a snapshot view of the health status fluctuations between visits. This could allow the physician to spend less of that original 20 minutes fishing for information and instead give them more time to listen to and educate the patient.

**Patient Interface**

Patients are the most important member of the care team. Given the appropriate care plan, even if the care team do their job perfectly, it’s ultimately the behaviors of the patient that determine their health. In the care of HF, these behaviors are well understood, and include medication adherence, proper diet and collecting the data requested by the provider. Monitoring this data can help catch problems early, and inform adjustments to the care plan. Our tool will enable patients to remain engaged with their care by simplifying data collection, keeping them in touch with their care team, and ensuring they understand their condition and what is required of them.

The patient’s interface into the system is via a mobile application (fig. 6a), or in the absence of a smartphone, a web interface with the same design. Since most of the HF patients are older adults, the interface should be adjustable based on user comfort level (fig. 6b). The application contains all the details of the care plan, and serves as a point of reference for the patient. The reminders and alert escalation scheme are configured in collaboration with the provider at the time of enrollment and are unobtrusive in order to avoid alert fatigue. Data entry is extremely simple, and the patient is able to monitor their own trends inside the application (fig. 6), which helps drive engagement, according to numerous stakeholders.
In addition to medication, appointment and exercise reminders, the patient will occasionally be prompted to answer questions relating to their care. This is to assess their understanding, and if necessary, they will be provided with education materials (e.g. videos) to improve their understanding. Finally, the patient will be able to communicate with the care team using the messaging feature.

**Care Partner Interface**

In many cases, patients have someone to help them manage their condition. This person could be a spouse or family member, or professional home care provider. These individuals sometimes enter data for the patient, but at least need to be able to monitor trends and be alerted when something goes wrong. The care partner also needs to be able to communicate with the care team (fig. 8). The care partner interface is essentially identical to the patient interface. The care partner is able to enter and review patient data, as well as interact with the rest of the care team using the secure messaging feature. The primary difference between the care partner interface and the patient interface is the alerting. Care partners will receive different alerts than the patient, according to the alert escalation configuration set up by the provider.

**Nurse/Care Coordinator Interface**

Nurses at the UWMC HF clinic have a high case load. Among many other things, they are responsible for scheduling appointments, making sure patients understand their care plan, and remote monitoring during the 30 days following discharge. They are also responsible for answering questions from the patient and their care partner. The current process to perform these tasks is to pull up individual patient records in the EHR, and call patients if necessary.

The nurse home screen (fig. 9) is a dashboard of the panel of patients they are responsible for, which is embedded in the EHR using SMART on FHIR to minimize workflow disruption. The screen gives an overview of the status of all patients, along with any triggered alerts and questions submitted via the messaging feature. From this screen, the nurse is able to click on any patient and view the same detail available to the cardiologist. More importantly though, they are immediately able to respond to alerts or messages without navigating away from the dashboard screen, which will decrease the amount of time it takes to complete a round of remote monitoring. The nurse will also be able to use the messaging feature to contact all members of the care team.

**Technology**

There are a number of technical factors that we considered when designing Heart on FHIR. EHRs are not designed to capture patient generated data, and therefore our solution uses the HIPAA compliant Microsoft HealthVault cloud infrastructure. This gives us the additional benefit of being able to iteratively implement application logic rapidly, without being constrained by the lack of extensibility of the Epic implementation at UWMC. However, we want to integrate as tightly as possible into existing provider workflow to minimize disruption. The SMART on FHIR (SoF) standard allows us to do exactly this, by exposing a custom web application (hosted anywhere) as a new screen within a conformant EHR. When viewed inside the EHR, the SoF specification allows the application access to data both on our servers, as well as in the EHR. Alternative approaches considered include a completely separate standalone application, as well as an Epic customization. The former would disrupt workflow and require special access to FHIR APIs, while the latter would require development resources and timelines beyond our control, and the developed solution would not be translatable to other settings.

HIPAA compliance is always a concern when dealing with patient data, and in order to ensure security and privacy, Heart on FHIR application servers will be hosted using the HIPAA approved Microsoft Azure cloud infrastructure, and use industry standard transport layer security for all API requests. Additionally, any data access will require a secure password, and full audit trails will be maintained.
Implementation

From a technical perspective, to enable adoption of Heart on FHIR at scale, it would be necessary to publish the application in vendor SoF app stores, like Epic’s App Orchard\(^3\). This would allow “one click” installation analogous to installing an app on a smartphone. It may be necessary to charge a fee or subscription for the application to cover cloud infrastructure costs. From an organizational perspective, to gain acceptance from hospital executives and IT departments, we would be required to demonstrate HIPAA compliance by providing our Azure HIPAA compliance certification, as well as our documented standard operating procedures. Acceptance by clinical staff would be gained by having a well designed application that does not disrupt workflow and provides value by having a positive impact on clinical care. We have collaborated with clinical staff during our design phases to achieve these goals.

Alternative Solutions

One-size-fits-all

A more basic data visualization tool that does not accept physician configuration. Essentially an extension of the EHR, this simplifies use by the provider, but does not necessarily fit the customized approach of each care plan. We decided we could design a simplified interface and avoid this route.

Machine learned feedback and plan generation

The data we collect has the potential to be automatically matched to care plans based on protocol, effectively automating the assignment and variation of the care plan without provider input. After some discussion with providers, it was clear that buy in would not be universal, and proofing such a system would require its own project.

Research and Evaluation

Our evaluation plan attempts to address the following questions:

1. In what ways does the actual workflow of our users vary from expected workflows?
2. Does our intervention reduce encounter times?
3. Is use of our intervention associated with a difference in 30 day readmission compared to usual care?
4. How does our intervention impact satisfaction experienced by clinical staff managing HF patients?
5. How does our intervention impact satisfaction experienced by HF patients?
6. What are the barriers to technology acceptance by our users?

The first question will be addressed in qualitative interviews with each member of the care team. These structured interviews will ask participants to recount their actual usage along many scenarios, which will then be compared to the intended workflow. In order to address the second question, we will survey cardiologists who manage HF patients at the UWMCH before and after the intervention.

Our 3rd research question can be tested by comparing outcomes in our intervention and control groups, controlling for provider and severity of the condition. Although it would be best to randomize both patients and providers, this may not be possible at UWMCH due to HF clinic staffing and operations. Our 4th and 5th aims will be addressed by a pre and post intervention survey. The instrument will attempt to assess success as defined in the DeLone and McLean Information System (IS) Success Model (fig. 10). This model attempts to capture the relationship between the quality of the IS and its information, and the user’s satisfaction and capacity to realize benefit. Our 6th aim can be assessed during the same surveys as the prior two aims. However, questions will be based on the technology acceptance model (fig. 11), and focused on capturing both technical and experiential barriers to acceptance and use of the IS. Taken together, our evaluation should provide insight for future design iterations, as well as preliminary approaches for studying effectiveness and efficacy in reducing 30 day readmissions and, ultimately, the efficiency and accuracy in tracking heart failure patients throughout the course of illness.

\(^3\) https://apporchard.epic.com/
Voice Enabled Framework to Support Post-Surgical Discharge Monitoring
Kevin Blansit, M.S. 1, Rebecca Marmor, M.D. 1, 2, Beiqun Zhao, M.D. 1, 2, Dan Tien, M.P.H. 3

1 Department of Biomedical Informatics
University of California San Diego
La Jolla Ca, 92093

2 Department of Surgery
University of California San Diego
La Jolla Ca, 92093

3 Department of Health Policy and Management
University of California Los Angeles,
Los Angeles Ca, 90095

Abstract:
Unplanned surgical readmissions pose a challenging problem for the American healthcare system. We propose to combine consumer electronic voice recognition technology with the FHIR standard to create a post-surgical discharge monitoring app to identify and alert physicians to a patient’s deteriorating status.

Introduction:
Of the nearly 50 Million patients that receive surgical care every year, almost one in five of them will have an unplanned readmission to the hospital[1]. Despite many of these readmissions being preventable, these events cost the American health system over $17 Billion in the Medicare population alone[2]. In addition to these high costs, surgical readmissions have been previously investigated by federal regulators as a potential quality metric for reimbursement[3]. As the healthcare system attempts to realign with these changing reimbursement models that emphasize health outcomes, hospitals have struggled with implementing strategies to reduce these preventable readmissions. Previous work has shown that additional nurse follow-up and monitoring improves surgical outcomes and prevent unplanned readmission in prostatectomy, total joint, and cardiac surgeries[4-6]. While these solutions have shown promising results, the care model requires a specialized nursing staff, limiting the scalability of these care units across multiple service lines. In response to these limitations, health researchers have recently investigated the potential to utilize patient generated data (PGD) to extend care monitoring in the post-surgical care setting[7].
Despite an acute need for PGD to be integrated into the electronic healthcare record (EHR), interoperability and secure authentication methodologies remain an active area of design and research[8]. Specifically, Mandel, et al. highlighted these challenges in the development of the pivotal SMART on FHIR (Fast Healthcare Interoperability Resource) framework which brought standardized authentication methods to FHIR specifications[9]. However, to date there has been scant utilization of these resources to bridge the digital divide between patient home care and the EHR.

Despite comprising only 13% of the overall population, an estimated third of the aggregate national health care spending is concentrated among the elderly. It is therefore imperative for patient-focused devices that enable PGD EHR integration to consider that a significant proportion of users may not be as technologically adept to newer technologies. Previously, detailed user interviews have highlighted the potential benefits of using voice-enabled frameworks to interface with this population.

Given the goal of reducing unplanned surgical readmission, we propose a voice enabled framework to monitor these patients in the post-discharge setting using consumer internet connected smart home devices. Using the FHIR based standards, we aim to provide EHR connectivity to structured patient interviews performed by the smart home device, prompting clinician in potential medical crises.

METHODS:

Colorectal surgery was identified by the collaborating surgical residents as a primary clinical use case given the high degree of post-surgical intervention for this patient population. User screening questions and potential question responses were developed. These were then compiled into a dialogue tree, which was used to inform the application state machine design. Additional clinical applications could be added with unique dialogue trees to reflect the heterogeneity of surgical complications.

We considered all major consumer-grade home voice-command assistive device platforms including Apple Siri, Google Home, Microsoft Cortana, and Amazon Alexa. Given the strong market share and wide range of developer tools necessary for rapid prototyping, the Amazon Alexa (and associated Voice Skills application framework) was chosen[10]. This framework provides developer specified intents (invocation programming triggers) and utterances (mapping to user dialogue), which are mapped together in the server application. Incoming intents trigger the Amazon Voice Service server to send a JSON
POST request to the application server. Using the flask-ask extension for the flask python microframework (https://github.com/johnwheeler/flask-ask), these intents are mapped to state dependent functionality, providing context aware and seemingly ‘intelligent’ communication abilities.

FHIR was chosen given its conformance to modern REST-like messaging standards and providing HTTP verb functionality such as ‘GET’ and ‘POST’. With an appropriately authorized oauth token, secure and interoperable transmission of patient information to the application server is enabled.

Results:

The proposed schematic is displays the technical and interactive details using this framework (Supplemental materials). (Step 1) When a patient is being discharged, the surgeon would select a relevant surgical indication that would specify the type of questions the framework would ask. (Step 2) Upon the patient’s activation of the Amazon Alexa app, he would be prompted to login through their patient portal credentials. This would create a session level oauth token allowing the patient server to request their medical information from a FHIR enabled server. Using the Encounter FHIR resource specifying the patient’s surgery from an appropriate ‘GET’ request search, surgery-specific screening questions are then mapped to the current session

(Step 3) The application then confirms that this application is okay to proceed, and asks birth date information to further verify the identity of the patient. Failure to appropriately authorized would result in quitting of the application and deauthorization of the token. After each question is answered, it is similarly saved to the application until the final question is answered. The patient is then asked to confirm any abnormalities that were found during the screening session before alerting the clinician. (Step 4) The screening application then structures all recorded answers as a FHIR’s ‘QuestionnaireResponse’ resource, and ‘POSTS’ these data to the hospital’s FHIR enabled server. (Step 5) Once these data are integrated into the hospital’s EHR, the attending surgeon’s email would receive a starred alert message contained structured results of the patient interaction, prompting the doctor to follow up with a clinical care team to further investigate the patient.
Source Code for the alpha prototype is available from https://github.com/KBlansit/Alexa-Discharge-Monitoring/tree/master. The video describing the patient interaction is located at https://youtu.be/05MgJHWA-ns.

Conclusions:

Developing new modalities to extend care monitoring to the home is a current and active area of research. In this project, we elected to utilize a consumer home assistant device. While using this platform permitted the focus of development on the underlying server application logic, it limited the scope and capabilities of the project to curated interactions, which may not fully cover the landscape of potential medical needs from patients. Since this platform uses both consumer technology and interoperable health web standards, there is high potential for scalability of this framework to reduce unplanned surgical readmissions. Future efforts should seek to formally evaluate the potential of this framework to aid in reducing unplanned surgical admissions and increase patient recovery self-efficacy. Partnership with a surgical department to pilot these efforts would be necessary to aid in the study coordination. Further considerations must be given the integration into each medical site’s EHR as robust APIs from vendors do not come standardized.

References:


10. eMarketer. Amazon controls 70% of the voice-enabled speaker device market
Home Behavior Monitoring Module in OpenEMR: Use of home sensors as Patient-Generated Data (PGD) for elderly care

Regina A Casanova-Perez, BSc1, Pierre G Padilla-Huamanitnco, BSc1, Catharine I De Freitas-Vidal, MBBS1, Yong K Choi, MPH2

1Biomedical Informatics in Global Health Research Unit, School of Public Health and Administration, Universidad Peruana Cayetano Heredia, Lima, Peru; 2Department of Biomedical Informatics and Medical Education, School of Medicine, University of Washington, Seattle, WA, USA

Definition of the selected challenge related to leveraging of patient-generated data for improving patient care grounded in deep understanding of an identified health problem

People older than 60 years old represent the twelve percent of world population, and the treatment of elder-related diseases represents an important percentage of health expenditure. The amount of older adults is growing every year due the better quality of life and the advancement in medical technologies. This phenomenon occurs on high and low-and-middle income countries predicting that, by 2030, elder population will be more than 25 percent of North American population, and 17 percent of Latin America and the Caribbean people.

Many diseases, such as type 2 diabetes mellitus, high blood pressure, coronary artery disease and neurologic disorders are prevalent among older adults, and, for many of them, physical activity (PA) is a cornerstone in their management and prevention. PA has shown to diminish mortality and improve quality of life in older adults with such diseases. By its own, a great amount of low level energy expenditure activities during the day, increases the mortality up to 20% according to a systematic review.

Assessing physical activity in an older adult is challenging due to several factors. First, defining how active an older adult is will depend on the comorbidities that an older adult may have such as musculoskeletal or neurological impairment. Additionally, self-reports of physical activity rely on how well they can recall, which is also influenced by their older age. Studies have shown that older adults tend to report less sitting time compared to sitting time measured objectively. Despite the importance of the measurement of the level of physical activity and sedentary behaviors, patient-generated data (PGD) related to physical activity has not yet been integrated to the patients’ Electronic Health Record (EHR).

The rapid development and proliferation of digital devices, apps, and services made it possible to capture and record different parameters that are relevant to human health. The advances in data processing and storage, communication infrastructures and networks, and the increased capacity of batteries allowed for continuous tracking and monitoring of various parameters regarding people’s health and illness. The use of smart home sensors to detect and record individuals’ activities and behavior patterns is one such example. In addition, we can also track environmental factors such as temperature, humidity and luminosity reports, or on/off control data for the lamp. The information from these home-based sensor technologies make it possible to learn about different patterns associated with a particular behavior and the health status of people.

Home-based sensors have been implemented in several previous projects. However, the previous efforts have focused on producing and displaying the sensor measurements to patients and have not looked into how this information should be reported to the health providers. Therefore, we propose to implement a new module into an EHR that will allow the follow up of the home environment and the physical activity in older adults through home-based sensors. This module will let health providers know in an objective way the characteristics of home behavior of patients ranging from home characteristics to indoor physical activity.
Description of the proposed solution

Our solution concentrates on the system OpenEMR (http://www.open-emr.org/) which is an open source Electronic Health Record (EHR) for international use. As the team is composed of 4 graduate students, 3 living and studying in Peru and 1 living and studying in the US, we wanted to focus on a possible solution that can be used in both countries. In Peru the EHR is not yet used because the regulations are not yet finished, but its implementation is imminent for the future, so we are getting ahead with this proposal showing how an EHR could integrate the Patient-Generated Data (PGD); and in the US case, we are showing just an improvement of the EHR with a new Module.

For our proposal, we plan to collect data generated by the user in their home environment using the home-based sensors specifically that can track movement, open/close signal from door and windows, humidity, temperature and luminosity. With the use of these sensors we plan to know the behavior patterns of an older adult, so that timely interventions can be given when a sudden or peculiar change in pattern have been detected. The data could also be used to identify slow, gradual changes of their habits by making comparisons with the historical data generated longitudinally.

In this solution, we show the PGD in a new tab inside each patient record, specifically we added one option just below the patient’s name called ‘Patient Data’ where the physician would find easy-to-read charts about the patient’s motion around their house, the luminosity, humidity and temperature of their house and the data about when the doors and windows have been opened or closed. With this information, the physician can make a medical decision regarding the patient’s health with the use of this charts. For example, if one patient has a lung condition, decreasing or rising the humidity or temperature can help to improve the patient’s lifestyle. Or a patient that has been through orthopedic surgery need to decrease their activity levels within their house so that it can heal properly. All those small and non-intrusive sensors can help physicians keep track of their older patients, so that they can have better quality of life. For the part of showing the charts, we provided three different solutions of how the information would be arranged within the EHR. Also, a summary of the main data would be available for every patient, this would help the physician to make decisions about whether they need to review the data from each sensor or if the older adults continues taking care of their health as before. For this part, only has one solution and it would be shown as the home page of the ‘Patient Data’ tab.

When the patient arrives to the doctor’s office, the doctor would look up the summary report of the ‘Patient Data’ tab, in this report we tried to generate some interpretations of possible health indicators that may be of interest to the physicians, this indicators will show the amount of time the patient is currently sleeping at night, the average time that they are spending outdoors and how much time they are spending doing low level energy expenditure activities, such as reading or watching tv. With these indicators, the physician can have a rough idea of how well the patient is doing regarding their mobility; even so, physicians will be able to look at the specific charts if they need more information before giving any recommendation. We expect that, by looking at the summary, physicians will be able to obtain objective measure of patient’s home behavior, making the interaction go more fluid and giving the physician more time to concentrate on other useful questions. With this information, we expect that both, physician and patient, will agree on the best and realistic PA goal based on accurate measurement of the patient’s activity patterns.

Discussion of alternative solutions considered

The evolution of mobile technology has changed different aspects of our lives. Data and knowledge are more affordable and manageable. It happens because digital technology allows us not only to carry or place devices but also wear them. Wearable technology has potential applications in healthcare, especially in patient monitoring. Some reports say that this technology would make health systems more efficient and responsive than the current ones17. Unfortunately, these devices are a challenge for elder people. If the wearable device presents a complexity in attachments or management (e.g., limited battery life), older patients may be unable to carry it continuously during their daily lives18. Some studies show that the interventions with wearable technologies work fine under supervision because it is hard for elder people to understand the user interfaces19.
Another solution considered was a hand-written activity journal where patients can log their daily activities such as how many hours they have slept or how many hours they have gone out. This type of journal relies heavily on the patients’ memory and their responsibility to log the activities in timely manner before they forget. This solution is not so much different from the feedback the patients give to their doctors during their appointments. As we are trying to objectively measure the patients’ activity levels and the environmental characteristics of where they live, the manual activity logging would not be an ideal solution.

**Discussion of the strengths and weakness of the chosen solution as compared to the alternatives.**

Our proposal seeks to integrate into the Electronic Health Record (EHR) new information from sensors installed in patients’ homes. These sensors make the elder people a data source and follow the approach of Internet of Things (IoT). We decided to integrate this information because patients produce many data day by day, and the IoT revolution is redefining modern healthcare. Sensors can track and monitor people 24/7 in any environment. It allows that any information related to healthcare (logistics, diagnosis, therapy, recovery, medication, administration, finance and daily activity) can be stored, managed and shared. Getting the maximum benefits from this data means having a better understanding of people’s health. We just need to use it properly and distribute it efficiently.

Setting up smart homes solutions is not usually affordable, the cost is the main reason. Many people would live at smart homes if the transformation of their environments were not expensive. Open source solutions become more relevant to face this challenge while they promote the integration of new data for healthcare. Our solution will use OpenEMR an open source EHR platform because we plan to implement it in a high-income country (US) and LMIC country (Peru). The main advantages of open source are accessibility, flexibility, and documentation. Open source solutions are gaining more traction in the scientific community by the evidence of superior technical levels and less expensive than proprietary solutions.

**Proposed implementation and dissemination plan**

For implementation, previously we need to identify and involve the principal actors from the health system in both countries. Companies like Google and its Health service failed because they wanted to implement the service without providing added values to the stakeholders of the US health system. Then, we must select intervention sites like patient homes and health centers. The selected institutions should already have OpenEMR. The proposal does not pretend to create a new workflow or modify the existing IT infrastructure. In the case of patients, they have to be older adults who live alone in their homes. Patients should have agreed to be installed a set of sensors before the implementation. They will receive an informed consent where we mention the study protocol, the privacy, and potential security risks of using the sensor devices to collect the PGD. The acquisition and implementation should be assigned to IT area from health centers. The information about installation and local providers would be available on a Wiki page. We will ensure the providers give a special discount to health centers if they want to start with a pilot. The project will offer training in home sensors installation.

**Implementation and Integration of New Module**

With this implementation, three stakeholders of the medical system would benefit. The care providers would benefit because they would make better medical decisions with better understanding of the patient’s health, without having to rely only on the patient’s self-report of their physical activity. The hospitals or clinics would benefit as a result of the better medical decisions that the doctors would make, making the management easier because the resources and equipment won’t be used in an inadequate way. Finally, the patients would benefit because it would be easier for them to identify their patterns in their daily lives by seeing the charts and won’t need to stress about remembering about their habits. This add-on to the OpenEMR system would be easy to install and won’t need nothing more than an update, all the data collected by the sensors, would be daily updated to the servers via internet connection so the doctors would have updated data.
Home Sensors Description

The set of sensors selected is from Arduino IoT platform. Arduino IoT guides anyone who wants to get started tinkering with the Internet of Things. The sensors will measure parameters like movement, temperature, humidity, luminosity, and open/close doors and windows. These devices must have strategic positions if patients have pets or frequent guests. Interoperability plays a major role in ensuring a better adoption. Our proposal follows recommendations from the Continua Health Alliance (CHA), a major standardization body working on device level interoperability, which has recommended to use standards like the Bluetooth Health Device Profiles and USB Personal Healthcare Device Profile. These formats have been specified by the ISO/IEEE 11073 family.

Dissemination plan

Open source lets people replicate, modify and upgrade different technologies. Our purpose focus on creating a community in Public Health and more institutions can add new data for giving patients a better service. The project keeps the same license from OpenEMR, the documentation and a tutorial will be available on the Wiki page. The team will ensure these instructions may be read by someone whose expertise is different from engineering. Github will host the code and firmware required to implement the project. Also, a publicity campaign would be used both in hospitals and in residential homes where elderly patients are known to live alone, this would help us recruit the subjects to install the sensor in their homes.

Proposed evaluation plan

After the implementation, we will conduct a usability study with the participant physicians. Physicians that will receive the implementation of the module will be those in charge of patients that already have sensors deployed within the home, which may vary depending on preference of the patient.

All the physicians involved in the implementation will be invited to participate in the usability study. It will consist of semi-structured interviews asking participants to give feedback on the presented designs and to comment on which sensor data visualization was considered more useful, and to generate ideas for new alternatives to the visualization.

Once subjects have agreed to participate in the study and have signed the consent form, we will collect socio-demographic data such as age, gender and time of experience. Then, we will conduct an interview with the participant that will take 30-45 minutes, addressing the following questions:

- We would like to hear about your general thoughts on the new module of the EHR.
- How do you think the new module in the EHR plays a role in your clinical practice, if at all?
  - What are some benefits that you might find in using the new module in your clinical practice?
  - What are some drawbacks that you might have with using the new module in your clinical practice?
- Here we have some of the charts that you have been working with in the new module. What are your thoughts on the way they are displayed?
  - What do you find interesting about the displays?
  - What do you find confusing about the displays?
  - How might you use these displays, if at all?
  - What indicators would you like as feedback on the visual display?
  - How useful were the indicators presented to you in the new module, if at all? What kind of clinical recommendation did you give based on the information reported in the new module, if at all?
  - What suggestions would you have for improving the displays?
  - Are there alternative ways in which you would like the sensor information to be presented?
  - How frequently did you look at this information in your clinical practice?
- Is there anything else that you would like to add regarding the new module?
References

Abstract

Cardiovascular disease and diabetes are epidemic in the United States, and efforts to shift this trend have been largely ineffective. The greatest challenge that health care practitioners face is inspiring the lifestyle changes necessary to prevent or reverse these conditions. New evidence suggests that minimal activity, such as simply standing up periodically and moving around can reduce biomarkers of cardiovascular disease and diabetes. Given the challenge and temporary nature of inspiring habitually sedentary individuals to take on intensive exercise routines, this is an exciting prospect. With this new information, the question becomes: “How do we inspire individuals and populations to get up and move?” Dr. Matthew Buman and his group at Arizona State University are addressing this question through researching methods to inspire individuals and groups to stand and move at work. In support of Dr. Buman’s work, we have leveraged subject generated postural data gathered by his group to create a pipeline that processes and analyzes the patterns of subject movement with the prompts they received in order to identify the most effective prompt that elicits standing and moving behavior. The pipeline helps researchers in his group visualize their data in an interactive way and to help inform statistical analyses. In future directions, this pipeline structure can be adopted by various aspects of clinical work such as diagnosis, selection of treatment options, monitoring for changes in a patient’s conditions over time, evaluating efficacy of different treatment options, and promoting shared decision-making between providers and patients. They can range from novel ways to visualize PGD to help clinicians and patients identify important and actionable trends, to novel computational solutions, to using these data together with the traditional EHR data to provide clinical decision support that may or may not include visual presentations. Importantly, in this challenge, our focus is on integrating PGD with EHR to improve care within the clinical context, rather than on using these data outside of traditional care settings to promote health and wellness.

Introduction

According to the American Heart Association, cardiovascular disease (CVD) is the leading cause of death for both men and women in the United States (1), and there is ample evidence to support a link between CVD and diabetes (2). At least 68 percent of people age 65 or older with diabetes die from some form of CVD, and adults with diabetes are two to four times more likely to die from heart disease than adults without diabetes. This major health crisis can be linked, in great part, to an increasingly sedentary lifestyle in the US population (3).

Moderate-to-vigorous physical activities are effective in reducing biomarkers of CVD (4), but it is very difficult to inspire habitually sedentary individuals to take on a regular practice of moderate to vigorous physical activity. Matthew Buman et al., 2014, found that for sedentary individuals, periodically replacing sedentary activity with light-intensity activity such as standing up and moving around significantly lowered biomarkers of CVD and diabetes (5,6). Subsequently, Dr. Buman and his group set out to create a mechanism by which they could encourage large groups of individuals who are sedentary (sitting) during the work day to stand or stand and move at periodic intervals.

Using patient generated data (PGD) gathered by Dr. Buman and his team, we created “Stand and Move at Work” (the name given to the project by Dr. Buman’s group), a pipeline that analyzes PGD from a device that detects sitting behavior, standing behavior, and stepping behavior, and establishes which of several prompts most effectively inspires individuals and groups to stand and move at intervals during their work day. Our pipeline includes visualization tools for postural data for individuals and selected groups that can be linked to an electronic health record to assist researchers and practitioners in a variety of fields such as occupational health, wellness, and nutrition.

Design Process
Dr. Buman’s group provided us with files containing postural data collected over a four-week period on subjects working in the transportation department at Arizona State University (ASU). All subjects were given a sit/stand desk a few weeks prior to the start of the data collection. Their postural information was collected using an activePAL™ device which is a proprietary instrument worn on a subject’s 18 subjects during the work-week during the time frame of the study. We were also given a file containing prompts that were sent to subjects by email each hour throughout the work day. The prompts were divided into three assessment groups: prompts that tell individuals to stand vs prompts that tell individuals to move, prompts that contain an authoritative tone vs those that do not, and prompts that include a goal vs those that do not.

**Problem Clarification:** Dr. Buman’s group is hoping to use this data to determine if sending subjects prompts over email will encourage them to stand and move at work. Further, they want to know which prompts were most effective. Our job was to create a reproducible pipeline that uses the data we were given to determine which subset of prompts most effectively influenced the subjects to stand and/or move.

We carried out three meetings over Skype with the ASU group. Our first meeting was designed to clarify their study design and pinpoint the portion of the study we were to address. During our second meeting we clarified some data cleaning issues and they gave us more information on the type of data visualization and analyses they needed in order to help them understand the data and outcomes. During our third meeting we discussed ways in which we could extend our pipeline to include more analyses.

**Pipeline Design:** Our pipeline was created using the R statistical programming language. We also used the R package, “Shiny”, which is a platform used to translate R to HTML code in order for users to interact with the data visualizations through a web interface. This method is especially important when there is a high volume of data because it makes it faster and easier for researchers to view the data. The Shiny application is launched through a local server at OHSU. We uploaded the activPal™ events files provided to us by the ASU group. These files contain information about changes in postural position (sitting, standing, or moving) associated with the dates and times that these changes occurred. We also uploaded a prompts file containing a list of prompts, such as “Try to stand for the next 5 minutes”, which were queued every hour during the workday for the length of the study, 50% of which were emailed to subjects and 50% percent were not emailed (this allowed for a comparison between activity following a prompt and activity with no prompt). There were a total of eight different prompts. The prompts file also noted which subject the prompts were sent to, and what date and time they were opened. Our pipeline merges these files together resulting in a dataset that allows us to calculate the time interval between when a prompt was opened by a subject and when a positive action (standing to standing, or standing to moving) took place.

Our pipeline performs a non-parametric test, the Kruskal-Wallis rank sum test, on log transformed data to assess whether there are significant differences in the time interval described above relative to each of the eight prompts. We also used a Chi-squared test to determine whether significant differences existed in the proportion of subjects that responded positively to each prompt. There are various normalization methods available in the Shiny app.

We conducted a secondary analysis that observed the subject’s sitting and standing behavior within the first half hour each prompt was sent. From this data, we computed statistical analyses to see whether the average time spent standing or moving was significantly higher in prompt versus no prompt events, goal versus no goal prompts, authoritative versus non-authoritative, and stand versus move prompts. We found the average time of standing was significantly higher when subjects received a prompt compared to when subjects did not receive a prompt (p-value=0.03).

**Alternative Solutions Considered**

We considered the possibility of creating an application that would ring or chime periodically throughout the day to remind subjects that they should get up and move. After discussing this idea with Dr. Buman, we learned that these kinds of applications had been tested previously and found to be ineffective.

**Results:**

Our pipeline determined that there was no statistical difference between the interval of time it took the subjects to positively respond to each of the eight prompts.
Figure 1. Image of the first tab of the data visualization app showing the summary by week and by day for each subject across the study. Users can select specific starting and end dates and download this data as a .csv file in order to flag specific days.

Figure 2. Image of activePAL™ events file data visualization page from our Shiny application showing the percentage of time each subject spent in each of the three postural activities detected by the activPAL™ device.

Figure 3. Shiny application visualization of the distribution of time to positive action for each prompt and percentage of subjects that positively responded to each prompt.
Use Case:

Shared provider-patient Decision Making and Novel Data Visualization

Mr. Smith works in the transportation department of local university and spends most of his work day sitting for long periods of time. He is overweight and has developed diabetes. He is here for routine follow-up in your primary care clinic. He tells you he has participated in an employer sponsored project that looks at workplace interventions that could improve his diabetes control. He brings in data from an activity tracker that he has been wearing. He explains that based on preliminary results evaluated for him the nurse running this employer-sponsored project, he was assigned to receive an email alert that said “Try to stand for the next 5 minutes” every hour during his work day. He proudly shows you his data that indicates he had increased his standing/moving time at work by 45 minutes. You bring up the HgA1C he had drawn before his clinic visit and find it has decreased significantly, but it is still slightly elevated. Together, you review his latest HgA1C and activity data and suggest increasing his daytime activity by another 20 minutes. He goes back to his employer sponsored project and requests that they increase the frequency of his reminder to stand/move alerts.

Feedback and Evaluation Plan:

We created a survey that was distributed to Dr. Buman and the researchers who are working on this project. Our questions were focused on the usability of our application within an electronic health record in a clinical/patient care setting and in an occupational/employee wellness setting. The overarching response was that being able to visualize an individual’s physical activity and sedentary behavior is a valuable tool for providers which will allow them to co-create personalized interventions that are geared toward disease prevention and health promotion.

Our survey also included a question regarding what could be added to the visualization to support shared decision making for providers and patients or employees. One suggestion for future application of this tool would be to have it provide a recommendation for health improvement based on the individual’s current level of activity and health promotion goals.
We plan to further our evaluation by using our survey to gain feedback from providers in the health care and occupational health settings. We would also like to test our pipeline and visualization on a new set of data provided by a group in a different occupational setting.

Summary:

We have created a pipeline that we believe can be adapted for use by any group interested in using activePAL™ events data to examine intervals between interventions and positive changes in postural behavior. This data can be incorporated into both occupational and individual electronic health records, and can be used in partnership with patients to determine the best interventions to improve their health. More broadly, this data can also be used to guide employer health programs.

Special thanks to graduate student Lawrence Hsu for initial input and support for this project.

References


2. http://www.heart.org/HEARTORG/Conditions/More/Diabetes/WhyDiabetesMatters/Cardiovascular-Disease-Diabetes_UCM_313865_Article.jsp/#.WUHqYTPznPA


Visualizing the Patient-Reported Outcomes Measurement Information System (PROMIS) Measures for Clinicians and Patients

Lisa V. Grossman1,2*, Elliot G. Mitchell1*

*These authors contributed equally to this work
1Department of Biomedical Informatics, Columbia University, New York, NY
2College of Physicians and Surgeons, Columbia University, New York, NY

Abstract

Congruent with the nationwide movement toward patient-centered healthcare, an increasing number of organizations collect and assess patient-reported outcomes (PROs). The standardized NIH PROMIS measures represent one of the most widely used PRO questionnaires, but organizations still face challenges with conveying PROMIS outcomes to clinicians in clinically relevant ways. Our proposed solution, the ProVis application, uses visualizations to engage heart failure patients with PROMIS questionnaires in the waiting room, and conveys PROMIS data to clinicians through longitudinal visualizations in iNYP, our institution's electronic health record (EHR) interface. Here, we discuss the design and development of ProVis, the alternative strategies we considered, the strengths and weaknesses of ProVis, and our future dissemination and evaluation plans.

Background and Definition of the Challenge

Patient-reported outcomes (PROs) constitute an important subset of patient-generated health data. When collecting PROs, healthcare organizations use standardized survey questions to understand patient's experiences, and the data may inform each patient's care plan individually. Recently, the National Institutes of Health (NIH) created the Patient-Reported Outcomes Measurement Information System (PROMIS) measures.1,2 PROMIS is a validated and standardized set of PRO measures applicable to a range of chronic conditions. The PROMIS scales map directly to PRO measures such as the PHQ-9 depression questionnaire. As such, PROMIS measures occupy an important position at the center of national "Meaningful Use" efforts to incorporate PROs into patient care. Epic Systems now offers PROMIS measures through the MyChart patient portal, and the NIH's EASI-PRO initiative promotes integration of PROMIS into major electronic health records systems, including Epic and Cerner.

Despite PROMIS's success, challenges remain. An ideal PRO measure is efficient (minimizes the number of items without compromising reliability) and precise (minimizes error in estimate).2 Despite efforts to maximize efficiency of individual scales, clinically useful combinations of PROMIS scales may take too long (>10 minutes) for older patients and patients with low technology literacy. Despite rigorous effort to increase precision, patients with a lower literacy level may demonstrate lower precision.3 Such limitations may decrease clinician's confidence in PROMIS outcomes, especially clinicians who work with disadvantaged populations. One untested solution is to visualize each PROMIS item's answer choices to help patients select answers more precisely and efficiently.

Another challenge is defining appropriate cutoff points for clinical action based on PROs. Indicating the patient-reported symptom's severity is necessary, but insufficient. Clinician-facing systems must also report changes in each symptom's severity. For example, a patient with longstanding severe depression known to a psychiatrist is less clinically concerning than a patient with newly-reported depression that requires a psychiatric consult. Currently, physicians use clinical notes to track symptom changes over time, a practice that presents challenges due to copy-paste conventions and the time needed to re-read the note. Overcoming these challenges will require better presentation of longitudinal PROs in the EHR to support clinical decision making.

In collaboration with the Columbia University Center for Advanced Cardiac Care at NewYork-Presbyterian Hospital (NYP), our proposed solution is to 1) visualize PROMIS answer choices for heart failure patients to increase PROMIS's precision and efficiency, especially in disadvantaged populations, 2) incorporate these visualizations into an application, ProVis, for heart failure patients to report PROMIS measures from the waiting room, and 3) visualize ProVis data for cardiologists in the EHR, in a manner that predicates clinical action and facilitates longitudinal symptom tracking. We use the heart failure population to demonstrate our application's utility for multiple reasons.
Heart failure affects more than 5.7 million Americans and is the leading cause of 30-day hospital readmissions in the United States.4 Ineffective symptom recognition and management is the primary cause of repeated hospital readmissions in heart failure.5 Clinicians typically spend the majority of the outpatient visit discussing symptoms, although time constraints may prevent the full elucidation of clinically relevant symptoms. As such, PROMIS measures hold great potential to transform clinical decision making in heart failure care.

**Proposed Solution**

At this time, the patient-facing component of ProVis is fully developed and communicates with the EHR, and the interface for the clinician-facing component is designed. During the design process for ProVis, we applied techniques from biomedical informatics (see Integration with the EHR section), human-computer interaction (see Patient-Facing Interface section), and information visualization (see Clinician-Facing Interface section) to enhance usefulness and usability. To integrate clinical expertise, we conducted participatory design sessions with two clinicians to ensure each visualization's clinical relevance, including card sorting to discover clinicians' mental models of heart failure symptoms. To integrate patient expertise, we refined the application after usability testing with 6 hospitalized patients (poster accepted to 2017 AMIA proceedings).

**Conceptual Framework:** We used the updated technology acceptance model (TAM) to guide the development of the patient-facing interface.6 TAM is a validated model that considers how users come to accept and use technology. Figure 1 displays the TAM, and indicates which factors the 4 patient-facing features of ProVis specifically support.

![Figure 1. Technology Acceptance Factors](image)

**Use Case Scenario:** When the heart failure patient arrives, the receptionist enters the patient's MRN into the ProVis application, installed on an iPad (see video demonstration). The MRN triggers ProVis to display the patient's identifying information. The receptionist hands the iPad to the patient, and the patient verifies their identifying information. Then, the patient completes the PROMIS measures and views their summary visualization on ProVis while in the waiting room. When the patient presses the "submit" button, the application prompts the patient to return the iPad to the receptionist. Then, provider and patient review the PROMIS outcomes together, in conjunction with medication data, through the EHR interface, during the actual clinic visit.

**Innovation:** As the use case scenario describes, the proposed solution: (1) integrates PROMIS measures into the EHR with medication data to inform clinical decision making, (2) presents PROMIS outcomes to patients and clinicians using novel visualizations, (3) applies PROs to a population, heart failure, where symptom management is critical, and (4) facilitates communication between the patient and provider during the clinic visit.

**Patient-Facing Design Process:** We applied material design to create an industry-standard interface for use in populations with low technology literacy (see Supplementary Materials). Launched in 2014, Google's material design is among the most widely accepted design frameworks. Advantages over previously popular minimalist frameworks include three-dimensional layering effects which enhance system usability for older users, as well as updated design patterns. A design pattern is a highly-tested solution to a commonly occurring problem in interface design. ProVis uses a responsive mobile-first framework, Bootstrap 3, for viewing on various screen sizes, including both the vertical and horizontal iPad orientations. To improve the interface's learnability and memorability for older adults, we use clearly identifiable buttons, progress bars, consistent linear navigation, large font, colorful icons, high contrast, and system response times tailored to older users. To increase precision and efficiency, we coupled previously published infographics tested in low health literacy populations with PROMIS answer choices.7 Additionally, we created visualizations to summarize PROMIS outcomes for the patient directly after survey completion. In creating patient-facing summary visualizations, we provide immediate feedback to patients, engage them in their care, and highlight the survey's usefulness to them. Within the summary visualization, we used locally developed materials to educate patients about their symptoms and enhance their situational awareness.
Integration with the EHR: The NewYork-Presbyterian (NYP) hospital enterprise uses multiple EHR vendors, including Epic, Cerner, Athena, and Allscripts. Instead of working with each vendor separately, we will integrate our clinician-facing visualizations into iNYP, an EHR-viewing system interoperable across multiple commercial EHRs within NYP. 12,000 physicians use the iNYP daily to view 4.5 million patient records. iNYP uses "clinical dashboards" to summarize and synthesize patient information for the provider. Our department oversees the development and maintenance of the dashboards. To prepopulate patient demographics in the ProVis interface, a back-end programmer integrated it with iNYP using Microsoft's .Net. To integrate patient-reported data into the dashboard, iNYP will query the ProVis database server (see Hardware section) for the patient's MRN.

Clinician-Facing Design Process: Clinicians view PROs within discrete packets of information, termed "cards," within the iNYP clinical dashboard. The cards separate data into portions and support one particular clinical action each, to prevent information overload. One card summarizes recommended clinical actions for the provider based on the novelty and severity of each symptom. Because heart failure symptoms differ dramatically between individuals, and clinicians may wish to track different symptoms in different patients, we offer clinicians the ability to customize cards to individual patients. We anticipate that customization will increase utilization of PROs in the EHR. We used visualization techniques for time-oriented data to visualize PROs longitudinally in cards. For example, we visualize how patients' symptoms scores vary according to medication dose and type (see Supplementary Materials).

Training: The Patient-Centered Outcomes Research Institute (PCORI) user's guide on integrating PROs in EHRs recommends training administrators and clinicians prior to implementing PROs to prevent non-use. To help clinicians understand how to manipulate provider-facing visualizations and discuss them with patients, we will offer training at monthly cardiology team meetings. Additionally, we will train the clinic receptionists to engage patients in ProVis and provider user support. We will provide a phone number to report hardware problems.

Hardware: ProVis will be accessed using iPads located in cardiac clinic waiting rooms. The iPads access secure WiFi through an existing hospital network. To protect the iPads, the hospital uses rubberized cases and installs theft-deterrent software. The information systems, including the ProVis application itself and the databases housing patient-reported data from ProVis, use secure, encrypted, HIPAA-compliant servers. Physical access to the servers requires keycard authentication, and we secure electronic access via the Secure Shell (SSH) protocol.

Alternative Solutions, Strengths, and Weaknesses

We choose to implement ProVis in the waiting room for five key reasons. First, although our institution offers an online patient portal, monthly use for patients with access lies below 2%, consistent with the low adoption of portals across the US. Second, physicians can immediately act on information collected in the waiting room, decreasing legal liability. For example, if the patient screens positive for depression in the waiting room, the physician can immediately place them on an antidepressant trial. Third, because every patient the provider sees reports PROs from the waiting room, the provider can integrate PROs into their standard clinic workflow, which may increase PRO utilization. As such, we do not need to prompt providers with notifications to view PROs, and this may prevent alert fatigue. Fourth, the clinic receptionist can assist patients with the application in the waiting room, potentially increasing reliability and therefore clinician's trust in the patient-reported information. Fifth, restricting the frequency of PRO collection to clinic visits reduces the burden of reporting on patients. Although collecting PROs in the waiting room demonstrates many strengths, one weakness is the inability to monitor symptoms between visits.

We chose to use an external, or "wraparound," system for PRO collection, rather than collecting PROs using an application within the EHR system itself. External systems offer several strengths over internal systems, including the easy customization necessary to create our visualizations. However, external systems also demonstrate several weaknesses. First, although our system does not require any username or password, the clinical receptionist must links the external system directly to the EHR using the patient's MRN. Second, our external system uses a unique protocol to communicate with iNYP, which is not easily generalizable to other EHR systems.

We chose to normalize patient-reported symptoms relative to the population distribution in only one clinician-facing visualization. In participatory design sessions, our clinician participants described population-level data as irrelevant to the heart failure clinic, which sees primarily vulnerable populations. Clinician participants placed more emphasis on patient perceptions of severity and changes to severity over time, so we emphasized that instead. Furthermore, the current version of ProVis does not normalize to the population distribution in patient-facing visualizations. This
is a weakness, since seeing their information relative to the population may motivate behavior change in some patients. However, it is also a strength, since participants in usability testing worried that population information may prompt complacency ("I'm where everyone else is") or despair ("I'll never be able to fix this").

Implementation and Dissemination Plan

Prior to implementation and dissemination, we will demonstrate the usability and usefulness of our visualizations to patients and clinicians, as well as their impact on PROMIS's precision and efficiency (see Evaluation Plan section). As discussed above, we will work with the Center for Advanced Cardiac Care to integrate ProVis into electronic intake systems in our hospital's heart failure clinics. Going forward, the iNYP clinical dashboard team at Columbia University Department of Biomedical Informatics will govern the cardiology PRO system in collaboration with the Center for Advanced Cardiac Care. Independent of implementation of the ProVis application, the visualizations hold potential for adaptation and use in any system incorporating PROMIS measures. Ideally, we would translate our visualizations to other systems through close collaboration with institutions participating in the NIH EASI-PRO initiative. We will share all of the tested visualizations and interfaces on GitHub.

Dissemination beyond the clinic to hospital administrators and policy makers is an important consideration. For example, Franklin and colleagues recently demonstrated that incorporating PRO data dramatically improved their readmission risk prediction model (c-statistic of 0.86 vs 0.65). Hospitals use readmission risk prediction models to distribute resources, and the Centers for Medicare and Medicaid Services uses such models to standardize readmission rates between hospitals. Making PRO data available to administrators and policy makers may improve readmission risk prediction, among other pursuits.

Evaluation Plan

In addition to the evaluations our team already conducted (see first paragraph of the Proposed Solution section), we plan to complete the evaluations described below.

Further Usability Testing with Patients: We will recruit 20 patient users from the heart failure clinic for laboratory-based usability testing. The research coordinator will guide the end-users through an IRB-approved usability assessment protocol. The users will complete goal-oriented tasks such as reporting specific symptoms. The coordinator will video-record the screen during each interview using QuickTime Player and ask predetermined questions to encourage concurrent think-aloud. After each task, the coordinator will assess satisfaction, ease-of-use, and confidence with retrospective probing techniques. Quantitative measurements will include time to task completion (approximating efficiency) and the task error rate (approximating precision).

According to the ISO / IEC 9126-9 quality model, usability involves 4 dimensions: (1) understandability, (2) operability, (3) learnability, and (4) attractiveness. Only 32% of usability-related publications focused on mobile health applications assessed all 4 dimensions. However, assessing all dimensions is critical. For example, attractiveness, the least assessed dimension, plays the greatest role in a user's decision to being using an application. To ensure we assess all dimensions, we will administer a field-based usefulness, ease-of-use, and ease-of-learning questionnaire to 20 randomly-selected patients after their clinic visit.

Validation of PROMIS Visualizations: We will conduct a two-arm randomized trial to evaluate the impact of visualizing answer choices on PROMIS's precision and efficiency. A 20 person pilot study prior to the trial will generate parameters to estimate effect sizes in an a priori power analysis. General medicine patients will be block randomized to complete PROMIS measures with visualizations or without visualizations. For the gold standard, all patients will complete "legacy" questionnaires (well-validated and widely-used measures of the same concepts). At baseline, both arms will complete a survey to assess demographics, health literacy, and prior technology use. The primary outcome will be efficiency (time to completion) and precision in comparison with the gold standard.

Semi-Structured Interviews with Clinicians: We will conduct semi-structured interviews with clinicians to assess the predisposing factors (attitudes towards use), enabling factors (skills and support required for use), and reinforcing factors (perception that use makes a difference) for use of clinician-facing PRO visualizations in the EHR. We will analyze the interviews in multiple steps using a qualitative descriptive approach to uncover common themes. First, 2 researchers with training in qualitative methods will independently read each transcript, and define codes in a
We will conduct 1 round of inter-coder comparison queries. The coders will meet to review, discuss, and arrive at consensus for the content coding. To enhance rigor, a third researcher with extensive experience in qualitative research will conduct an audit trail of the coding processes. To enhance confirmability, we will use “member checks” and peer debriefing among the data collection team. We will share summaries of the coded data with 3 clinicians and ask for their confirmation or revisions to interpretation.

**Conclusion**

Our proposed solution, ProVis, uses visualizations to display PROMIS measures and outcomes to heart failure patients in the waiting room, and display PROMIS outcomes to clinicians in the EHR. We believe ProVis will impact heart failure symptom management and prevent readmissions, as well as increase PRO utilization.

**Acknowledgements**

This work is supported by the National Institute of Nursing Research (K99NR016275, PI: Masterson Creber). Thank you to Ruth Masterson Creber, Drashko Nakikj, Meghan Reading, and Maria Hwang for their valuable comments.

**References**

Sensi-steps: Using Patient-Generated Data to Prevent Post-stroke Falls

Angela Smith, MS, Ada Ng, Eleanor R. Burgess, MSc,
Noah Weingarten, MA, Jennifer A. Pacheco
Northwestern University, Chicago, IL, USA

Abstract

We present Sensi-steps, an application using patient-generated data (PGD) to prevent falls for geriatric and especially post-stroke patients. The Sensi-steps tool incorporates a wearable wrist device, pedometer, pressure and proximity sensors, and tablet. PGD collection occurs through Timed Up and Go (TUG) tests and collection of physiological data, which is integrated into the EHR. Fall risk factor active tracking encourages new ways of shared decision-making between patients, caregivers, and practitioners. PGD will be managed at the primary care nurse or Care Manager level (see 3-tier PGD service proposal), presenting a novel way to incorporate PGD into clinical decision-support systems. We expect our solution to be easier to use routinely by the patient at home than other fall risk tracking solutions. Sensi-steps has the potential to improve patient care, help patients make informed decisions, and help clinicians understand patient-generated, environmental, and lifestyle information to deliver personalized, preventative healthcare.

Background

Problem Definition Patient falls are a major health concern in geriatric care and rehabilitation. In several geriatric populations, approximately 80% to 90% of patient injuries are caused by falls. The risk of falls is abnormally high among stroke patients, and falling is one of the most common recurrent complications among stroke patients in rehabilitation. Stroke patients are susceptible to falls due to several factors: reduced efficacy of postural responses, diminished sensory acuity, impaired musculoskeletal, neuromuscular, and/or cardiopulmonary systems, deconditioning associated with inactivity, depression and low balance self-efficacy, polypharmacy and a host of environmental factors. Clinicians assess for the presence of these factors by: asking questions about medical history (e.g., do you remember falling in the last 6 months?), obtaining self-report measures (e.g., fear of falling scales), and administering performance based tests (e.g., Timed Up and Go test and ambulation speed). Clinicians synthesize this information to identify specific causes of fall risk that can be addressed.

The “Timed Up and Go” (TUG) test provides a reliable and valid assessment of functional mobility in the population at large. It has been specifically validated as a predictor of fall risk in community-dwelling older adults and patients with stroke. A fall is an event in which a patient unintentionally comes to rest on the ground or other lower supporting surface, unrelated to a medical incident or to an overwhelming external physical force. Since falls are common in stroke rehabilitation, a major challenge is to reduce the number of falls and injuries without limiting patients’ physical activity. Many falls are caused by intrinsic mechanisms and occur during transfers and changes in position -- and in most cases by the patient acting against given instructions. Consequently, a major goal of stroke rehabilitation is to improve postural stability, motor performance, and the patient’s self-perception and awareness of the risk of falls. Achieving these goals enables patients to ambulate more safely. This project seeks to collect ongoing patient-generated data from stroke patients to identify those at risk to implement appropriate preventative measures and therapies in place and decrease complications.

The AMIA Challenge Our team presents the Sensi-steps application designed to decrease fall risk in patients recovering from strokes by collecting patient-generated data (PGD) predictive of fall risk. PGD reports are shared with patients and clinicians in a user-friendly format. The app displays data in both the EHR and the app itself in a way that establishes common ground between patients and clinicians, supporting shared decision making. Our application satisfies clinicians and patients alike by collecting and displaying the following PGD: patient TUG score, ambulation speed, and basic physiological data relevant to stroke rehabilitation. This data can be collected automatically or manually through a series of tools: pressure/load and proximity sensors, Alvita Wireless Activity Tracker (pedometer), and the Omron Heartvue wearable wrist device. Sensi-steps rewards patient’s test compliance with a virtual garden, a type of performance feedback that provides a motivating context for therapy. The Sensi-steps physiological data displays in corresponding fields in the electronic health record (EHR) and high-risk TUG scores are sent as an EHR note to primary care nurses and Care Managers. Patients’ information is logged in the Sensi-steps application for trend and progress overview. Sensi-steps integrates into the patient’s EHR where clinicians can see clinical, evidence-based information related to predicted fall analysis and imbalance, with the aim of improving clinical interpretability.
Solution Design and Development

Design Process Overview: Our user-centered design process focused on understanding and building PGD solutions for the needs of stroke patients and their team of providers. Our initial point of contact was Dr. Chen Lin, a practicing neurologist and Stroke Research Fellow at Northwestern University who researches technological initiatives for improving stroke recovery. He detailed what he saw as the major needs and challenges for patients in stroke recovery. This foundational knowledge allowed our team to plan our design process in three phases: (I) needs analysis, (II) design, and (III) iterative human-centered evaluation.

(I) Needs Analysis: First we carried out a literature and current technology solutions review, and then we conducted interviews with multiple relevant stakeholders. Our initial literature and technology research delved into existing tools that supported common stroke recovery symptoms such as aphasia, cognitive and motor functionality decrease, mood and anxiety disorders, as well as overall decreased quality of life. Information gleaned from our literature review guided our semi-structured interview protocol with clinicians and patients. For a list of questions and general themes asked, please refer to the Interview Protocol in Appendix A. In total, we interviewed eight practitioners, including two primary care physicians (PCP), one physical therapist (PT), one speech-language pathologist (SLP), one occupational health physician, and three neurology fellows at a teaching hospital in a large Midwestern city. We, additionally, interviewed with a number of patients with and without their caregivers to aid in this design process.

The patients’ needs analysis findings identified the following concerns and suggestions: 1) discerning what PGD and patient concerns warrant seeking immediate/emergency medical care; 2) simplifying technology usage; 3) general reminders to slow down. Similarly, our clinicians noted the following issues during our phase 1 assessment: 1) biological trends in relation to prescription regimens; 2) various risk management factors (glucose, blood pressure, heart rate, cognition, exercise, etc.); 3) data that provides an actionable outcome (and that the alerts themselves be actionable with pre-formatted buttons enabling quick decision-to-action pathways); 4) desire to have primary care nurses or Care Managers manage PGD reports; 5) the legal reality that ongoing PGD must be managed at the primary care level.

(II) Design: Our tool was developed to fulfill the identified needs of our participants and reflects overall needs assessment insights. Our interviews indicated a lack of general knowledge about stroke recovery rehabilitation among specialists and PCPs. Day-to-day stroke rehabilitation management is left in the purview of primary care physicians, yet those to whom we spoke explain that they are largely unfamiliar with how to meet the specific needs of stroke patients. Here, we present our design of the Sensi-steps a) novel technology and b) novel tiered PGD and EHR management process for primary care support staff.

a) Novel Technology: Our proposed solution, Sensi-steps, is a tablet-based application bundle designed to decrease fall risk in patients recovering from stroke. The tool has four main technological components comprised of currently available market technologies: (1) an ambulation speed calculator, (2) a TUG performance tool, (3) a patient portal add-in for medical adherence and notes to the clinician, and (4) practitioner-facing PGD EHR data visualizations. We focus on ambulation speed and the common clinical Timed Up and Go test because the time taken to complete this test is strongly correlated to the patient’s level of functional mobility, (i.e. the more time taken, the more the patient is likely to be dependent for daily living activities). Older adults who take longer than 12 seconds to complete the TUG have a high risk for falls. 

(1) The Sensi-steps ambulation speed calculator utilizes a combination of either a) the Omron Heartvue, a wearable wrist device that accurately captures blood pressure and tracks other heart health data, such as sleep patterns, together with the Alvita Wireless Activity Tracker (pedometer) or b) three household sensors placed within the patient’s home. One is a pressure sensor located in the patient’s seat and the other two are proximity sensors mounted on the wall in corridor (see Appendix E for household sensor setup). (2) The Timed-up-and-go (TUG) performance tool responds to the Omron Heartvue or the proximity and load sensors via Bluetooth connectivity. The Omron Heartvue TUG, hereby known as the “manual TUG”, begins when the seated user taps the ‘Start’ button on the watch application and ends when the user has risen from his chair, walked 10 feet, returned to his seat, and tapped the application’s ‘End’ button. Upon completion, the user is presented with a new seedling, which is planted in the virtual garden, the application’s background. The user has choice to use the pedometer as well, which will provide more accuracy and allow capture and calculation of the ambulation speed. Upon continuous usage, the seedlings continue to grow and the user’s garden expands. The sensor option, hereby referred to as “automated TUG”, tracks every time the patient rises from the chair configured with a load sensor (Sensor A). As the patient begins to rise from the chair, Sensor A detects a negative change in pressure and starts the TUG timer. The test is
logged in the Sensi-steps application when the Omron Heartvue connects via Bluetooth to the wall-mounted sensor (Sensor B) and the chair’s pressure is returned to zero, indicating that the patient has sat down again. The test will time out at 30 seconds. If the test times out before final pressure is returned to the chair, the tool will prompt the user with a dialogue window providing the option to call for support if the user has fallen. Upon clicking the app, the user is prompted with a disclaimer: “For your safety, caregiver supervision is highly recommended while you take this assessment.” If a patient opts out of sensor integration in their home, ambulation speed can only be calculated using the manual TUG test. Sensi-steps also incorporates (3) medical prescription compliance and messages to the clinician through the patient portal. Our interviews with clinicians identified prescription compliance as a point of action, and we incorporated this into automated data. The use of algorithms will populate automated trend information below the user’s graphic display, e.g. improved performance with continued use of Drug A or measurements at certain times of day (See Appendix D, p. 10 and Appendix D-3, p. 16). (4) Lastly, information visualizations for patients and for clinicians showing physiological data, ambulation speed, and TUG trends. This directly integrates actionable information for clinicians. The main Sensi-steps dashboard provides an array of this information (see Appendix D, p. 9). The Fall Risk Score is visualized on a spectrum from 0 to 6 and is calculated with an algorithm that incorporates TUG time and ambulation speed. Left of the score display are five metrics, all of which are synced passively from the Omron Heartvue and sensory data, blood pressure, heart rate, step count, distance travelled, and last TUG. Clicking an individual metric provides a graphical screen illustrating the specific metrics trend over time. The banner across the top of the screen allows the patient to view and share information from the tethered personal health record (PHR) or patient portal view and write specific notes that will be linked to relevant EHR information.

Because approximately 70% of the one million stroke patients discharged annually from United States hospitals continue living in their homes15, we envision our solution being used in-home as: a routine part of physical therapy, a shared information delivery system, and a tool for primary prevention of falls, for patients aged 60 to 101 years. A longitudinal study was performed using the TUG tool at a patient’s 1st week to 3 months, 3 to 6 months, and 6 to 12 months after discharge. Research shows that the TUG is a responsive test for capturing improvements in mobility during the first three months after stroke16. Responsiveness was established in relation to statistically significant changes in mobility.

b) Novel Service Provision: Key to ongoing management of PGD is to identify which members of the healthcare system are most able to closely track ongoing PGD “red flags.” We present a three-tiered system for PGD management to ensure long-term sustainability of not just Sensi-steps, but also the many other future PGD applications. Tier 1 is for information or actions that a Medical Administrative Assistant could accomplish, such as rescheduling a patient meeting. Tier 2, the main PGD ongoing management tier, is for information or actions that could be under the purview of a primary care nurse or Care Manager. In the attached appendix, the Sensi-steps PGD notification would be sent to the primary care nurse or Care Manager. Tier 3 is for high-priority notifications and sign-offs at the level of the primary care physician.

(III) Iterative Evaluation Once an initial wireframe was developed, we conducted a more formal needs assessment (phase 2) with additional provider types and researchers at multiple institutions and with potential patients. We asked for feedback on our initial wireframes (see Appendix B and Appendix C for the list of questions asked as well as the initial wireframes). For example, for the wireframe provider-side EHR message, we conducted three iterations based on clinician feedback. This process showcased the importance of including the PGD data graph together with decision pathway buttons. We also interviewed patients prior to our initial design, which guided our development from the initial wires to the prototype design. We presented Sensi-steps, our newly developed prototype, to aid fall prevention in stroke recovery patients to PCPs and various patients to gauge their conceptual understanding, usability, and sustained usage.

Alternative Solutions and Solutions Comparison

During outpatient sessions, physical therapists provide stabilization exercises and monitor patient progress utilizing validated indexes and tools, such as Berg’s balance scale, the Barthel Index, and the SitBAT assessment tool. SitBAT is not considered a clinical tool, so we have not included it within the strengths and weaknesses assessment. Sensi-steps and its use of the timed up and go assessment succeeds because the previous tools lack global interpretability, acceptability, and feasibility. There are no common standards for the interpretation of these scores despite the establishment of a minimum score for balance impairment and a lack of standardization across individuals. The rating scales associated with each item, while numerically identical, have different operational definitions for each number or score. For example, a score of 2 in one item is defined differently and has a different
associated level of difficulty when compared to a score of 2 in another item. Use of an overall score that adds ratings with different meanings having no common reference point may not be appropriate as interpretation is difficult and very little functional information is provided about the individual patient. Table 1 summarizes the alternative solutions as well as their strengths and weaknesses.

<table>
<thead>
<tr>
<th>Solution</th>
<th>Strength</th>
<th>Weakness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensi-steps</td>
<td>● Integration with EHR&lt;br&gt;● Can be completed outside of the clinical context&lt;br&gt;● Requires little to no training to use and administer assessment&lt;br&gt;● High inter-rater reliability&lt;br&gt;● Simple and quick assessment</td>
<td>● May be too technologically complex for certain members of our target population&lt;br&gt;● Wearable may cause discomfort for some wearers&lt;br&gt;● Cost of sensors</td>
</tr>
<tr>
<td>Berg Balance Scale</td>
<td>● Quantitative assessment of balance&lt;br&gt;● Measures static and dynamic aspects of balance</td>
<td>● Requires direct clinician observation&lt;br&gt;● No common interpretation of scores&lt;br&gt;● Some item ratings are not used at all or are underutilized, collapsing the rating scale.</td>
</tr>
<tr>
<td>Daily Gait Index</td>
<td>● High inter-rater reliability&lt;br&gt;● Focuses on changes in balance and ability to respond to tasks while walking</td>
<td>● Ceiling effect noted at 60 years of age[^9]&lt;br&gt;● Several unknown difficulties in attempts to normalize data[^9]</td>
</tr>
</tbody>
</table>

Table 1. Comparison of risk tools for fall prediction and imbalance.

**Implementation and Dissemination**

We have spent over three months iteratively developing our current prototype with input and feedback from provider and patient stakeholders. We will continue to iteratively improve our tool based on further needs assessments and feedback and add additional features, such as educational resources and community support connections. Our next version will be ready by November 6, 2017, and will include changes and improvements resulting from usability testing, including visualization of risks for providers/patients/caregivers (e.g. infographics, dashboards), and context-sensitive links or InfoButtons directing patients to clinical guidelines and knowledge resources (e.g. American Stroke Association resources). Future versions of Sensi-steps will integrate: users’ assistive devices into their Fall Risk score, TUG test extensions to incorporate cognitive and manual tasks such as counting backwards or pouring a glass of water, and assessments reported by patients and caregivers. These iterations will help demonstrate that Sensi-steps can be generalizable to other measures of function, health risk, and quality of life and applicable to other conditions, which increases the impact of our solution.

To disseminate our application to others, we will collaborate with stakeholder groups such as Dr. Chen Lin and his clinic, and the associated clinical and research community. We strive to solve a problem that not only satisfies the goals of the AMIA Design Challenge, but also aligns with Dr. Lin’s focus to develop app-based recovery tools to improve outcomes in post-stroke patients. After pilot trials at Northwestern University (NU) we will capitalize on NU’s partnerships with other hospitals and rehabilitation clinics. We may also partner with the Omron company to integrate our successfully trialed Sensi-steps application within all Heartvue devices. Additionally, we will leverage organizations such as the American Stroke Association, the National Stroke Association, and the American Stroke Foundation to promote and gather feedback on our tool. The shared mission of these organizations is to empower survivors and their families to overcome the challenges of life after stroke. Our tool promotes this mission by actively engaging stroke patients and their caregivers in the complex decision making and self-management required during the stroke rehabilitation process.

**Evaluation**

The proposed Sensi-steps tool will be evaluated through usability testing, and comparative effectiveness evaluation of similar balance and fall risk tools. Usability testing has been recommended as an important part of the evaluation of health risk communications, and is a recognized method to achieve user-centered improvements in health information, and research-based improvement in web information resources[^8]. Usability testing will follow recognized methodologies, and allow us to observe actual users and evaluate if Sensi-steps meets user needs. Results will enable identification of potential usability problems, and inform redesign and / or improvement of the fall prevention tool.

**Conclusion**

Falls are common and potentially debilitating for patients who have had strokes and for the geriatric population, at large. There is growing demand for patient-focused preventative efforts. There is also a great need for tools that help clinicians parse through and handle the complexity of clinical and biomedical “Big Data”. Sensi-steps has the
potential to improve patient care, help patients make informed decisions, and help clinicians understand patient-generated, environmental, and lifestyle information to deliver personalized, preventative healthcare to improve patient and population health.

References


2281
Quantifying and Visualizing Medication Adherence in Patients Following Acute Myocardial Infarction

Jane Wang, B.S.1, Shannon Wongvibulsin, B.S.1, Katharine Henry, B.S.2, Saki Fujita2
1 Johns Hopkins University School of Medicine, Baltimore, MD, 2 Johns Hopkins University, Baltimore, MD

Abstract
Medication adherence is a critical component for recovery following acute myocardial infarction (AMI). Currently, numerous smartphone applications are capable of tracking medication adherence through patient-generated data (PGD), but few are integrated with the electronic health record (EHR). Integration of medication adherence PGD into the EHR can give both healthcare providers and patients increased insight into patterns of missed doses, effects on vital signs, and correlation with side effect symptomology to inform healthcare decisions. We propose the generation of a medication adherence “vital sign”, a score calculated based upon the patient’s reported doses taken collected through a smartphone application and streamed into the EHR on a daily basis. We also propose the creation of Patient Health Reports that incorporate relevant patient history, information from previous visits, and the medication adherence scores to give providers a comprehensive view of patients’ health prior to clinic visits. These features are intended to incorporate PGD into clinical care to inform decision making in ways that streamlines patient visits, reduces healthcare costs, and improves health outcomes.

Introduction
Medication non-adherence has been reported in over 60% of patients with cardiovascular disease (CVD), the leading cause of morbidity and mortality worldwide,1, 2 and in approximately 50% of patients with chronic diseases.3 Typical medication regimens for CVD are complex and involve drugs from multiple classes that may require intake at numerous different time points4. Medication non-adherence contributes to poor achievement of treatment goals, increased patient morbidity and mortality, and an estimated $300 billion annually in U.S. healthcare expenditures.5 Verification of proper patient usage of prescribed medications is important for assessing treatment efficacy. In most outpatient settings, clinicians rely upon the patient’s self-report through recall.6 However, studies have shown that the patient’s recall for medical information may be poor, and patients generally over-report adherence.7 Suboptimal assessment of patient adherence leads to medication errors, such as over-prescription and dosing errors.

With the growth of health information technology, the Electronic Health Record (EHR) has become central to the practice of medicine. According to the Office of the National Coordinator for Health IT (ONC), approximately 74% of physicians have adopted certified EHR systems.8 However, approximately 51% of physicians are only using basic EHR functionalities, such as inputting patient demographic information.8 EHRs also introduce work-flow disruptions for clinical staff that have been shown to lead to productivity losses of up to 20% in internal medicine clinics within the first month of implementation.9,10 These studies illustrate that EHRs are not yet being used to their greatest potential.

Recent developments in smartphone applications and wearable technologies have led to the concept of patient-generated data (PGD), which are health metrics captured by the patient outside of the context of formal clinical care.11 The integration of PGD into the EHR has been suggested as a method of allowing providers to monitor patient health behaviors and key health metrics, enabling them to make more informed decisions on patient care.11 However, incorporating PGD into the EHR is hindered by numerous challenges including the following: the capacity to handle large volumes of data in a manner that neither overwhelms the EHR11,12 nor the providers,12,13 and the presentation of PGD in a way that is easily interpretable and relevant to clinical decision making.

Proposed Solution
A. Medication Adherence “Vital Sign”
A group of clinicians, engineers, and hospital administrators have developed the first Apple CareKit smartphone-based cardiology application, Corrie, to reduce hospital readmission rates following acute myocardial infarction (AMI).14 Corrie aggregates all components of follow-up care surrounding AMI, including medication trackers, vital
sign intake, care team contacts, and educational modules into a single mobile application. Corrie’s current platform enables patients to log their adherence to each prescription on the basis of dose and time.

We propose the creation of a medication adherence score and corresponding graph to be streamed into the EHR daily as a “vital sign”. The medication adherence score will be calculated for each medication as (#doses taken)/(#doses prescribed) *100. The visualization will be generated based upon the patient’s medication tracker that captures the patient’s adherence to each medication, temporal and prescription-specific patterns of adherence, and medication side effects (Figure 1). This feature will also have the capacity to generate visualizations that highlight time points throughout which the patient has not been adherent to specific medications (Supplemental Materials). Clinicians may hover over the plot in order to see precise data specific to time, day, and medication. The medication adherence “vital sign” will be incorporated into the EHR as a streaming data set that can be accessed for any day in the patient’s history, and as a cumulative report that summarizes the information into an adherence score for a specified window of time. Additionally, to evaluate whether the current medication regimen requires titration adjustments, adherence to medications targeted at outcomes such as blood pressure (BP) and heart rate (HR) can be viewed alongside the recordings of BP, HR, and other vital signs as well as reports of side effects or symptoms to determine trends between adherence and outcomes (Figure 2). Clinicians who are concerned about patient compliance to medications and the consequences on health outcomes may check on patient medication adherence and directly contact the patient through the mobile application’s secure platform. These visualizations will also be available to patients to increase patient engagement and empowerment, which has been shown to be linked to patient satisfaction and compliance, which in turn can lead to improved health outcomes.

Figure 1: Medication adherence plot emphasizing completed doses

Figure 2: Medication adherence plot (bottom panel) with corresponding daily changes in vital signs, mood, and steps (top panel)
B. Patient Health Reports

To streamline clinical workflow, we furthermore propose the creation of a Patient Health Report (Figure 3) that clinicians may generate and view before a clinical encounter. Clinicians may specify the timeframe of interest (e.g., since the last clinical encounter) for the Patient Health Report, which will include key components of every patient’s medical record such as past medical history, current medication regimen, concerns raised during the previous clinical encounter, a longitudinal graph of the patient’s medication adherence, and the overall medication adherence score for each medication. Since many clinicians prepare for patient visits by refreshing their knowledge of their patients’ medical history, needs, and active problems the night before or the morning of a clinical visit, we propose the health report to expedite pre-encounter preparation by providing clinicians with information that will inform the clinical encounter. In a process mapping study conducted by the Veteran’s Health Association, physicians agreed that reviewing recent patient information prior to a patient encounter helped to 1) build trust and rapport to establish a basis of familiarity and knowledgeability about the patient, 2) determine whether pre-visit tests or consultations are required, 3) remind clinicians about care plans from previous visits. By incorporating the Patient Health Report with the medication adherence information into the workflow that many clinicians already use, we hope to enhance the clinical workflow experience through the centralization of relevant information and the addition of the medication adherence vital sign. An example of the Patient Health Report can be accessed here: https://kathen.shinyapps.io/amia_visual3/. Further details may be found in the Supplemental Materials.

Figure 3: Sample Patient Health Report with pertinent patient demographics, history, and medication adherence
Alternative Solutions Considered

A. Data collection methods for medication adherence: As an alternative to using the medication logging functionality of the Corrie app, we considered using data collected from smart pill bottles17 or smart pills with digestible circuits18 that can automatically keep a record of the patient’s medication adherence.

B. Medication adherence visualization: As an alternative to the heat map of medication adherence, we considered presenting the visualizations as longitudinal bar graphs or line graphs to display the trajectory of the adherence score over time.

C. Medication adherence “vital sign” and comprehensive summary: As an alternative to the medication adherence score and Patient Health Report summary, we considered presenting the adherence information in a separate tab in the EHR where the clinicians can view the adherence history.

Strengths and Weakness

A. Medication adherence data collection through Corrie mobile health app: Corrie already incorporates a comprehensive recovery plan for AMI patients, so logging their medication intake through this app will facilitate ease of use and centralize the patients’ self-care tasks to a single app. Using a smart pill bottle and smart pills with digestible circuits would add another level of unnecessary complexity to the data collection process. Additionally, using smart pills could limit which medications we could track. Nevertheless, smart pill bottle and smart pills could help address the issue of the validity of PGD from self-reported loggings of medication adherence through the Corrie mobile app. Yet, if patients intend to falsify their adherence, they could also manipulate the records from smart pill bottles (by simply opening the pill bottle but not taking the medication) and smart pills (by dissolving the pill rather than taking the medication).

B. Medication adherence visualization: Presenting the adherence score as a heat map allows clinicians to easily visualize the general patterns of medication adherence. Visualizations as longitudinal bar graphs or line graphs would allow for exact values to be plotted and displayed over time. However, with patients commonly on dozens of medications, presenting the scores as bar or line graphs in different colors for different medications can become complex and challenging to interpret. In our proposed solution, clinicians may hover over each box in the heat map to see the actual value of the adherence score and how it was calculated.

C. Medication adherence “vital sign” and comprehensive summary: Incorporating the adherence score as a vital sign stresses that medication adherence is an essential component of medical care that must be considered frequently. One weakness of including the medication adherence as a vital sign is that this requires the adherence of all the medications to be summarized into a single value. To overcome this limitation, we have also included the options for clinicians to view the adherence score for each medication separately under the medications list section of the EHR. Our proposed solution would be more helpful clinically than presenting the adherence information in a separate tab in the EHR since our method integrates the medication adherence information into a clinical workflow that clinicians are already accustomed to.

Implementation and Dissemination Plan

Implementation: Our implementation plan includes feasibility and usability studies as the initial step. We propose to pilot our proposed solution in the cardiology clinic, specifically with Corrie patients and providers. The feasibility and usability studies will begin with clinician training of the new EHR features and patient education on medication adherence and the medication logging function on Corrie. Then, the patients and clinicians will be surveyed (additional details are provided in the Evaluation Plan section below).

Dissemination: To disseminate our proposed solution, we will develop open-source software that is modular, scalable, and easily adapted to a variety of EHR systems. Additionally, we propose to publish the results of our small-scale implementation and feasibility-usability studies to inform future large-scale implementations across various institutions and EHR systems. The Shiny app demonstration of the implementation as well as videos with step-by-step user instructions for patients and providers will help facilitate the dissemination and adoption of this proposed solution at a large scale. The code will also be published on GitHub, so individuals interested in adapting it to their systems can easily access and modify the code for implementation at their institutions.

Evaluation Plan

The efficacy of our proposed solution may be evaluated through the following methods:

1. User satisfaction and feasibility-usability surveys: These surveys will be incorporated into the Corrie application and administered at 1 month intervals to patients; providers will receive monthly surveys via email; they will evaluate the frequency of Corrie use, ease of use, medication compliance following Corrie use, and efficiency of use during clinical encounters.
2. Comparison of medication adherence rates of CVD drugs (e.g. aspirin, clopidogrel, simvastatin, etc.), hospital readmission rates, and blood pressure and heart rate control within a target range between Corrie users and the national average at the time points of 1 month, 3 months, and 6 months.

3. Comparison of medication adherence rates of CVD drugs, hospital readmission rates, and blood pressure and heart rate control within a target range between high-frequency Corrie users (log at least 5 medications/week) and low-frequency Corrie users (log 4 or fewer medications/week) at the time points of 1 month, 3 months, and 6 months.

Conclusion
Medication adherence is one of the most frequently cited causes of hospital readmissions following AMI. We propose streaming recorded medication compliance PGD from the Corrie application directly into the EHR that can be visualized and assessed in the same capacity as a vital sign. Providers and patients may monitor and conduct early modifications to the medication regimen. We furthermore propose the creation of a Patient Health Report that streamlines clinical encounters by incorporating all relevant information (including the medication adherence score) that providers may need to prepare for patient visits. The integration of medication adherence PGD has widespread implications in CVD management and can be easily translated to other disease entities. We hope that our proposed solution will improve medication adherence rates and ultimately, long-term patient health outcomes at reduced healthcare costs.

References
8. Monegain B. More than 80 percent of docs use EHRs [Internet]. Healthcare IT News. 2015 [cited 2017 June 30].
18. Enríquez J. FDA Clears First Ingestible Device for Medication Adherence [Internet]. Med Device Online. 2015 [cited 2017 June 30].
The BD2K Training Coordinating Center’s Data Science Education Platform

José Luis Ambite, Ph.D.¹, Lily Fierro, S.B.¹, Florian Geigl, Ph.D.¹, Jonathan Gordon, Ph.D.¹, Gully A.P.C. Burns, Ph.D.¹, Kristina Lerman, Ph.D.¹, Jeana Kamdar, MPH², Crystal Stewart, M.A.², Avnish Bhattrai, M.S., Xiaoyu Lei, M.S.², Sumiko Abe, Ph.D.², John Darrell Van Horn, M.Eng, Ph.D.²

¹Information Sciences Institute, Univ. of Southern California, Marina del Rey, CA, USA; ²USC Mark and Mary Stevens Neuroimaging and Informatics Inst., Los Angeles, CA, USA

The National Institutes of Health (NIH) has made an unprecedented investment in biomedical data science through its Big Data to Knowledge (BD2K) initiative (datascience.nih.gov). Through a series of K01 career development awards, T32/T15 institutional training awards, R25 data science training research awards, as well as the training components of fifteen U54 BD2K Centers of Excellence, the NIH has placed a premium on the development of a new generation of biomedical data science professionals¹. These efforts are systematically producing unique training materials which seek to introduce everything from the basics of data science to in-depth examinations of applied biomedical analytics in the investigation of health as well as disease. In addition to these BD2K training resources, a large number of high-quality data science educational resources are available online, such as massive open online courses and video lectures. Consequently, the BD2K Training Coordinating Center (TCC) is developing a data science educational resource discovery index (ERuDIte) and an online education platform with the capability to automatically discover, describe, organize, and personalize training materials for the varied learning objectives of its users.

In this system demonstration, we will first provide a tour of the core functionality of the BD2K TCC web portal (http://www.bigdatau.org), including the exploration, search, and recommendation of learning resources tailored to a user’s learning goal. Figures 1–3 show three key site features: the Concept Map, the Faceted Search Interface, and the Educational Plan Interface. The system is already available for beta testing. Second, we will describe how we use data science techniques to build the ERuDIte index that powers the web portal. This system demonstration aims to inform the audience on the educational value of http://www.bigdatau.org for the medical informatics community, while also showing the generality and extensibility of the education platform to other domains.

Figure 1: Concept Map

Figure 2: Faceted Search

Figure 3: Customizable Educational Plan to collect resources significant to individual learning goals

References


Systems Demonstration: NIH Common Data Elements Repository

Liz Amos, MLIS, Christophe Ludet, MS, Lisa Lang, MPP and Vojtech Huser, MD, PhD
U.S. National Library of Medicine, National Institutes of Health, Bethesda, MD

Description of purpose: Ability to share and re-analyze data from clinical trials is a well-recognized challenge in Clinical Research Informatics (CRI). Similarly, aggregation of corresponding data across several trials can facilitate larger meta-analyses that may lead to new discoveries. To facilitate data aggregation, research relevant Common Data Elements (CDEs) have been defined by number of initiatives1. The National Library of Medicine (NLM), with initial funding from an HHS Patient Centered Outcome Research Trust Fund, launched a beta version of the NIH Common Data Element Repository in January 2015 and has been involved in active pilot projects throughout 2015-2016.

The NIH Common Data Elements Repository has been designed to provide access to structured human and machine-readable definitions of data elements that have been recommended or required by NIH Institutes and Centers and other organizations for use in research and for other purposes. Each NIH Institute or Center (IC) classifies their CDEs in collections or sets belonging to specific clinical domains. The process of developing CDEs varies from initiative to initiative, but generally, a designated working group—or other set of subject matter experts—collectively decides the core variables or measures that should be recommended for researchers studying in that domain or disease area. The finished set is sent back to the IC and there is a period of public review and commentary. Once vetting is complete, the IC returns its data elements and classifies per the IC’s internal policies and formatting. The most established of these initiatives developed their own access portals and formats. While large repositories exist (most notable, NCI), the ability to search across all NIH CDEs, and to link clinical research to clinical data standards in active use in EHRs, seemed critical. For examples of current collections of CDEs from NIH ICs, see the NIH CDE Portal2.

CDEs have many benefits for research: facilitating cross-study comparisons of results, aggregation of data, speeding study start time by selecting existing measures, improving data quality via use of validated measures, and improving replication and reproducibility3. Perhaps most pressing, though, is the increase in language from NIH funding announcements which encourages or recommends the use of CDEs: “NIH encourages the use of common data elements in basic, clinical, and applied research, patient registries and other human subject research to facilitate broader and more effective use of data and advance research across studies...[and] investigators are encouraged to consult the Portal and describe in their applications any use they will make of NIH-supported CDEs in their projects”3.

A demonstration of the tooling and functionality will highlight the ability to search across all collections from initiatives in a centralized place, annotate with clinical terminology(ies), and variety of form creation and export capabilities. A brief discussion of the pilot projects will incorporate feedback from working groups and feature improvements based usage. The demonstration will also describe challenges and future steps for development and implementation.

Degree of deployment: The repository has been deployed and available since 2015 with new web-based functionality added regularly. It contains a total of 42,123 data elements and 1724 forms from 12 different initiatives (e.g., PhenX, PROMIS, National Eye Institute or AHRQ). For full export functionality (JSON, ODM/XML, Structured Data Capture/XML, REDCap), users must use/obtain a free Unified Medical Language System (UMLS) license.

Acknowledgements: This work is supported by the Intramural Research program of the U.S. National Library of Medicine.

References

CKMS – An Integrated System for Managing Institutional Clinical Knowledge Assets

Dominik Aronsky¹, MD, PhD, Dirk Wenke¹, MS, Priyaranjan Tokachichu², MD, MS, Christopher J. Vitale², PharmD, Margarita Sordo², MSc, PhD, Saverio M. Maviglia², MD, MSc, Roberto A. Rocha², MD, PhD
¹Semedy, Zug, Switzerland; ²Partners Healthcare System, Boston, MA

Abstract

Many healthcare institutions maintain a large quantity of clinical knowledge assets, such as rules, pathways, order sets, templates, and dictionaries, frequently within multiple proprietary clinical applications, resulting in “knowledge silos.” A central, dedicated and independent repository of knowledge assets with systematic lifecycle processes for managing institutional knowledge often does not exist. Knowledge silos result in inconsistent and error-prone workflows, requiring the wasteful application of specialized resources to update and maintain redundant knowledge assets. This demonstration showcases a novel clinical knowledge management system that enables a structured and systematic knowledge curation process, based on a strict semantic validation framework that ensures the integrity of assets at different lifecycle stages.

System Description and Functionality

The clinical knowledge management system (CKMS) provides a flexible and scalable framework for creating various types of knowledge assets, including detailed semantic linkages, while also guaranteeing the referential integrity of each asset through validation processes with a rule-based semantic reasoning engine. The external knowledge assets can be imported, extended, and linked to other assets within and across different namespaces. An authoring workbench supports a strict curation lifecycle and enables knowledge and terminology engineers to create, revise, clone, extend, validate, publish, audit, or retire knowledge assets. During the authoring and maintenance process, the participation of subject matter experts is supported by a collaboration workbench, which facilitates the communication and documentation required for the definition and maintenance of knowledge assets. An end user, such as a clinician, can search and view assets using faceted queries available via a read-only portal. Knowledge assets can also be displayed in a customizable, asset type specific view. Multiple import and export options support the need to embed and connect to clinical systems. These integration options allow standard terminologies, content from knowledge providers, and locally maintained assets to be consistently curated, including the effective management of dependencies across successive revisions (Figure). Knowledge-driven applications, such as an Infobutton Manager, can utilize CKMS as a backend system.

CKMS is the result of a multiyear collaboration between Partners and Semedy. The system is a web-based application using standard technologies (Java, JDBC, Apache Wicket, HTML, JavaScript, JQuery, jQueryUI, XML) and stores data in a MS SQL server database. IRIS™ serves as the reasoning engine. User authorization occurs via LDAP and a fine-grained role permission system governs asset control. REST-like web services provide access for external applications, enabling seamless integration with other systems. The presentation will provide in-depth information about the system’s architecture, structure, knowledge modeling approaches, etc. CKMS has been deployed at Partners for more than 24 months, and more recently at another large academic medical center. The system is currently being used to catalog and preserve legacy knowledge assets, as well as to catalog and manage assets used by a commercially available EHR. The system also includes curated versions of standard terminologies, such as SNOMED CT, ICD-10-CM, LOINC. The CKMS repository at Partners includes over 11 million discrete knowledge assets, corresponding to 138 distinct asset types.
Vicinity Exploration: Enabling User-Driven Visual Search of Multiple Machine Learning Models for Precision Medicine

Suresh K. Bhavnani PhD1, Archana Ayyaswamy MS1, Tianlong Chen BS2, Shyam Visweswaran MD PhD3, Gowtham Bellala PhD4, Kevin E. Bassler PhD2

1Institute for Translational Sciences, University of Texas Medical Branch, Galveston, TX; 2Department of Physics, University of Houston, Houston, TX; 3Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA; 4C3 IoT, Redwood City, CA

Abstract

Although machine learning methods (e.g., cluster analysis) are increasingly being integrated into visual analytical applications for identifying complex patterns (e.g., patient subgroups) in large data sets, such approaches typically generate a single best model based on optimizing an objective function. However, comparison of models in the vicinity of the optimal model can enable analysts to explore tradeoffs among key model parameters, important when exploring large datasets. Here we describe the user interface features critical for vicinity exploration, and demonstrate its efficacy in the use of a network layout algorithm to explore a large dataset related to precision medicine.

Introduction

Machine learning (ML) methods such as cluster analysis and feature selection are increasingly being integrated into visual analytical applications to enable the rapid identification and comprehension of patterns in large datasets. However, such methods typically output a single best model determined by optimizing an objective function. For example, the ExplodeLayout algorithm attempts to enhance the comprehension of networks that have significant but highly overlapped clusters (Fig. 1A). The algorithm takes as input a network layout (e.g., generated from a layout algorithm such as Fruchterman Reingold), and node clusters (e.g., generated by a modularity algorithm), and maps each cluster’s centroid onto equidistant points on an imaginary circle. The algorithm searches for a circle radius such that the ratio of the total non-overlapped area among the clusters, and the total space required for the entire network layout is maximized, resulting in an optimal cluster separation within a compact area (Fig. 1B). Here we describe the user interface (UI) design features and utility of vicinity exploration, an interaction approach that enables users to explore multiple ML models in the vicinity of the optimal model such as that generated by ExplodeLayout, to better comprehend complex patterns in large datasets.

Method and Results

As shown in Fig. 1, we designed five UI features for enabling vicinity exploration of an optimal model generated by ML algorithms such as ExplodeLayout: (i) vicinity graph implemented as a line graph of how the objective function changes with respect to the circle radius, (ii) optimal model indicator implemented as a dotted line on the graph to help recall the circle radius of the optimal model, (iii) single model viewer implemented as a scroll bar to select a circle radius and display a high resolution image of the corresponding exploded network, (iv) model significance implemented as progressive analysis and display of significance of vicinity models to address the high computation time in large datasets, and (v) multiple model viewer implemented as side-by-side views of user-selected models in the vicinity of the optimal model to reduce working memory loads during model comparisons. The above features were used by a geriatrician to comprehend patient subgroups within all 30-day readmitted hip fracture patients (n=6150), extracted from the 2010 Medicare database that had at least one of the 8 significant comorbidities shown. The optimal layout (Fig. 1B, radius=0.5) revealed 2 levels of heterogeneity in each of 7 patient subgroups: patients with one comorbidity on the outer side, and patients with more than one comorbidity on the inner side of each cluster. A higher separation of the clusters (Fig. 1C, radius=1), while not optimal based on the objective function and with lower resolution of the clusters, enabled the comprehension of inter-cluster edges showing how patients shared comorbidities across clusters. This led to hypotheses of how renal disease and diabetes interacted with other comorbidities precipitating hospital readmission in each patient subgroup. Future research will test the generality and usability of the above features in other ML-integrated visual analytical applications.

References

RxMix – Use of NLM drug APIs by non-programmers

Olivier Bodenreider, M.D., PhD, M.S., Lee Peters, M.S.
National Library of Medicine, National Institutes of Health, Bethesda, Maryland, USA

Contact information: RXNAVINFO@LIST.NIH.GOV

Motivation

Subject matter experts, including pharmacists, pharmacy benefit managers, health researchers and health data analytics specialists, are generally knowledgeable in drug resources, such as RxNorm, and have developed interesting use cases for it. However, processing datasets often requires programming skills, e.g., making calls to a service, such as the RxNorm Application Programming Interface (API). To enable users who do not have programming skills to leverage National Library of Medicine (NLM) drug APIs, we developed RxMix, a web application designed for non-programmers to create queries for complex use cases and execute batch queries against our APIs.

NLM drug Application Programming Interfaces (APIs) available through RxMix

Over the past decade, NLM has developed a variety of publicly available drug APIs to expose some of its resources through SOAP and RESTful web services for users to integrate into applications (https://rxnav.nlm.nih.gov/). API usage reached 1 billion queries in 2015. We also make these APIs available through RxMix.

- The RxNorm API enables users to access RxNorm drug names and codes. The most common use is for mapping drug identifiers (e.g., NDC codes) to RxNorm. It can also be used to map drug names from local formularies to RxNorm using normalization and approximate matching. The APIs is useful for traversing the rich graph of relations among drug entities provided by RxNorm. Recently developed functions help access historical drug codes for analytics purposes.
- The RxClass API enables users to access drug classes and drug members for a number of different drug class types, including ATC, DailyMed and NDF-RT. It can be used to find the classes for a particular RxNorm generic or brand name drug, or can list the ingredient members of a specified drug class.
- The Interaction API enables users to find drug-drug interaction (DDI) information for a specified drug or find the interactions from a list of drugs. Sources of DDI include DrugBank and the ONC high-priority list of DDIs.
- The RxTerms API provides access to the physician-friendly display names for clinical and branded drugs in e-prescribing applications.
- The NDF-RT API enables users to access clinically-oriented information associating drugs with their pharmacologic classes and other properties (e.g., mechanism of action and physiologic effect).
- The DailyMed API provides access to drug information from the FDA Structured Product Labels.
- The RxImageAccess API provides access to pill images as well as their related properties.

RxMix – Complex queries and batch processing

Using the RxMix graphical interface (https://mor.nlm.nih.gov/RxMix/), users can select an API function apply it in batch mode to a list of input values. For example, a list of identifiers from the National Drug Code found in a claims database can easily be mapped to RxNorm identifiers using the function RxNorm:findRxceByld. The batch query is executed on NLM servers and the results are returned to the users after completion.

Users can also create complex workflows (i.e., sequences of API functions) for specific use cases. RxMix “knows” which functions are interoperable and helps guide users in workflow development. Users can also select existing workflows from a library and save their own workflows for future use. Use cases suggested by users include mapping drug codes found in observational datasets to drug classes for analytics purposes and identifying non-opioid analgesics in a list of analgesics that includes multi-ingredient drug products.

All queries and workflows can be tested interactively in RxMix, which makes it a powerful educational resource. RxMix receives 150-200 batch queries and 1800 interactive queries each month on average.
NLM3D: an Open-Source Library of Medical-Imaging-derived 3D Polygonal Models for Use in Public Health, Medical and STEM Applications

Kristen Browne, MSc¹, Anne Altemus, MA²
¹,²Audiovisual Program Development Branch, Lister Hill National Center for Biomedical Communications, National Library of Medicine, Bethesda, MD, USA

System Description:

The NLM3D: Anatomy site has been developed to address multiple health and science issues. First, while a wealth of 3D data is being generated by imaging professionals in the medical and research communities, distribution is often limited to either 2D renderings of 3D objects or to a selection of images from large image stacks; this leaves a gap between the audience’s ability to interpret these images and the true meaning of the data. A library of optimized 3D scientific models derived from raw data can bridge these communication gaps, especially if they are delivered in conjunction with educational information explaining the imaging technology and how to use 3D models in practice. As a starting point, our library is focused on human anatomy derived from medical imaging data.

It is currently difficult to find accurate and high quality 3D scientific models. This can significantly delay the production of health and science communication resources. To address this issue, our base models are derived from segmentations of medical imaging data, then optimized to industry standards using professional digital sculpting and retopology software. To date, we have a nearly complete Visible Human Male, as well as a significant portion of the Visible Human Female. Future imaging sets could include the NCI Cancer Imaging Archive, the Visible Korean, and embryonic development series such as the Carnegie Embryology Collection ¹,²,³,⁴.

Models are uploaded to a MS SQL database that is managed with a Drupal content management system. A custom user interface requires inclusion of all model information (including structural modifications) as well as relations to custom anatomical vocabularies and imaging collection details. The UI also requires creation of ALT tags for all media including images and models to improve resource accessibility. All included metadata are autocomplete keyword searchable using an Apache Solr powered engine. Faceted and hierarchical searching are also available using the vocabularies and other metadata as filters.

NLM3D strives to provide the educational resources necessary for both communications professionals and the general public to utilize our assets. This includes relevant material describing the types of medical imaging used to create our models, how to process imaging data into 3D models, and how to use these assets in gaming, illustration, and animation environments. We also provide fully functional example projects, such as 3D printer-ready model sets, and interactive gaming environments.

NLM3D is a primary effort to create scientifically accurate and industry standard 3D models for public use, deployed through an engaging interface with interactive tools. This fills a different niche that other initiatives which are community-driven, where the quantity of models available can be high, but the quality control is more difficult to implement. All modeling and site development work has thus far been completed with a staff of one research fellow and one summer student per year. We therefore project that with the addition of 1-2 qualified medical illustrators we will continue to grow and maintain NLM3D with minimal resources. NLM3D is therefore well-poised to become a sustainable and trusted source of assets to populate novel health systems, treatment and training applications, and educational resources well into the future.

References:


Statement of Deployment:

As of June 15th, 2017 NLM3D anatomy exists as a functional prototype with most resources described above; educational content is still under development. We hope to have a complete prototype (with all NLM Visible Human Male and Female Models) in mid-2018.
Person-Centric vs. Problem-Centric: Data-driven Heuristics for using Machine Learning to Improve Health Outcomes

John B. Coles, Ph.D.¹, Brett F. Cropp¹, Jacob M. Ekstrum¹, Jacob T. Fijas¹, Michael D. Moskal¹ CUBRC Inc. Information Fusion Group, Buffalo, NY, USA

Abstract
In the world of big data, health outcome and personalized medicine research is in the process of shifting from a hypothesis, problem-centric approach to a data-driven, person-centric approach. In this talk we highlight how data-driven approaches to developing predictive models can open the door to improving health outcomes with cross-sector collaboration and data fusion, leveraging a diverse set of datasets and measures for model development and validation. Specifically, we will discuss how multi-variate heuristic filtering and semi-automated machine learning techniques can be used to select the right subset of variables and hyperparameters when developing predictive models. We will demonstrate the capability of the Heartwood Analytics™ Software, with a focus on the Machine Learning (ML) AutoPilot system, to rapidly develop new predictive models across a range of problems from open datasets, highlighting the breadth of data that can be rapidly assessed when developing new predictive models.

Introduction
This presentation will build on ongoing work addressing problems in outpatient treatment and case management, with the goal of providing guidance to stakeholders through prediction of intervention efficacy including short-term and long-term outcomes¹. Here we will build on promising results in clinical outcome prediction using Neural Nets, with advanced methods for temporal data². The Heartwood Analytics™ Software (Figure 1) was developed using open source software packages such as Apache Kafka, Storm, NiFi, and Fuseki RDF store. Heartwood’s semi-supervised machine learning component, ML AutoPilot, was built using Google’s TensorFlow Library to automatically identify driving variables based on heuristic filtering, combine a myriad of machine learning models, and automatically tune each model’s hyperparameters using Bayesian Optimization³. Finally, ML AutoPilot exports self-contained, problem-specific models that can be deployed as a web application via Flask.

Figure 1. Fusion of data with semantic enhancement and predictive modeling.

Current Deployments: The current system is in its Beta phase and is deployed on AWS and in internal cloud for secure customer data analytics. Additionally, it has been installed at multiple partner sites for evaluation and feedback for retrospective and prospective analysis to support clinicians and case managers in understanding and engaging challenging populations. We will highlight current work including the impact for customers while comparing and contrasting existing systems with Heartwood Analytics™ for capability, flexibility, and ease of integration.

References
Improving the Care of Women through Health Information Technology

Caitlin M. Cusack, MD MPH FACOG

1 Contract Physician Informaticist, Offices of Informatics & Information Governance and Women’s Health Services, Veteran’s Health Administration, Silver Spring MD

Abstract

The Veterans Health Administration has undertaken two projects to improve the documentation and tracking of care for Women Veterans: The Teratogenic Drugs project and the System for Mammography Results Tracking (SMART). These projects improve the ability to record and store discretely a woman’s ability to conceive, and pregnancy and lactation status. Refinement of order checks and alerts around known or potentially unsafe medication or imaging use in reproductive aged or breast feeding women increases the clinical relevancy and mitigates alert fatigue. Clinical reminders are used to increase conversations around family planning, preconception planning, and contraception use. SMART introduces the ability to track the management of a mammogram from the point of ordering, through additional testing and treatment, to a return to routine age-based mammogram screening. In combination, these projects will improve the efficiency in documenting care to women Veterans, as well as improve the quality and safety of that care. This system demonstration will highlight features of these projects.

Introduction

Many electronic health records have taken a gender-neutral approach to clinical decision support tools and documentation of care. An inability to capture reproductive data elements as structured data results in clinical reminders, order checks, and alerts that are indiscriminate and frequently clinically irrelevant, risking alert fatigue. One evaluation showed that drug-pregnancy alerts had the lowest acceptance rates of all drug alerts presented to clinicians. Across all genders, delayed follow up of abnormal results, immature tracking systems of preventive health care. Journal of the American Medical Informatics Association. 2006 Jan 1;13(1):5-11. Women’s Health Services, Veteran’s Health Administration, Silver Spring MD

Table 1. Overview of System Demonstration

<table>
<thead>
<tr>
<th>Teratogenic Drugs</th>
<th>SMART</th>
</tr>
</thead>
<tbody>
<tr>
<td>Easy capture of ability to conceive, pregnancy and lactation status as discrete data.</td>
<td>Linking of mammography results notification to BIRAD-specific documentation templates.</td>
</tr>
<tr>
<td>Clinical reminder dialogs capturing pregnancy intentions and method of pregnancy avoidance.</td>
<td>BIRAD-specific templates with typical next steps, linking to orders and consults.</td>
</tr>
<tr>
<td>Medication/image review when a status of ‘yes’ pregnant or ‘yes’ lactating is entered in the system.</td>
<td>Single view of all actions taken against an index mammogram, from the point of order through to a return to age-based screening.</td>
</tr>
<tr>
<td>Medication/imaging order checks based on a woman’s ability to conceive, use of contraception, and pregnancy and lactation status.</td>
<td>Floating windows ideal for easy viewing opened from within the templates containing additional information on any un-reviewed breast images, actions taken against an index abnormal mammogram, and last 3 breast images.</td>
</tr>
<tr>
<td>Generation of notifications to ordering providers regarding potentially unsafe prescribed medications and ordered images when pregnancy or lactation status changes to ‘yes’.</td>
<td>Next actions taken and method of patient notification captured as discrete data allowing for easy identification of those patients for whom an action hasn’t taken place or who haven’t been notified of results.</td>
</tr>
</tbody>
</table>

References

AMICUS: A Metasystem for Interoperation and Combination of UIMA Systems

Gregory P. Finley, PhD¹, Benjamin C. Knoll¹, Reed McEwan, MS, MSSE¹, Genevieve B. Melton, MD, PhD¹, Hongfang Liu, PhD², Hua Xu, PhD³, Serguei V.S. Pakhomov, PhD¹
¹Institute for Health Informatics, University of Minnesota, Minneapolis, MN; ²Mayo Clinic, Rochester, MN; ³University of Texas Health Science Center at Houston, TX

Introduction

There are many publicly available systems for natural language processing (NLP) of biomedical and clinical text to facilitate data mining, information retrieval, and natural language understanding. These systems have their respective strengths and unique capabilities, and in some cases it is desirable to combine annotations from different systems, or even to chain together these systems’ components. A practical and reproducible means of doing so has remained elusive, however, due to differences in type systems—i.e., what types of NLP annotations are available in each system and how they are defined.

Many of these systems are built around the Unstructured Information Management Architecture (UIMA) or have UIMA wrappers available. Though UIMA does not define a standardized type system, it does provide a robust and mature framework for working with annotations of any arbitrary type. We introduce AMICUS, an open-source “metasystem” that facilitates working with any number of UIMA-based NLP systems. AMICUS takes annotations from other systems and can inspect, compare, translate, and combine them according to flexible, user-defined configurations. (We previously released NLP-TAB, a tool for visually browsing UIMA annotations¹; together, NLP-TAB and AMICUS constitute a natural workflow to identify and operate on annotations of interest.)

AMICUS is unique in that it works across systems without relying upon or demanding standardization of types—indeed, part of its design philosophy is to celebrate the strengths of other systems and give end users a simple yet powerful framework to make use of the best options available for any NLP task.

System details

AMICUS itself does not perform NLP, but rather imports annotations from other systems’ outputs into its own analysis pipeline. A wide variety of pipelines can be designed from a small set of configurable and repeatable modules, where each module is designed for a specific task, such as: filtering annotations or changing their values, converting annotations to other types, combining annotations across systems, and exporting annotation values to CSV and other formats. Everything can be configured in a single text file—no programming experience is required.

The AMICUS distribution includes a “cookbook” of example pipelines for numerous applications. Users can apply these recipes to their own data, draw upon them for inspiration, or write their own from scratch. Examples of tasks that can be accomplished using a single pipeline include: converting annotations between type systems to enable serial processing by multiple other systems; filtering annotations based upon their values and/or the values of others; pooling annotations across several systems, possibly according to various voting schemes; generating statistics of annotation values; and measuring precision and recall of annotations against a manually annotated standard.

AMICUS is an open-source, cross-platform application that includes a documented Java API for developers to extend the functionality of analysis components. Additionally, all modules are implemented with uimaFIT and can be instantiated independently of the pipeline in other applications.

Development status

A public beta release is available at https://github.com/gpfinley/amicus. We have published experiments demonstrating the utility of merging disparate annotations from three different clinical NLP systems for an abbreviation disambiguation task² (the cookbook includes a recipe to repeat these experiments).

All work was supported by the National Institute of General Medical Sciences (GM102282).

References

mCerebrum and Cerebral Cortex: A Real-time Collection, Analytic, and Intervention Platform for High-frequency Mobile Sensor Data

Timothy Hnat, PhD; Syed Hossain, PhD; Nasir Ali, PhD; Simona Carini, MA; Tyson Condie, PhD; Ida Sim, MD, PhD; Mani Srivastava, PhD; Santosh Kumar, PhD

1University of Memphis, Memphis, TN; 2University of California, San Francisco, San Francisco, CA; 3University of California, Los Angeles, Los Angeles, CA

mCerebrum and Cerebral Cortex

The Center of Excellence for Mobile Sensor Data to Knowledge (MD2K) has developed two open-source software platforms (for mobile phones and the cloud) that enables the collection of high-frequency raw sensor data (at 800+ Hz for 70+ million samples/day), curation, analytics, storage (2GB/day) and secure uploads to a cloud. mCerebrum supports concurrent collection of streaming data from multiple devices (including wearables: Microsoft Band, research-grade MotionSenseHRV, EasySense, and AutoSense), phone sensors (e.g., GPS), Omron scale and blood pressure monitors and a smart toothbrush. mCerebrum continuously assesses data quality to detect issues of sensor detachment or placement on the body so they can be addressed. Data science research conducted by MD2K has resulted in 10 mHealth biomarkers: stress likelihood, smoking via hand gestures, nicotine craving, eating, lung congestion, heart motion, location, physical activity, driving, and drug use. Several of these biomarkers are computed in real-time on the phone to support biomarker-triggered Just-in-Time Adaptive Interventions (JITAI).

Cerebral Cortex is the big data companion of mCerebrum designed to support population-scale data analysis, visualization, and model development. It currently supports thousands of concurrent mCerebrum instances and provides machine-learning model development capabilities on population-scale data sets. Cerebral Cortex supports the same mHealth biomarkers that our phone platform computes, but does so on the entire dataset by using an Apache Spark powered abstraction that is designed for batch-mode computations as well as near real-time analysis via a data science dashboard. The MD2K platform uses Open mHealth mobile data exchange standards that are being harmonized with FHIR for EHR data exchange. All MD2K software is available under the BSD 2-Clause license.

By collecting and storing high-frequency raw sensor data, our approach enables external validation of computed biomarkers as well as computation of new biomarkers in the future. This benefit is akin to how biomedical studies archive biospecimens in biobanks so that they can be reprocessed to take advantage of future improvements in assays and support discoveries not possible at the time of data collection. Our platform addresses the challenges of high frequency, large volume, rapid variability and battery life limitations to enable long-lasting digital biobanks or digibanks as research utilities for mobile-driven translational research.

This demonstration will showcase data collection and visualization from multiple wearable sensor platforms and just-in-time user engagement will demonstrate how the platform can react in real-time to biomarker events. The demonstration will show how Cerebral Cortex’s capabilities for near real-time visualization of mCerebrum data are currently utilized to manage multiple concurrent clinical study participants at several sites across the United States. Finally, we will explore the interactive data science interface and showcase biomarker generation and model evaluation.

Study Deployment and Ongoing Activity

mCerebrum and Cerebral Cortex are currently deployed in ten studies at sites worldwide. These studies will result in a total of 106,806 person-days (1.3 million usable hours) of high-frequency sensor data from 2,251 unique participants. We estimate that the net data generated, processed, stored, and transmitted will be at least 150TB, consisting of more than 4.7 trillion data points.

References

1. MD2K [Internet]. 2014 [cited 16 June 2017]. Available from: https://md2k.org/
2. MD2K documentation [Internet]. 2016 [cited 16 June 2017]. Available from: http://software.md2k.org/
Enterprise Level Data Restructuring in the Military Health System – Building the Foundation for Optimization and Advanced Analytic Capability

Andrew J. Holdaway, MD, Chris Nichols, MSN, Eric A. Shry, MD
1Madigan Army Medical Center, Tacoma, WA; 2Defense Health Agency, Falls Church, VA

Introduction

The Military Health System (MHS) has historically provided data services within silos defined by branch affiliation (Army, Air Force, Navy), as well as by differing customer business requirements. This has resulted in redundancy and limited crossover of capabilities. In order to increase efficiency of medical services in the MHS, a joint directorate named the Defense Health Agency (DHA) was launched in 2012. From the outset one of the DHA’s key roles has been to standardize the collection, aggregation, storage and application of clinical and business data generated by the MHS. One of the key challenges has been to provide accurate, timely, meaningful data and business intelligence solutions. These solutions must be flexible enough to meet a variety of business, clinical and tactical needs, all while remaining standardized enough for broad implementation across a large integrated health system.

The consolidation of data services is far from complete, and adoption of a new commercial EHR is proving to be a forcing function in modernizing the data architecture. As presently constructed, the data schema for the MHS is an amalgamation of products designed to meet niche requirements derived from both clinical and operational circumstances. Specifically, the current data infrastructure is driven by 23 source systems which feed 52 distinct data domains. These systems at the source layer are joined to an integration layer by over 450 interfaces to 5 largely independent data repositories. Much of the data processing for the reporting layer is still done by manual extraction and manipulation. This in turn limits the capacity to employ meaningful real-time analytic capability.

Project Description

To address these issues, a recent project was launched within DHA to streamline these products and services. The goal is to organize these disparate data collection methods into a streamlined schema that will systematically collect, process and analyze data in near real time. The organization is informed by the concept of a layered approach to data collection, storage, aggregation and reporting. The end-state will allow for modern predictive capabilities at the application layer. This will allow leaders at all levels to access up-to-date information in near real-time. Ultimately, the desired end-state is to move the entire enterprise closer to becoming a learning healthcare system.

Demonstration Aims

This project has practical implications beyond the MHS. In the setting of hospital acquisitions and health system mergers, the challenge of migrating, integrating, and optimizing data acquisition and use is a considerable challenge. If not executed in the context of a principled framework, usability of data, and ultimately the ability to monitor and manage is diminished. In order to discuss the MHS data modernization in context, the panel will present data consolidation and management principles driven by academic reports and industry best-practices.

Specifically, this demonstration panel proposes to:

1) Describe the framework and requirements of a modern data architecture that is capable of delivering actionable data at the analytics layer for multiple customers with unique requirements.
2) Summarize the current state of the data schema of the MHS.
3) Describe the proposed MHS data modernization project in the context of industry best practices, with a focus on the potential for cost savings and enhanced reporting and analytic capabilities.

As of proposal submission, the project is in the development/planning stage with proposed execution within 3-6 months. At the time of presentation, early results will be available for discussion.
SMART on FHIR Apps from the University of Utah Interoperable Apps and Services (IAPPS) Initiative: Extending the EHR to Optimize Patient Care

Kensaku Kawamoto, MD, PhD, MHS¹, Guilherme Del Fiol, MD, PhD¹, Benjamin Brooke, MD, PhD², Charlene R. Weir, PhD, RN¹

¹Department of Biomedical Informatics, University of Utah, Salt Lake City, UT
²Department of Surgery, University of Utah, Salt Lake City, UT

Abstract

The SMART on FHIR interoperability framework enables healthcare organizations to extend the functionality of electronic health record (EHR) systems such as Epic and Cerner to improve patient care and enhance the provider experience. At the University of Utah, which uses the Epic EHR platform, a multi-stakeholder effort known as the Interoperable Apps and Services (IAPPS) initiative is developing a variety of SMART on FHIR applications to improve patient care in an interoperable, scalable manner. In this demonstration, leaders in the IAPPS initiative will demonstrate various SMART on FHIR applications that have been developed, including a neonatal bilirubin management application that has received several awards through the HHS/ONC Provider User Experience Challenge, as well as a surgical referral and communication dashboard that is being developed through the ONC High Impact Pilot program. Both the bilirubin management application and the surgical referral dashboard are available for free under an open-source license. We will also demonstrate additional SMART on FHIR applications that complement and extend the EHR through data aggregation, data visualization and clinical decision support.

Description

The mission of the University of Utah IAPPS initiative is to extend the Epic EHR to optimize patient care and the provider experience, with the additional objective of using a standards-based, interoperable approach that can be scaled to other healthcare institutions and EHR platforms. This demonstration will showcase selected applications that illustrate this approach, including a neonatal bilirubin management application, a surgical referral dashboard and several other SMART on FHIR applications designed to improve care in specific clinical areas. The neonatal bilirubin management application has won several awards.¹ Its purpose is to enable efficient and effective management of neonatal hyperbilirubinemia (Figure). Prior work by our group identified knowledge gaps in goals of care shared by primary care physicians (PCPs) and surgeons managing complex patients during the referral and follow-up period after surgery. This led to the design of a surgical referral dashboard, which facilitates bidirectional communication to support closed-looped referrals between PCPs and surgeons and improvements in patient care. The bilirubin management app is in production clinical use, and the surgical referral dashboard is under active development, with a plan for pilot use starting in the summer of 2017. Additional applications that we plan to demonstrate include a diabetes management dashboard that uses predictive modeling to guide pharmacotherapy (1.0 system completed and undergoing Epic integration), an enhanced growth chart application (physician-requested enhancements completed and undergoing Epic integration), and various other applications for data aggregation, data visualization and clinical decision support.

¹ https://www.challenge.gov/challenge/provider-user-experience-challenge/
Web-Service-Enabled Apps for Research: SMART-on-FHIR for OMOP and PCORnet
Jeffrey G. Klann, PhD1,2,3, Kavishwar B. Wagholikar, MBBS, PhD1,3, Lori C. Phillips2,
Matthew A.H. Joss2, Shawn N. Murphy, MD, PhD1,2,3
1Harvard Medical School, Boston, MA; 2Partners Healthcare, Boston, MA; 3Massachusetts General Hospital, Boston, MA

Interoperable Web Services for Research Data Repositories
Nationwide, hospital systems’ clinical data are becoming available to researchers for comparative effectiveness research. Various national initiatives are engaging IT departments to consolidate clinical data sources into institution-wide data warehouses using one of several emerging standard data formats, such as PCORNet’s Common Data Model (CDM) and the Observational Medical Outcomes Partnership (OMOP) CDM. Use of a CDM format is critical for efficient data exchange across organizations. Both PCORnet and OMOP CDMs are coded using standard terminologies and support common data elements in observational medical data. Unfortunately, there is still significant difference between each data model, in that they favor different standard terminologies and have a somewhat different set of supported data fields. Moreover, each model has its own suite of analytical programs, incompatible with the other models.

Informatics for Integrating Biology and the Bedside (i2b2), an open-source clinical data warehousing and analytics platform developed for the past decade and in use at presently over at over 100 institutions nationwide, provides a flexible web services layer which runs on top of its data model and abstracts the actual database structure into clinically meaningful XML objects, such as the Patient Data Object (PDO). These XML objects are agnostic to the underlying data model and can provide information on any data that can be modeled in i2b2’s “ontology” structure. We have previously described how this structure can be used to represent a generic Information Model to express a wide variety of data representations. [1] Therefore, i2b2 has the theoretical capability to query a variety of data models through the common i2b2 web services.

Reusable Apps Running on i2b2
The Fast Healthcare Interoperability Resources (FHIR) data format is an increasingly popular standardized messaging language for encapsulating patient data. It has added traction because it is an HL7 standard, and their official FHIR wiki presently lists 50 organizations interested in using FHIR. Electronic Health Record (EHR) systems are being outfitted with FHIR interfaces, and apps are being written to understand FHIR. Such apps will run interoperably on any FHIR interface. The SMART (Substitutable Apps, Reusable Technology) app platform, which is spearheading this approach, features, at the time of writing, twenty-five functional apps in their gallery: https://gallery.smarthealthit.org/.

We have developed a component (called a “cell”) for i2b2 which translates its XML data objects into FHIR format. [2] These can then be consumed by apps that would like to interface to the data warehouse. This interface allows apps to be developed directly on the data warehouse layer, enabling apps for research. Additionally, it serves an important use case in developing clinical apps, because it can frequently be difficult to gain access to the EHR FHIR interface, especially during app development.

Description of System
We made small changes to i2b2 to allow it to query other data models, configurable through the existing ontology system. We next adapted a previously-developed PCORnet CDM ontology to interface directly with the PCORnet data model tables. [1] We also developed an OMOP ontology which implements a subset of the OMOP data model (drug exposure, conditions, and procedures), and we also modified this to interface directly with the OMOP data model. At the time of this writing, we are able to query our OMOP and PCORnet data models directly through i2b2. We are now testing the i2b2-FHIR cell for executing SMART apps on these data models.

This enables a suite of interoperable apps to engage popular research data models without requiring institutions to transform their data. Our demonstration will feature SMART on FHIR apps running against OMOP and PCORnet via i2b2 web services. All tools used in this demo will be made available open-source.

References
CDS Connect: A New National Repository for Clinical Decision Support Knowledge Artifacts

Edwin A. Lomotan, M.D.1, Robert McCready, M.S.2, Steven Bernstein1

1Agency for Healthcare Research and Quality, Rockville, MD; 2The MITRE Corporation, Bedford, MA

Background
Clinical decision support (CDS), as a component of quality improvement and when properly implemented, has the potential to accelerate the movement of evidence into practice. This process may be further expedited by tools and infrastructure that facilitate sharing of CDS among developers, implementers, and target audiences such as patients, clinicians, health care organizations, and other stakeholders dedicated to improving health care quality. In 2016, the Agency for Healthcare Research and Quality (AHRQ) launched an ambitious multi-component initiative, focused on patient-centered CDS, with two primary aims: 1) to accelerate the movement of evidence into practice through CDS and 2) to bring CDS closer to becoming a shareable, standards-based, and publicly-available resource.

Description of Demonstration
One of the initiative’s components, known as “CDS Connect”, is a contractual effort through the MITRE Corporation to build prototype infrastructure for CDS authoring, sharing, and dissemination. CDS Connect is a combination of web-based infrastructure and software, including a repository of CDS knowledge artifacts and a CDS authoring tool. This system demonstration will feature use of both.

The CDS repository stores and makes available for sharing CDS knowledge artifacts developed by AHRQ and others. CDS knowledge artifacts can range considerably in scope, purpose, and format. They include computer-interpretable, formal logic statements derived from clinical practice guidelines to results of CDS testing, implementation guides, links to primary research evidence, and other resources designed to leverage CDS as a means for quality improvement. Attendees will learn how to access the repository and the CDS artifacts that are available. Since user feedback and interaction with the CDS community are key drivers for improving CDS Connect, attendees will also learn how to contribute feedback and participate with ongoing development efforts.

The CDS authoring tool enables the design of standards-based CDS artifacts that can be more easily shared among disparate electronic health record platforms and other technologies. The CDS authoring tool leverages the Clinical Quality Language (CQL), a Draft Standard for Trial Use under HL7. CQL is the expression language that harmonizes specifications for electronic clinical quality measures (eCQMs) and CDS. When working together, eCQMS and CDS can be powerful tools to address gaps in care and improve health care processes. Attendees will learn how to access the CDS authoring tool and will view a demonstration of its use.

CDS Connect also includes application programming interfaces (APIs) that allow for more automated interaction with the CDS repository, the CDS authoring tool, and other software sources designed to produce and consume CDS artifacts. Attendees will also learn how MITRE developed CDS for cholesterol management as part of the AHRQ initiative, piloted the CDS in a live clinical setting, and made CDS artifacts publicly-available as a proof-of-concept. Finally, attendees will learn how CDS Connect relates to other, existing platforms and about plans for maintenance of the CDS, governance, and sustainability of the infrastructure.

Author Statement
CDS Connect has been available since early March 2017. The CDS repository is undergoing constant improvement, with a more formal second release scheduled for September 2017. The CDS authoring tool is currently being built and is scheduled for its first release in September 2017. The initial design of the CDS for cholesterol management is complete, with formal piloting and refinement to occur from the spring through summer 2017. Information about the AHRQ initiative, including access to CDS Connect, can be found at http://cds.ahrq.gov.
CANARY: An Information Extraction Platform for Clinical Researchers

Shervin Malmasi, PhD¹, Nicolae L. Sandor, DM, PhD², Naoshi Hosomura, DDS, DMS²¹, Matt Goldberg³, Stephen Skentzos, MA⁴, Alexander Turchin, MD, MS, FACMI¹

¹Brigham and Women's Hospital, Boston, MA; ²Boston University, Boston, MA; ³Harvard University, Boston, MA; ⁴Kiip, Inc., San Francisco, CA

Abstract

Information extraction is a key Natural Language Processing (NLP) task for discovering and mining critical knowledge buried in unstructured clinical data. Nevertheless, widespread adoption of NLP has yet to materialize; the technical skills required for the development or use of such software present a major barrier for medical researchers wishing to employ these methods. In particular, no free NLP software geared towards biomedical researchers with little technical expertise has been made available. To address this issue we have developed CANARY, a free and open source solution designed for users without NLP, software development or engineering experience. The software allows users to model their target information using lexicons and detection criteria, ranging from simple to complex. It was designed to be fast and work out of the box via a user-friendly graphical interface.

Introduction

Unstructured, free-text documents contain vital clinical data and this has led to the development of computational methods to process and mine them for information of interest, a key task in clinical research. Despite the task’s prevalence, no readily usable tools for biomedical researchers are available. Ideally, such tools should be easy to use for researchers, even those without any NLP or software development experience. However, no such free or open source solution exists. This need prompted the development of CANARY, an NLP-based information extraction platform designed to meet these criteria. It has been developed (using Perl and Microsoft .NET) for processing clinical documents to support the extraction of text snippets via user-defined knowledge discovery criteria and lexicons.

Demonstration Description

We first outline the software’s motivation and design process. The components forming the NLP pipeline and information identification mechanism for identification of concepts of interest will be described. Next, we present an end-to-end demo covering: (1) installation; (2) a tour of the graphical user interface; (3) data acquisition; (4) text normalization & preprocessing; (5) ontology definition; (6) information discovery parameter definition; (7) structured value extraction; (8) model testing and empirical evaluation; (9) data filtering; and (10) scaling up to process big data.

This functionality will be interactively demonstrated using complex, real-world models designed by members of our research team without technical expertise. The models process large-scale (Gigabytes) datasets focusing on drug side effect discovery, medication adherence, surgical procedures, and radiology report processing, amongst others. The models have been used to publish numerous studies and we cover how the structured output generated by our software can be analyzed to obtain study results ranging from prevalence statistics to temporal analysis of patient outcomes.

Project Management

Text Normalization and Preprocessing

Ontology Definition and Management

Target Information Models and Parameters

Structured Output Criteria

Data Filtering

Runtime and Execution Settings

http://canary.bwh.harvard.edu/
Sync for Science (S4S): Helping patients share EHR data with research

Joshua C. Mandel, MD\textsuperscript{1,2}, David A. Kreda, BA\textsuperscript{1}

\textsuperscript{1}Harvard Medical School, Boston, MA \textsuperscript{2}Verily, Cambridge MA

Abstract: Sync for Science (S4S) is a new form of patient-mediated digital workflow that enables donating electronic health record (EHR) data to research projects such as the Precision Medicine Initiative's All of Us Research Program. S4S offers a streamlined, ecommerce-like user experience for donating data from any EHR system by making use of HL7 FHIR\textsuperscript{®} to represent clinical data and OAuth 2.0 for authorization. S4S leverages the HIPAA requirement that patients must be able to access their own EHR data, and the Meaningful Use Stage 3 and 2015 EHR Certification programs that require patients be able to share a Common Clinical Data Set with software apps of their choice. S4S design incorporates input from the National Institutes of Health (NIH), the Office of the National Coordinator, and the Office for Civil Rights to be sure it satisfies rigorous technical, operational, ethical, and legal requirements for patient to release their personally identifiable health information to research studies.

Purpose: Using Sync for Science (S4S) technology, research participants can donate their EHR data to research projects by executing a simple workflow between a participant-facing mobile (or web) app and a healthcare provider's patient portal (see Figure). This is particularly useful when participants are recruited from outside of traditional institutional settings, where collection of EHR data is otherwise hard to automate. To create this patient-mediated solution, our NIH-funded team worked with partner teams from several major EHR vendors to implement S4S technology consistently into their individual patient portal products.\textsuperscript{1} Each EHR's patient portal represents data using Argonaut-profiled FHIR resources including patient demographics, allergies and intolerances, immunizations, lab results, medications, documents, problems, procedures, vital signs, and smoking status.\textsuperscript{2} Each app is registered with provider systems ahead of time. Subsequently, a patient's in-portal approval causes the provider EHR to generate an access token that the app uses to retrieve clinical data. The app issues requests to retrieve historical patient data, and periodic requests for updates during a patient-determined time interval. See http://synfor.science for details.

Deployment Status: Real-world testing of S4S-enabled patient portals will be performed at twelve healthcare sites. In accordance with the All of Us Research Program IRB protocol, as many as 2,000 participants will be invited to join by using the All of Us research app to self-enroll, to consent the use of their EHR data, and, using the portal-based S4S workflow, to instruct their healthcare provider to share EHR data with the research app.

References

Genome Dashboard: An Interactive Tool for the Exploration of Human Genomic Variants and Phenotypic Associations

Katherine Miller, PhD1, Rajeswari Swaminathan, MS1, Soheil Moosavinasab, MS1, Robert Strouse, MFA1, Matthew Bailey, BA1, Yungui Huang, PhD1, Simon Lin, MD1

1The Research Institute at Nationwide Children’s Hospital, Columbus, Ohio

Introduction

Whole genome and exome sequencing (WGS/WES) are increasingly being utilized to establish the genetic basis of disease in patients. Typically, extensive bioinformatics experience is required to use complex tools to sort through thousands of sequence variants and identify potentially disease-causing variants, thus preventing many clinicians/researchers from exploring sequence variant interpretation on their own in light of new symptoms or in light of new information on genotype-phenotype associations that are constantly being discovered and curated in public databases. We developed an interactive, easy-to-use web application Genome Dashboard (GD) that enables users to easily explore sequence variant interpretation.

System Description

GD is designed as a search engine to uncover known disease-gene associations using phenotype keyword(s) search to match with entries in public knowledgebases against uploaded genome sequencing data (variant call format; VCF). We integrated knowledge from public archives of gene-phenotype relationships (Online Mendelian Inheritance in Man [OMIM], ClinVar, NCBI GeneRIFs, and Mouse & Rat Genome Database). We implemented various filters (e.g., for location in gene, molecular consequence, and variant allele frequency using gnomAD1 database) to help users refine and narrow their search. Elasticsearch is used to index annotated VCF files (via ANNOVAR2).

System Evaluation

We tested Genome Dashboard using WES data set from CLARITY 1 Challenge3 (Figure 1). We input phenotype terms (“cardiac conduction defects” and “familial heart block”) from the patient of interest. Additional filtering for location in gene (exonic) and minor allele frequency <1% identified the same variant in the TRPM4 gene reported by 13/23 groups as disease-causing in CLARITY 1 challenge1.

System Deployment

The system was developed and tested internally at Nationwide Children’s Hospital (NCH). GD has been demonstrated internally to researchers in the NCH Institute for Genomic Medicine and to physician-scientists in the Center for Cardiovascular Research at NCH. Next steps include beta-testing to genetic counselors and medical geneticists in central Ohio.

Conclusion

Genome Dashboard will give clinicians/researchers opportunity to dive through sequencing data and explore clinically relevant variants, which may help lead to diagnosis, assessment on recurrence risk, insight to prognosis, medical surveillance for suspected complications, and/or tailored approaches to clinical management. We will provide a URL during our demonstration for a public version of Genome Dashboard.

References

Systems Demonstration: Fall Prevention and Fire Departments

Andrew Muth, MD\(^1\), Christopher Hollweg, MD, MPH\(^1\), Pamela Garcia-Filion, PhD\(^1\), Michael Smith, RN\(^2\), Hamed Abbaspazegan, MD, MBA\(^2\)

\(^1\)University of Arizona College of Medicine, Phoenix, Arizona; \(^2\)Phoenix VA Healthcare System, Phoenix, Arizona

Problem
Accidental falls continue to be a leading cause of injury among the elderly population and a common reason to activate the 911 system.\(^1\) Fall screening is not routinely done, and interventions are not tailored to patients’ specific needs. There is a gap in access to standard preventative care, and the unintentional fall rate in the elderly is rising. Coupled with this is the significant associated morbidity and mortality.

Purpose
The Clinical Informatics Department at the Phoenix VA had previously established successful partnerships with both the City of Tempe Fire Department’s and the City of Chandler Fire Health & Medical’s Community Paramedic (CP) programs to identify chronically ill Veterans in need of interim care and intervene. With this project, we sought to expand the target cohort by identifying Veterans at high risk of falling or adverse outcomes from a fall. To mitigate fall risk, our objective was to intervene and connect Veterans to services for Occupational Therapy (OT) and Physical Therapy (PT).

Description
Using a novel, weighted sum model, Veterans living in the municipalities of Chandler or Tempe are identified as high risk of falling and/or adverse outcomes based on pre-defined criteria. These criteria include: a history of fall, high Morse Fall Scale score (>45), certain diagnoses (e.g. stroke, blindness), problems with ambulation, taking certain medications from Beers list (e.g. benzodiazepines, antipsychotics), recent ED visit/hospitalization, and Care Assessment Needs (CAN) score, a VA tool predicting high utilization of the system.\(^2\) Points are assigned to each criterion, with a higher value given to a more significant factor, and then summed. Patients with high scores are then scheduled via CP programs for a telehealth visit with a Nurse Practitioner. The use of telehealth allows qualified providers to assess the patient, as well as his/her environment, perform a thorough medication reconciliation, and manage the patient within the VA system in real-time.

Degree of Service Deployment
Barriers to the implementation included the coordination of a multidisciplinary team and the complexity of risk factors involved. Furthermore, many Veterans that likely would have benefited from visits declined them. Regardless, at the time of presentation, the project will have been running for almost 2 years: 1 year based solely on CAN score, and since 2017 with the current modified model. During this time, we have been able to successfully identify Veterans needing fall preventative services. The post-fall initiative cohort was much more likely to have an abnormal gait and history of fall, and the number of orders placed for all categories increased, most notably for labs, imaging, and prescriptions (half of which were fall related). More orders were also written for mobility aids and PT/OT consults. Additionally, subjective experiences from both patients and providers continue to be positive. As medicine is becoming more focused on preventative care, we anticipate this care delivery model to continue and even expand to other areas. For example, identifying patients with risk factors for myocardial infarction (eg obesity, high blood pressure, diabetes, tobacco use) would allow for interventions like lifestyle counseling, medication reconciliation/adherence, and closer monitoring of chronic conditions.

References

Applications of the Ontology Abstraction Framework

Christopher Ochs, PhD¹, James Geller, PhD¹, Yehoshua Perl, PhD¹
New Jersey Institute of Technology, Newark, NJ¹

Abstract

Biomedical ontologies are large and complex knowledge representation systems. We have developed a software system called the Ontology Abstraction Framework (OAF) to create, visualize, and explore summaries of ontologies called abstraction networks. We have used abstraction networks to support comprehension of ontology structure, ontology quality assurance, and ontology change analysis. The OAF is composed of several modules and supports ontologies in various formats. We will demonstrate several applications of ontology summarization using the OAF. Several components of the OAF, such as the Visual Semantic Delta (VSD), which identifies and summarizes the effects of editorial efforts applied to SNOMED CT, and Live Abstraction Networks (LAbNs), which are structural summaries created “on the fly” as a user edits an ontology in Protégé, will be presented.

Introduction

Ontologies often contain many thousands of concepts and tens of thousands of semantic relationships; too much information to display on a single screen. Understanding the overall structure of an ontology, and how that structure changes, is important for ontology maintenance, development, and use. We have created a software system called the Ontology Abstraction Framework (OAF) [1] to derive, visualize, and explore ontology summaries called abstraction networks. An abstraction network provides a compact view of an ontology’s content and structure. The OAF can create different types of abstraction networks for ontologies in OWL, OBO format, and SNOMED CT’s RF2 format.

The OAF is currently available for download as a free and open source software (the current version, and its associated source code, are available for download at https://njitsaboc.github.io/). The OAF is released as standalone software tools for OWL or SNOMED CT and as an OAF plugin for the Protégé ontology editor.

Different OAF modules enable the summarization of a different aspect of an ontology’s structure. We will demonstrate how the different types of ontology summaries created by the OAF can support various applications, including ontology development. An overview of the architecture of the OAF system and three recently completed modules of the OAF will be presented. Illustrative examples from prior and on-going research studies will be presented using the OAF tool.

- **Visual Semantic Delta (VSD) module:** In a recent study [2] we introduced the visual semantic delta for identifying and summarizing changes in SNOMED CT. The VSD identifies editing operations that were applied to a selected subhierarchy of concepts. Editing operations reverse engineer the editorial processes applied to a given concept, based on the delta files provided with each SNOMED CT release. The VSD provides a visual summary of major structural changes, starting from which a user can investigate a selected subset of concepts, to review their individual changes. Examples from SNOMED CT’s Bacterial infectious disease subhierarchy will be shown.

- **Live Abstraction Network (LAbN) module:** Many biomedical ontologies have been developed using Protégé. We will demonstrate how LAbNs create dynamically changing summary views of an ontology’s structure as a user is editing the ontology in Protégé. We will demonstrate how LAbNs can be used to support ontology development.

- **Aggregate Abstraction Network (AAbN) module:** Aggregate abstraction networks were designed to enable comprehension of an ontology’s content in a visual way. The overall purpose of an aggregate abstraction network is to capture the “big picture” of an ontology in a scalable, flexible manner. Aggregate abstraction networks allow a user to control the granularity of summarization of an ontology.

References


A Scalable Approach to Federating Distributed Data: The Avec® Platform

Philip R.O. Payne, PhD1,2; Mark Vance1; Peter J. Embi, MD, MS1,3
1Signet Accel, Columbus, OH; 2Washington University in St. Louis School of Medicine, St. Louis, MO; 3Regenstrief Institute, Indianapolis, IN

Abstract
Researchers, clinicians, and organizational leaders recognize that the use of advanced data analytics can improve the quality, safety, outcomes, and cost of both biomedical research and healthcare delivery. However, achieving such a vision of a data-driven ecosystem requires surmounting substantive challenges related to the collection, storage, management, analysis, and dissemination of data assets that are often distributed across multiple systems, locations, and organizational boundaries. In response to this critical need, we have designed and delivered Avec®, a data federation platform that achieves the promise of truly interoperable healthcare data across and between traditional geographic, temporal, policy, and technical boundaries. In this systems demonstration, the creators of Avec® will provide real-world examples of use cases in which Avec® can be used to facilitate hypothesis-generating research, the creation of large-scale patient-focused registries, and the use of data federation technologies to create virtual integrative data repositories at an institutional level.

Introduction
Avec® provides organizations with a scalable, flexible, and resource-efficient solution to their data sharing and integration needs. In doing so, Avec® allows organizations to put their existing data assets and technology investment to work in new and innovative ways—generating demonstrable value, all while addressing critical data security, workflow, and human-computer interaction needs.

Key Platform Characteristics
The ability of the Avec® platform to enable such capabilities reflects a number of characteristics, including:

1. **Lightweight technology deployment**: Requires minimal infrastructure investment, and can be deployed on premise using a virtual appliance model and/or using cloud computing resources;
2. **Scalable distributed query execution**: Scales and evolves gracefully to support constantly changing biomedical big data standards, types, and models in a source data platform independent manner
3. **Compatibility with diverse technology**: Offers proprietary structural mapping compatible with a substantial cross-section of standard vocabularies;
4. **Faster/easier query building**: Enables end users to build and reuse complex queries that are executed in a distributed environment across multiple data sources; and
5. **Robust data security**: Delivers security mechanisms that enable data stewards/owners to maintain control over their data and determine how it is shared and for what purpose, addressing ownership, stewardship, and valuation concerns, as well as comprehensive authentication, access control, and auditing functionality.

Conclusion
Building upon the preceding capabilities, Avec® provides organizations with an efficient, scalable, and trustworthy technology infrastructure that can meet the demands of today’s challenging data security environment and simultaneously create value through enabling data sharing and liquidity.
The All of Us Research Program Researcher Portal:
Innovative access to Unprecedented Data

Andrea H. Ramirez, MD, MS; Anthony A. Philippakis, MD, PhD; Gonçalo Abecasis, PhD; Paul A. Harris, PhD; Joshua C. Denny, MD, MS

1Vanderbilt University Medical Center, Nashville, TN; 2Broad Institute, Cambridge, MA; 3University of Michigan, Ann Arbor, MI

Abstract

The Precision Medicine Initiative All of Us Research Program\(^1\) is a national effort to recruit and engage at least one million participants who consent to provide health information (including data from health surveys, electronic health records, and baseline physical measures), biospecimens, and to be recontactable in the future. The Data and Research Center (DRC) sits at the confluence of all these varied data and, in collaboration with the many All of Us partners, has been creating a framework to shape and support biomedical research in the future. This systems demonstration will present the Researcher Portal developed to bring diverse researchers to the data for hypothesis testing and experimentation in a cloud computing environment. The Researcher Portal integrates single authentication with access to raw data and advanced methodology including notebook based programming and statistics capability, tools to identify participants of interest within the larger group, and innovative data visualization techniques. This demonstration is a key part of efforts to engage world-class researchers including the AMIA community in testing and development of this intentionally dynamic framework.

Description

The All of Us Research Program Researcher Portal serves to fulfill the intent of the Precision Medicine Initiative (PMI) to bring traditional and nontraditional researchers to the wealth of data collected on diverse longitudinal participants. The PMI broadly aims to advance generation of knowledge supporting implementation of both disease treatment and health maintenance strategies aimed at an individual, by gathering an unprecedented depth and scope of data and providing broad access to diverse researchers while maintaining the highest standards of security and privacy. The Researcher Portal will serve as a catalyst to identify areas for improvement in medical research as state of the art informatics tools enable appropriate protection and analytics support so such valuable data may facilitate innovative methodology and solutions to health threats facing our population. The individual components of this system include adaptation of the Broad Institute’s FireCloud platform to enable hosting of Jupyter notebooks with capability for collaboration during a project and supporting efforts at reproducibility with shared documentation. Cohort builder tools querying structured data democratizes access to researchers of various backgrounds and sets a standard for interoperability challenges facing medical research and clinical care. The Researcher Portal will enable hosting of external data for analysis and visualization enabling genotype and phenotype correlations between populations. This platform will be a dynamic tool growing as research uses and needs are identified with active feedback mechanisms and training forums.

Statement

As of March 9, 2017, the Researcher Portal has been conceptualized including policy and architecture framework and testing of individual components and interoperability is underway. Ensuing development will include integration of tools to a unified platform with simulated data ahead of federal Authorization to Operate approval when All of Us participant data will be hosted.

\(^1\)Precision Medicine Initiative, PMI, All of Us, the All of Us logo, and The Future of Health Begins With You are service marks of the U.S. Department of Health and Human Services.
DocUBuild: A Collaborative System to Enhance Dissemination and Discovery of Genomic Clinical Content

Luke V. Rasmussen, MS¹, Casey L. Overby, PhD²,³
¹Northwestern University Feinberg School of Medicine, Chicago, IL; ²Johns Hopkins University, Baltimore, MD; ³Geisinger Health System, Danville, PA

System Description

As genomic medicine programs continue to expand, there is also a growing need to deliver sufficient and high quality information resources to aid patients and healthcare providers in their comprehension and decision-making. As part of the NHGRI-funded electronic Medical Records and Genomics (eMERGE) network (a national consortium developing methods and best practices for using the EHR for genomic medicine and research), we have created the DocUBuild system to support developing and sharing genomic medicine information that is accessible through context-sensitive links in the medical record (“infobuttons”).

A recent survey of sites involved with the eMERGE network and Clinical Sequencing Exploratory Research (CSER) consortium indicated that institutions continue to rely on locally-authored content (synthesized from external resources), and that similar content may need to be delivered as multiple modalities (e.g., printed handout, website, e-mail, printed along with an after visit summary)¹. In addition, some content may require local customizations (e.g., site-specific contact information, institutional branding), which precludes the reuse of existing content by others. While many document management and content management systems exist that allow rich editing and collaborative environments, many challenges to managing genomic information resources have not yet been addressed. A recent survey, for example, found that some genomic medicine resources only support keyword searching and for those that do allow searching, typically only use a single standard terminology (e.g., OMIM)². This highlights an opportunity to increase the amount of relevant metadata associated with individual resources (e.g., the use of other relevant terminologies to enable more robust searching). Furthermore, we believe there is a need to streamline sharing and rebranding of content for use across institutions, and a need to enable quick and easy deployment of content to multiple endpoints.

To address these gaps, we have developed DocUBuild (https://docubuild.fsm.northwestern.edu/) as part of the electronic health record (EHR) integration efforts within the eMERGE network. DocUBuild provides infrastructure to support: a) authoring content using structured templates, b) making content discoverable through the annotation of documents and individual sections with context elements according to the HL7 infobutton standard, c) sharing content and discovering content authored by other institutions, and d) deploying content in multiple formats in order to meet individual needs.

This system demonstration will show how DocUBuild may be used in different content authoring workflows, and how the output of DocUBuild can facilitate information retrieval. A particularly novel aspect to be shown is the ability to quickly and easily annotate an article with infobutton context elements, including the use of structured templates where infobutton context is already applied. Furthermore, we will demonstrate the ability to push or pull information resources to a variety of external systems, including OpenInfobutton⁵, and how this may reduce duplication and burden of keeping content in sync in disparate systems.

Deployment Status

The base DocUBuild system has been made publicly accessible as of the time of submission. Certain features are currently in the development phase and are not yet publicly accessible (OpenInfobutton responder integration), but will be prior to the system demonstration.

References

Creating a Cluster of Preclinical Science Lessons Using Pharmacogenomic-Focused Patient Case Presentations

Blaine Takesue, MD1,2, Maureen Harrington, PhD1, Bradley Allen, MD, PhD1, Debra Litzelman, MD1,2, Paul Dexter, MD1,2

1Indiana University School of Medicine, Indianapolis, IN; 2Regenstrief Institute, Indianapolis, State, IN

Introduction
Genetic testing and the use of genetic tests results in clinical decision making have become more prevalent over the past decade. Despite these advances, many physicians acknowledge their pharmacogenomic knowledge base is not adequate to take full advantage of genomics-based information in medication ordering.1 To address the gap between the state of the art pharmacogenomics knowledge and provider needs we created patient care vignettes through which students could learn pharmacogenomics principles.

The Indiana University School of Medicine (IUSOM) is home to the INGENIOUS (INdiana GENomics Implementation: an Opportunity for the UnderServed) trial.2 This NIH-funded study is investigating the implementation issues around pharmacogenomics recommendations in an urban hospital. As part of this study, the INGENIOUS team created pharmacogenomics clinical decision support (CDS) rules for the production Regenstrief Gopher computerized provider order entry system (CPOE). The team created 87 rules covering 28 medications and 97 medical problems (Figure 1).

In 2013, the Regenstrief Institute created a teaching electronic health record (tEHR). The tEHR is a branch of the production Regenstrief Gopher CPOE. As opposed to the clinical role of the Gopher, the tEHR is designed to support health professions education. The tEHR is a stand-alone application that uses anonymized, real patient data. The system allows students to provide virtually care for patients in a safe and realistic environment.

We imported the pharmacogenomic CDS rules from the Gopher to the tEHR). In the tEHR, we propose that alerts created by the CDS will prompt discussions of pharmacogenomic principles and the role of informatics in presenting pharmacogenomic data. Extending this model, we plan to use these vignettes to link pharmacogenomic patient cases to topics in biochemistry, genetics, pharmacology, and physiology.

System Demonstration
After discussing the design principles behind pharmacogenomics-focused patient vignettes, we propose to demonstrate a learner’s experience by working through a pharmacogenomics-focused case. In the demonstration, we will show 1) how background case information is delivered through the tEHR, 2) how orders trigger alerts based on the patient’s genomic data and medication selected 3) how educational alerts create opportunities for teaching and learning biochemical, genetic, pharmacologic, and physiologic content.

Relevance/Impact
There is a gap between the pharmacogenomic knowledge base and the knowledge base of many practicing providers. This podium presentation/demonstration will attract educators who are interested in pharmacogenomics training methods for providers early in their professional career. The presentation will also interest those who are contemplating the use of electronic health records and specifically the tEHR, as an instructional tool.

System deployment status
As of the submission date, the Regenstrief tEHR is deployed to 8 health professions colleges/schools.

References
1. Taber, K, Dickinson, B. Pharmacogenomic knowledge gaps and educational resource needs among physicians in selected specialties. Pharmacogenomics and Personalized Medicine 2014:7 145–162

2309
CLAMP – A User-Centric Clinical Natural Language Processing Toolkit

Hua Xu, PhD, Ergin Soysal, MD, PhD, Jingqi Wang, Min Jiang, Yonghui Wu
School of Biomedical Informatics, University of Texas Health Science Center at Houston, Houston, TX

Abstract
In this demonstration session, we will introduce a new clinical natural language processing (NLP) system called CLAMP (Clinical language annotation, modeling, and processing) to the medical informatics community. Similar to cTAKES and MetaMap, CLAMP is another comprehensive, general-purpose clinical NLP system, representing a significant amount of development effort by the NLP team at UTHealth. The motivation for developing CLAMP is to support non-technical end users to take advantage of the power of the state-of-the-art machine learning/hybrid approaches in clinical NLP, by implementing novel algorithms and designing user-friendly interfaces. CLAMP builds on a set of high performance NLP components that were proven in previous clinical NLP shared tasks such as i2b2, SHARE/CLEF, and SemEVAL challenges. In addition to the default Apache UIMA interfaces for system development, we developed graphical user interfaces that allow end users to easily manage their projects by annotating clinical text and managing clinical corpora, training machine learning modules for different components such as named entity recognition (NER), and building customized pipelines with different components. Currently, CLAMP version 1.0 is being used in both UTHealth and MD Anderson Cancer Center, and it is already downloaded and used by 80 users in more than 30 research institutes.

Motivations for developing CLAMP
Currently, there are many text processing tools specialized to clinical narratives, e.g., MetaMap, cTAKES, etc. Despite the great contribution and success of these existing tools, they generally require certain level of programming skills if a user wants to customize an NLP system for his/her specific tasks. For example, our experience showed that a machine learning-based smoking status detection system needed to be retrained in order to achieve desirable performance in a different institute; however, it requires knowledge and skills on machine learning technologies, which many end users may not have. Therefore, the goal of CLAMP is to enable non-technical users to utilize the state-of-the-art algorithms to build customized NLP systems for their applications.

Features of CLAMP
Highlighted features of CLAMP include:
2. **Annotation and corpora management**: Users can import clinical text corpora into the CLAMP workspace and annotate files using the built-in annotation tool. Then, this annotated corpus can be utilized by CLAMP projects, both for development as train and test data for a model, or in production running with the complete pipeline.
3. **Machine learning and hybrid approaches**: Depending on the task, users can train their own model for machine learning based components. CLAMP serves required functionality to train and evaluate custom models using a custom corpus for that component. Results can be evaluated by n-fold cross validation within CLAMP.
4. **Customized pipelines**: CLAMP offers all common components, like named entity recognition, assertion, and UMLS encoder to build an NLP pipeline. For some tasks, CLAMP may provide variations of a component that may utilize rule based methods, or machine learning algorithms such as Brown clustering, SVM, and CRF.
5. **Knowledge sources and sample clinical text**: All knowledge resources required by CLAMP components – e.g. dictionaries, section header list or medical abbreviation list – are also provided, in addition to a corpus with 400 annotated notes derived from MTSamples. So, user can start to develop and test NLP pipelines, immediately.
6. **Interoperability and Scalability**: CLAMP was built on the UIMA framework, so that it is compatible with other systems such as cTAKES. Moreover, it adopts cTAKES’ type system for lower linguistic level annotations, to increase interoperability. We also plan to develop a Hadoop version of CLAMP in next release to support distributed computing for big datasets.

Current Development Status
Developing general-purpose clinical NLP systems such as MetaMap and cTAKES requires substantial amounts of efforts and resources. Three programmers in our team have worked on CLAMP for almost two years. Prototypes of CLAMP have been used in various projects in UTHealth, as well as MD Anderson Cancer Center since 2013. Now we released CLAMP 1.2.1 to the public, with a specialized version for Cancer. A manuscript of this new tool is in preparation and will be submitted to JAMIA. We believe such a user-centric clinical NLP tool will greatly benefit the end users, thus facilitating clinical research and healthcare practice.